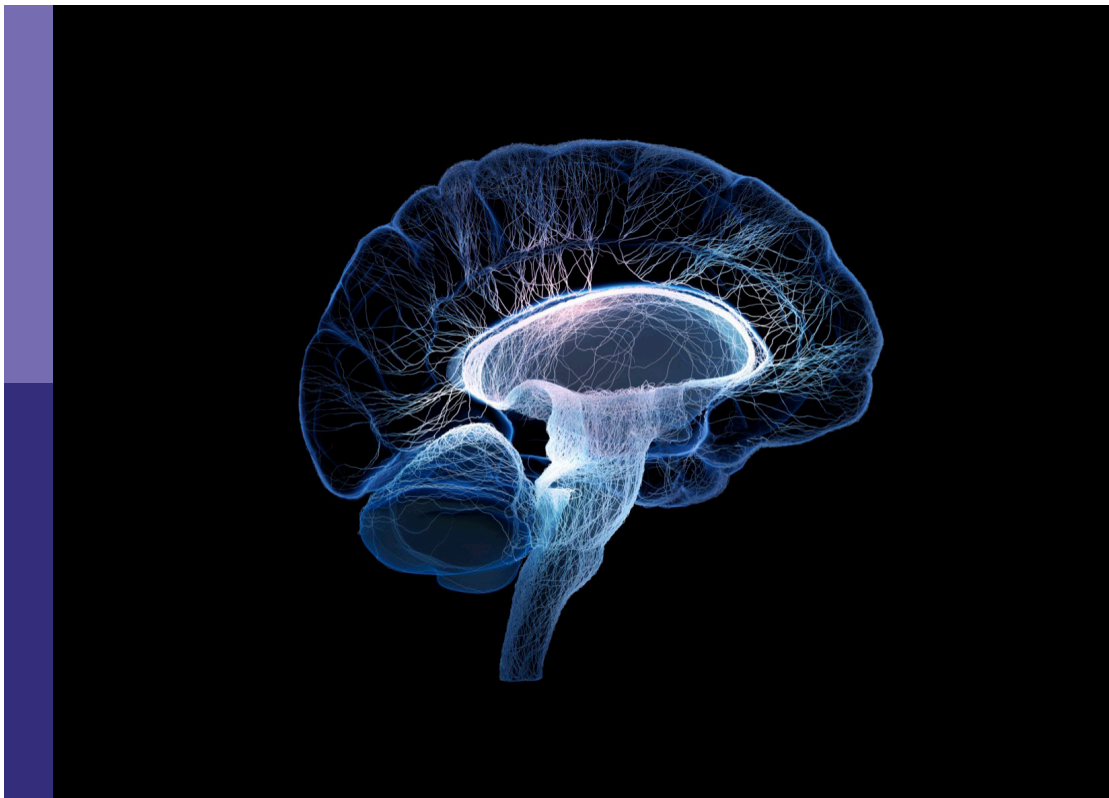
frontiers | Research Topics

Machine learning in neuroscience

Edited by
Hamid R. Rabiee, Ali Ghazizadeh, Mackenzie W. Mathis and Reza Lashgari

Published in
Frontiers in Neuroscience
Frontiers in Human Neuroscience
Frontiers in Computational Neuroscience
Frontiers in Neural Circuits



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83251-029-2
DOI 10.3389/978-2-83251-029-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Machine learning in neuroscience

Topic editors

Hamid R. Rabiee — Sharif University of Technology, Iran
Ali Ghazizadeh — Sharif University of Technology, Iran
Mackenzie W. Mathis — Harvard University, United States
Reza Lashgari — Shahid Beheshti University, Iran

Citation

Rabiee, H. R., Ghazizadeh, A., Mathis, M. W., Lashgari, R., eds. (2023). *Machine learning in neuroscience*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-029-2

Table of contents

06 Non-invasive Decoding of the Motoneurons: A Guided Source Separation Method Based on Convolution Kernel Compensation With Clustered Initial Points
Mohammad Reza Mohebian, Hamid Reza Marateb, Saeed Karimimehr, Miquel Angel Mañanas, Jernej Kranjec and Ales Holobar

20 A Regression-Based Framework for Quantitative Assessment of Muscle Spasticity Using Combined EMG and Inertial Data From Wearable Sensors
Xu Zhang, Xiao Tang, Xiaofei Zhu, Xiaoping Gao, Xiang Chen and Xun Chen

32 STEFTR: A Hybrid Versatile Method for State Estimation and Feature Extraction From the Trajectory of Animal Behavior
Shuhei J. Yamazaki, Kazuya Ohara, Kentaro Ito, Nobuo Kokubun, Takuma Kitanishi, Daisuke Takaichi, Yasufumi Yamada, Yosuke Ikejiri, Fumie Hiramatsu, Kosuke Fujita, Yuki Tanimoto, Akiko Yamazoe-Umemoto, Koichi Hashimoto, Katsufumi Sato, Ken Yoda, Akinori Takahashi, Yuki Ishikawa, Azusa Kamikouchi, Shizuko Hiryu, Takuya Maekawa and Koutarou D. Kimura

55 Stimulus-Specific Adaptation Decreases the Coupling of Spikes to LFP Phase
Mohsen Parto Dezfouli, Mohammad Zarei, Mehran Jahed and Mohammad Reza Daliri

66 Discovering Implied Serial Order Through Model-Free and Model-Based Learning
Greg Jensen, Herbert S. Terrace and Vincent P. Ferrera

90 The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines
David B. Parker and Qolamreza R. Razlighi

112 The Concept of Transmission Coefficient Among Different Cerebellar Layers: A Computational Tool for Analyzing Motor Learning
Saeed Solouki, Fariba Bahrami and Mahyar Janahmadi

129 Tonality Tunes the Statistical Characteristics in Music: Computational Approaches on Statistical Learning
Tatsuya Daikoku

141 Quantifying Neurodegenerative Progression With DeepSymNet, an End-to-End Data-Driven Approach
Danilo Pena, Arko Barman, Jessika Suescun, Xiaoqian Jiang, Mya C. Schiess, Luca Giancardo and the Alzheimer's Disease Neuroimaging Initiative

155 Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates
David Xing, Mehdi Aghagolzadeh, Wilson Truccolo, Erwan Bezard, Gregoire Courtine and David Borton

165 Corrigendum: Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates
David Xing, Mehdi Aghagolzadeh, Wilson Truccolo, Erwan Bezard, Gregoire Courtine and David Borton

167 Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia
Raymond Salvador, Erick Canales-Rodríguez, Amalia Guerrero-Pedraza, Salvador Sarró, Diana Tordesillas-Gutiérrez, Teresa Maristany, Benedicto Crespo-Facorro, Peter McKenna and Edith Pomarol-Clotet

176 Analyzing Neuroimaging Data Through Recurrent Deep Learning Models
Armin W. Thomas, Hauke R. Heekeren, Klaus-Robert Müller and Wojciech Samek

194 Using Deep Learning and Resting-State fMRI to Classify Chronic Pain Conditions
Alex Novaes Santana, Ignacio Cifre, Charles Novaes de Santana and Pedro Montoya

207 Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network
Zeinab Sherkatghanad, Mohammadsadegh Akhondzadeh, Soorena Salari, Mariam Zomorodi-Moghadam, Moloud Abdar, U. Rajendra Acharya, Reza Khosrowabadi and Vahid Salari

219 Detecting the Information of Functional Connectivity Networks in Normal Aging Using Deep Learning From a Big Data Perspective
Xin Wen, Li Dong, Junjie Chen, Jie Xiang, Jie Yang, Hechun Li, Xiaobo Liu, Cheng Luo and Dezhong Yao

231 Glioma Migration Through the Corpus Callosum and the Brainstem Detected by Diffusion and Magnetic Resonance Imaging: Initial Findings
Guive Sharifi, Amir Mohammad Pajavand, Saeedeh Nateghinia, Tohid Emami Meybodi and Hossein Hasooni

239 Deep Temporal Organization of fMRI Phase Synchrony Modes Promotes Large-Scale Disconnection in Schizophrenia
Tahereh S. Zarghami, Gholam-Ali Hossein-Zadeh and Fariba Bahrami

265 A Deep Learning Approach for Mild Depression Recognition Based on Functional Connectivity Using Electroencephalography
Xiaowei Li, Rong La, Ying Wang, Bin Hu and Xuemin Zhang

285 A Constrained ICA-EMD Model for Group Level fMRI Analysis
Simon Wein, Ana M. Tomé, Markus Goldhacker, Mark W. Greenlee and Elmar W. Lang

295 MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data
Matthias S. Treder

314 Automatic Reconstruction of Mitochondria and Endoplasmic Reticulum in Electron Microscopy Volumes by Deep Learning
Jing Liu, Linlin Li, Yang Yang, Bei Hong, Xi Chen, Qiwei Xie and Hua Han

327 The Importance of Anti-correlations in Graph Theory Based Classification of Autism Spectrum Disorder
Amirali Kazeminejad and Roberto C. Sotero

338 Gender Perception From Gait: A Comparison Between Biological, Biomimetic and Non-biomimetic Learning Paradigms
Viswadeep Sarangi, Adar Pelah, William Edward Hahn and Elan Barenholtz

349 Machine Learning Methods for Fluorescence Lifetime Imaging (FLIM) Based Label-Free Detection of Microglia
Md Abdul Kader Sagar, Kevin P. Cheng, Jonathan N. Ouellette, Justin C. Williams, Jyoti J. Watters and Kevin W. Eliceiri

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 02 April 2019
doi: 10.3389/fncom.2019.00014

Non-invasive Decoding of the Motoneurons: A Guided Source Separation Method Based on Convolution Kernel Compensation With Clustered Initial Points

Mohammad Reza Mohebian[1], Hamid Reza Marateb[1], Saeed Karimimehr[1,2], Miquel Angel Mañanas[3,4], Jernej Kranjec[5] and Ales Holobar[5]*

[1] The Biomedical Engineering Department, Engineering Faculty, University of Isfahan, Isfahan, Iran, [2] Brain Engineering Research Center, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran, [3] Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain, [4] Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain, [5] Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

OPEN ACCESS

Edited by:
Mackenzie W. Mathis, Harvard University, United States

Reviewed by:
Ping Zhou, University of Texas Health Science Center at Houston, United States
Alexander Boettcher, Werner Reichardt Centrum für Integrative Neurowissenschaften (CIN), Universität Tübingen, Germany

***Correspondence:**
Hamid Reza Marateb
h.marateb@eng.ui.ac.ir

Received: *20 January 2018*
Accepted: *26 February 2019*
Published: *02 April 2019*

Citation:
Mohebian MR, Marateb HR, Karimimehr S, Mañanas MA, Kranjec J and Holobar A (2019) Non-invasive Decoding of the Motoneurons: A Guided Source Separation Method Based on Convolution Kernel Compensation With Clustered Initial Points. Front. Comput. Neurosci. 13:14. doi: 10.3389/fncom.2019.00014

Despite the progress in understanding of neural codes, the studies of the cortico-muscular coupling still largely rely on interferential electromyographic (EMG) signal or its rectification for the assessment of motor neuron pool behavior. This assessment is non-trivial and should be used with precaution. Direct analysis of neural codes by decomposing the EMG, also known as neural decoding, is an alternative to EMG amplitude estimation. In this study, we propose a fully-deterministic hybrid surface EMG (sEMG) decomposition approach that combines the advantages of both template-based and Blind Source Separation (BSS) decomposition approaches, a.k.a. guided source separation (GSS), to identify motor unit (MU) firing patterns. We use the single-pass density-based clustering algorithm to identify possible cluster representatives in different sEMG channels. These cluster representatives are then used as initial points of modified gradient Convolution Kernel Compensation (gCKC) algorithm. Afterwards, we use the Kalman filter to reduce the noise impact and increase convergence rate of MU filter identification by gCKC. Moreover, we designed an adaptive soft-thresholding method to identify MU firing times out of estimated MU spike trains. We tested the proposed algorithm on a set of synthetic sEMG signals with known MU firing patterns. A grid of 9 × 10 monopolar surface electrodes with 5-mm inter-electrode distances in both directions was simulated. Muscle excitation was set to 10, 30, and 50%. Colored Gaussian zero-mean noise with the signal-to-noise ratio (SNR) of 10, 20, and 30 dB, respectively, was added to 16 s long sEMG signals that were sampled at 4,096 Hz. Overall, 45 simulated signals were analyzed. Our decomposition approach was compared with gCKC algorithm. Overall, in our algorithm, the average numbers of identified MUs and Rate-of-Agreement (RoA) were 16.41 ± 4.18 MUs and 84.00 ± 0.06%, respectively, whereas the gCKC identified 12.10 ± 2.32 MUs with the

average RoA of 90.78 ± 0.08%. Therefore, the proposed GSS method identified more MUs than the gCKC, with comparable performance. Its performance was dependent on the signal quality but not the signal complexity at different force levels. The proposed algorithm is a promising new offline tool in clinical neurophysiology.

Keywords: electromyography, kalman filter, motor unit identification, neural decoding, source separation

INTRODUCTION

A connectivity is generated between the motor-related areas of the brain during movement (Sporns et al., 2004; Rubino et al., 2006; Kim et al., 2017). A top-down structure in motor control is used particularly during an upper limb movement (with many degrees of freedom), where the brain and the muscles are functionally combined so that the muscle receives and electrically amplifies the resultant neural commands from the motor system (Haggard, 2008).

Over the last two decades, the non-invasive techniques for assessment of cortical and muscular activity have demonstrated significant progress and linear (e.g., coherence analysis) or non-linear (e.g., mutual information) methods have been used to analyze the relations between electromyographic (EMG), and electroencephalographic (EEG) signals (Chen et al., 2008; Meng et al., 2008). Due to the non-linear transfer functions of motor neurons, non-linear methods are more suitable for such a dependence analysis (Hashimoto et al., 2010; Ioannides and Mitsis, 2010) and for analysis of neural transfer functions between the central nervous system and pool of active motor units (MUs) (Negro and Farina, 2011; Gallego et al., 2015a,b).

Despite this progress in understanding of neural codes, the cortico-muscular coupling studies are still largely dependent on interferential EMG signal or its rectification for the assessment of motor neuron pool behavior (Gross et al., 2001; Schelter et al., 2009; Artoni et al., 2017). This assessment is not straightforward since the amplitude of surface EMG demonstrates several anatomical properties of the muscles, significantly interfering with the neural commands (Farina et al., 2010; Farina and Holobar, 2015, 2016). Indeed, amplitude or frequency of EMG are considerably affected by many factors, such as muscle anatomy, low-pass filtering of the subcutaneous tissues and MU Action Potential (MUAP) cancelation (Merletti and Farina, 2016). As result, they are not precisely related to the ongoing motoneuron activities and provide only a crude estimate of the central neural commands (so called neural drive) to the skeletal muscles (Farina et al., 2004, 2014a; Merletti and Farina, 2016). Therefore, the assessment of cortico-muscular connectivity and neuromuscular coupling via EEG–rectified EMG relationship is non-trivial and should be used with precaution.

Direct analysis of neural codes by decomposing the EMG (Webster et al., 2016), also known as neural decoding (Farina et al., 2014b; Karimimehr et al., 2017), is an alternative to EMG amplitude estimation (Gallego et al., 2017; Úbeda et al., 2017). It represents a paradigm shift, because it enables direct assessment of the neural drive to muscles (Farina and Holobar, 2015, 2016; Karimimehr et al., 2017). In fact, MU identification can be thought as a spike sorting algorithm (typical for computational neuroscience) applied to the outer layer of the human motor system (Balasubramanian and Obeid, 2011; Pani et al., 2016). The results of this MU identification can not only be used in EMG-EEG coupling analysis, but also in variety of research areas such as clinical neurophysiology for diagnosing neuromuscular disorders (Wheeler et al., 2006; Povalej BrŽan et al., 2017), sports and behavioral science (Merletti and Parker, 2004), movement science (Winter, 2009), robot-assisted rehabilitation (Savc et al., 2018), brain machine interface (Werner et al., 2016), and prosthesis control (Yoshida et al., 2010; Farina et al., 2014a).

In a typical experimental setup, EMG signals are detected using either conventional invasive intramuscular electrodes or non-invasive surface electrodes (Merletti and Parker, 2004). Both intramuscular (iEMG) and surface EMG (sEMG) signals include MUAP trains, superimposed into interferential signal patterns. Although iEMG provides some advantages, such as recording from deep muscles, it also has problems, such as discomfort and high selectivity. sEMG is thus a good alternative, particularly in sport sciences and studies of children (Ghaderi and Marateb, 2017). On the other hand, substantial MUAP superposition occurs in sEMG signals (Farina and Holobar, 2015, 2016). Also, surface MUAP shapes from different MUs are rather similar due to the volume conductor effects (Chen and Zhou, 2016). Thus, surface EMG decomposition is considered a very difficult task (Zhou and Rymer, 2004).

Variety of sEMG decomposition methods have been introduced in the literature. Generally, they either use shape-based algorithms, also called template matching (Xu et al., 2001; Gazzoni et al., 2004; Garcia et al., 2005; De Luca et al., 2006; Ren et al., 2006; Kleine et al., 2007; Nawab et al., 2008; Winslow et al., 2009; Siqueira Júnior and Soares, 2015) or the blind source separation (BSS) algorithm (Holobar and Zazula, 2004, 2007; Glaser et al., 2013; Ning et al., 2013, 2015; Chen and Zhou, 2016; Negro et al., 2016; Savc et al., 2018).

Although complementary fusion often results in more reliable findings (Durrant-Whyte, 1988), it is not yet sufficiently clear how to combine template-based and BSS sEMG decompositions, especially, as the aforementioned two methodological approaches differ substantially in the data model assumptions. For example, template-based algorithms assume a few recorded EMG channels and relatively low (or progressively increasing) muscle excitation levels in order to (progressively) identify the MUAP templates. They then use MUAP peel-off approach and combinatorial methods supported with artificial intelligence algorithms to identify each individual MU firing (Nawab et al., 2010). BSS approaches, on the other hand, model the mixing process of MUAPs in EMG, invert it and apply the inverse mixing procedure

directly to the recorded EMG signals. For this reason, they do not rely on the initialization of MUAP templates, but do require relatively large number of recorded EMG channels to identify all the MU firings at once. In other words, BSS approaches estimate the MU filter directly in the space of MU spike trains and, once MU filter is estimated, apply it very efficiently (in terms of computational costs) to the recorded EMG signals, yielding all the MU firings at once (Holobar and Farina, 2014). Combination of these two fundamentally different sEMG decomposition approaches is non-trivial as it imposes complex methodological steps in which many parameters must be fine-tuned.

In this study, we focus on a hybrid sEMG decomposition approach that combines the advantages of both template-based and BSS decomposition approaches. Briefly, we use the single-pass density-based clustering algorithm to identify possible cluster representatives in different sEMG recording channels. Unlike traditional BSS-based decomposition algorithms, in which the initial search positions are randomly selected and checked on a trial-and-error basis, we use cluster representative samples as initial points. Similar approach was proposed in Ning et al. (2015), where *k*-means method was used to cluster MUAPs and, therefore, improve the initial estimate of MU filter in classical CKC approach. On the other hand, Chen et al. (2016, 2018a) and (Chen and Zhou, 2016) proposed the progressive FastICA framework along with the advanced peel-off and valley-seeking process that efficiently enhances the initialization of the MU filters and speeds-up the decomposition.

We then use the Kalman filter and a modified Gradient CKC (gCKC) algorithm (Holobar and Zazula, 2008) to further increase the rate of convergence of MU filter identification. Moreover, we introduce a new non-linear soft-thresholding algorithm to reduce the False Positives (FP) and False Negatives (FN) in MU spike post-processing. Thus, the proposed hybrid algorithm can be considered a "Guided Blind Source Separation" method.

The rest of the paper is organized as follows: in the next section, information about the signals and formulation of methods used in this study is presented. Section Results provides the results of the proposed method. The discussion is provided in section Discussion, along with conclusions.

MATERIALS AND METHODS

Simulated Signals

A planar volume conductor model was used for generating synthetic sEMG signals (Farina and Merletti, 2001). Muscle, fat and skin tissues were used in the non-homogeneous and anisotropic volume conductor. It included an inimitable, and semi-infinite muscle layer with cross-section of 30 mm (transversal) × 15 mm (depth). The average fiber length, isotropic subcutaneous and skin layer thickness were 130 mm, 4 mm and 1 mm, respectively. Each MU had a random number of fibers uniformly distributed between 24 and 2,048 with the circular territories of 20 fibers/mm^2. The conduction velocities were normally distributed (4.0 ± 0.3 m/s). In the initial recruitment, each MU discharged at 8 pulses per second (pps) (Fuglevand et al., 1993). Its discharge rate increased linearly with excitation (0.3 pps per % of muscle excitation). A grid of 9 × 10 (9 columns, 10 rows) monopolar surface electrodes with 5-mm inter-electrode distances in both directions was simulated. Fifteen sEMG signals with length of 16 s were generated and sampled at 4,096 Hz. Muscle excitation was set to 10, 30, and 50% Maximum Voluntary Contraction (MVC), yielding 262, 388, and 446 active motor units. Noteworthy, regardless the decomposition algorithm used, we can identify only superficial motor units, whereas small and distant motor units contribute to the physiological noise. In addition to these simulated physiological noise, colored Gaussian zero-mean noise with the signal-to-noise ratio (SNR) of 10, 20, and 30 dB and the bandwidth of 20–500 Hz was added to the raw surface EMG signals. Overall, 45 simulated signals were analyzed in this study. The simulated signals are available online: https://doi.org/10.6084/m9.figshare.5808291.

The Proposed Algorithm

The structure of the proposed algorithm is depicted in **Figure 1**. Briefly, the signal was passed to the first-order band-pass Butterworth filter with the cut-off frequencies of 20 and 500 Hz in the forward and reverse direction. Furthermore, a whitening method was performed on sEMG signals (Thomas et al., 2006). Then, the whitened signal was used for initial point estimation as well as for the modified gCKC. Moreover, the convergence rate of the gCKC algorithm was improved by the Kalman filter. The detailed information about each aforementioned step is provided in the sequel.

Whitening

Whitening, referred to as "convolutive sphering" (Thomas et al., 2006) was used as the first step of sEMG decomposition. Assume the following signal model,

$$\mathbf{X} = \mathbf{WZ} \tag{1}$$

where $\mathbf{Z}$ is the $N \times M$ extended EMG signal matrix in which each row is a delayed repetition of one of EMG channels as proposed in Holobar and Zazula (2004) and Thomas et al. (2006) and each column corresponds to a time sample. In this study, the number of delayed repetitions of each signal is dependent to the sampling frequency and number of channels and in this work it was fixed to 30. The number of Whitening matrix $\mathbf{W}$ could be obtained, provided that the covariance matrix of $\mathbf{X}$ at time lag zero is equal to the identity matrix (Belouchrani et al., 1997):

$$W = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\,\mathbf{U}^T \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix obtained by the eigenvalue decomposition of the covariance matrix of $\mathbf{Z}$ and $\mathbf{U}$ is the modal matrix:

$$E\left\{\mathbf{Z}\mathbf{Z}^T\right\} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{3}$$

with $E\left\{\mathbf{Z}\mathbf{Z}^T\right\}$ denoting the covariance matrix of $\mathbf{Z}$.

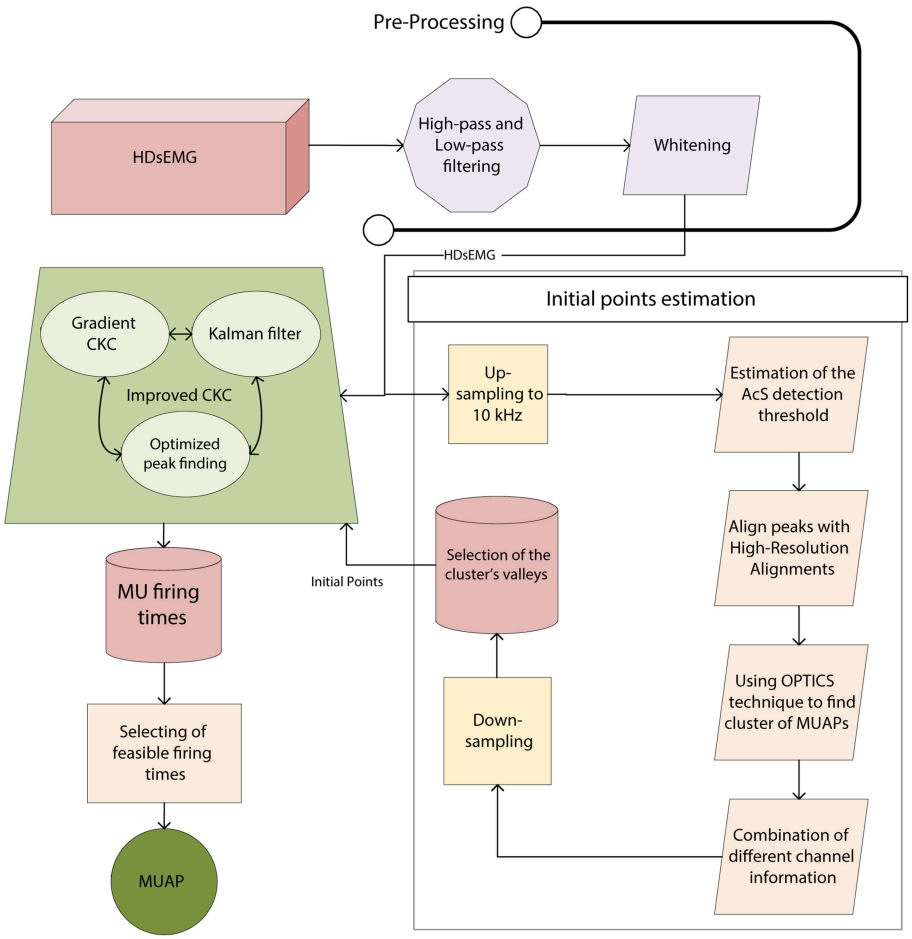


FIGURE 1 | The schema of the proposed decomposition algorithm. The sEMG signal is first filtered using band-pass and whitening filters. The template-based clustering algorithm is then used to identify the initial points for the modified gCKC algorithm. Such a clustering algorithm includes the segmentation and high-resolution alignment of the up-sampled signal. Then, a modified density-based clustering OPTICS algorithm is used to automatically locate cluster representatives (i.e., templates) in different recording channels. Such information is combined for different channels and the peak samples of the decimated templates are used to initialize the modified g-CKC algorithm. Finally, Kalman filtering and optimized peak finding is implemented to increase the efficiency of the algorithm and also to reduce the decomposition errors.

Initial Point Estimation

Similar to the shape-based sEMG decomposition approach in De Luca et al. (2006), the time samples of the sEMG signals were up-sampled to 10 KHz to increase the time resolution (**Figure 1**). The resampling was based on band-limited approximation of the original signal, produced by inserting zeros into the discrete Fourier transform of the waveform (Crochiere, 1979). We also included steps from Stearns and Hush (2011) to reduce the end effects caused by the zero-insertion.

Next, the signal detection threshold, required by our segmentation, was calculated as follows: the absolute values of the EMG samples were calculated and sorted (vector $\boldsymbol{s}$). The square root values of cumulative sum of $\boldsymbol{s}$ were then calculated ($\boldsymbol{ss}$). The maximum sample index i, where $\boldsymbol{ss}(i)$ multiplied by K exceeded the $\boldsymbol{s}(i)$ was found and $\boldsymbol{s}(i)$ was used as the detection threshold. In our study, K was empirically set to 4.0. The segmentation was then performed on each recording channel independently, using the estimated detection threshold on 2 ms intervals. Two-millisecond long signal segments, whose peak value was more than the detection threshold, were entitled as active segments (AcSs) (McGill et al., 2005).

Then, a high-resolution peak alignment method was used to align the detected AcSs on the highest peak (McGill and Dorfman, 1984). The time lags of the aligned AcSs were used as features for clustering. The density-based clustering was performed by using Ordering points to identify the clustering structure (OPTICS) algorithm (Ankerst et al., 1999; Daszykowski et al., 2002). This algorithm defines the clusters as areas of higher density. Among different density-based clustering algorithms, OPTICS was shown to be suitable for iEMG decomposition in which clusters have different dispersion (Marateb et al., 2011b). It is a single-pass clustering algorithm and unlike K-means or any other aggregative clustering method, does not require multiple runs on different predefined number of clusters (Daszykowski

et al., 2002; Larose, 2005). The original OPTICS algorithm requires two parameters, namely the minimal number of points in a cluster MinPts and the Euclidean distance ε. A point p is called a "core point" if there are at least MinPts other points in its ε-neighborhood (N_ϵ(p)). A point p is directly density-reachable from another point q, if q is a core point and p is in the ε-neighborhood of q i.e. $p \epsilon N_\epsilon$(q). Point p is density-reachable from another point q if there is a chain of points $p_1,\ldots,p_n$ where p_{i+1} is directly density reachable from p_i such that $p_1 = q$ and $p_n = p$.

By setting the parameter ε to the maximum distance between any points in the dataset, each point in the dataset will be the core point. Thus, the modified OPTICS algorithm only requires the parameter MinPts. In our case, MinPts was set to 40 for 10 s long signal epochs. OPTICS detects varying-density clusters (i.e., clusters with different dispersions) in the dataset, such that most similar points (AcSs) become neighbors. In the other words, similar points are grouped together into hierarchical clusters. In this procedure, the first AcS is selected as the current point. Its reachability distance (RD) is set to INF. Its nearest neighbor in respect to RD is then set as the current point. Its RD is calculated as the smallest density reachable distance from the previous AcS (Marateb et al., 2011b). This process is repeated until all the points are processed in this way. Using this method, the entire feature space is transferred to a two-dimensional plot in which x-axis corresponds to the ordered points and y-axis depicts the RD of the ordered points. Each RD valley in such plot is a possible cluster, with the point corresponding to the minimum RD in a valley as the corresponding cluster representative. In our study, the possible valleys were automatically extracted using the adaptive method proposed by Marateb et al. (2011b, 2012).

Modified CKC

Denote by $\mathbf{X}(n,m)$ the (n,m)-th element of the signal matrix $\mathbf{X}$ (the m-th sample of channel n) and by $\mathbf{X}(m)$ the m-th column of matrix $\mathbf{X}$. The rows in matrix X are EMG channels and their delayed repetitions (Holobar and Zazula, 2004; Thomas et al., 2006). In the CKC algorithm (Holobar and Zazula, 2004, 2007), the spike train $\check{\boldsymbol{t}}_j$ of the *j-th* MU is estimated as

$$\check{\boldsymbol{t}}_j = \mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_j}{}' \mathbf{R}_{\mathbf{XX}}^{-1} \mathbf{X} \tag{4}$$

where $\mathbf{R}_{\mathbf{XX}}^{-1}$ is the inverse of autocorrelation matrix of $\mathbf{X}$, $\mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_j}$ is the cross-correlation vector of $\mathbf{X}$ and $\check{\boldsymbol{t}}_j$ and the operator $'$ denotes the matrix transpose. The cross-correlation vector $\mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_j}$ is unknown, but can be iteratively estimated by gCKC (Holobar and Zazula, 2007, 2008):

$$\mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_{j,g}} = \mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_{j,g-1}} + \eta(g)\left(f\left(\check{\boldsymbol{t}}_{j,g}\right)\mathbf{X}'\right)' \tag{5}$$

where $\check{\boldsymbol{t}}_{j,g}$ is the j-th MU spike train, estimated in the *g-th* iteration, η is the step-size, $f(x) = \alpha(x)\log\left(1+x^2\right)$ and $\alpha(x)$ is the attenuation coefficient introduced herein to improve the FP and FN errors in gCKC. The initial value of α is set to one but is then determined for each MU firing in each gCKC iteration. Namely, we observed that sometimes gCKC iterations erroneously amplify signal artifacts. Thus, after each iteration, we search for a spike detection threshold that yields the maximum Pulse-to-Noise Ratio (PNR) (Holobar et al., 2014).

$$PNR(j,k) = 10 \times log\left(\frac{E(\check{\boldsymbol{t}}_{j,g})|_{\check{\boldsymbol{t}}_{j,g} \geq thr}}{E(\check{\boldsymbol{t}}_{j,g})|_{\check{\boldsymbol{t}}_{j,g} < thr}}\right) \tag{6}$$

where *thr* is the pulse detection threshold and PNR(j,k) is the Pulse-to-noise-ratio of the j-th MU in the k-th iteration. This is in fact a one-dimensional optimization problem and we efficiently solved it using a greedy search algorithm (Feo and Resende, 1995). The fitness function is the PNR while the optimization variable is the threshold *thr*. When this threshold is estimated, its 95% CI (Confidence Interval) is estimated and spikes within *thr* ± 95% CI of *thr* are detected (marginal spikes). The cluster representative (CR) of the spike samples surpassing the upper 95% CI threshold is formed and then if the corresponding sEMG AcS of a marginal spike has a high correlation with AcS of a CR, parameter α is set to one, otherwise it is set to 0.9. In this way, template-based methods are also used in the proposed modification of the previously introduced gCKC iterations (in addition to initial point estimation) (Holobar and Zazula, 2008). This operation is, in principle, equivalent to non-linear soft thresholding (Krzakala et al., 2016).

Kalman Filter

In our method, Equation (4) is substituted by expectation maximization Kalman filter (EM-Kalman filter), proposed by Bensaid et al. (2010). For this purpose, the gCKC is defined as the following state-space model joint to the EM-Kalman filter:

$$\begin{aligned} \check{x}_{k|k-1} &= \mathbf{F}_k x_{k-1|k-1} + w_k \\ &= \begin{bmatrix} \mathbf{I}_N & \eta(g)\left(\mathbf{X}f'(t_{j,g,k-1})\right) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\mathbf{X}t_{j,g,k-1}} \\ 1 \end{bmatrix} \\ &+ w_k \; \check{\mathbf{t}}_{j,g,k}(m) = \check{\mathbf{x}}_{k|\,k-1}\mathbf{H}(m) \end{aligned} \tag{7}$$

where $\check{\mathbf{x}}_{k|k}$ and k are estimated state and current Kalman filter iteration, respectively, g is related to the g-th iteration of the gCKC algorithm and $\mathbf{w}_k = \mathbf{C}_\mathbf{w}\mathbf{e}_k = \mathbf{C}_\mathbf{w}[\mathbf{e}_k(1), \mathbf{e}_k(2),\ldots,\mathbf{e}_k(N),0]'$ is the normally distributed noise vector of size $(N+1) \times 1$ with variance σ_k^2. $\mathbf{C}_\mathbf{w}$ is a diagonal $(N+1) \times (N+1)$ covariance matrix of noise and $\mathbf{e}_k = \mathbf{V}_k{}'\check{\mathbf{x}}_{k|k-1}$. According to the Bensaid et al. (2010), $\mathbf{V}_k$ is covariance matrix of $\check{\mathbf{x}}_{k|k-1}$ and due to the whitening process, $\mathbf{V}_k$ could be considered identity matrix. Thus, $\mathbf{w}_k = \mathbf{C}_\mathbf{w}\check{\mathbf{x}}_{k|k-1}$.

$\mathbf{F}_k = \begin{bmatrix} \mathbf{I}_N & \eta(g)\left(\mathbf{X}f'(\check{\boldsymbol{t}}_{j,g,k})\right) \\ 0 & 1 \end{bmatrix}$ is the $(N+1)\times(N+1)$ state (or system) matrix with $\mathbf{I}_N$ denoting the identity matrix and N being the number of channels. $\check{\mathbf{x}}_{k|k}$ and $\mathbf{H}(m)$ are defined by Equations (8) and (9):

$$\check{\mathbf{x}}_{k|k} = \begin{bmatrix} \mathbf{R}_{\mathbf{X}\check{\boldsymbol{t}}_{j,g,k}} \\ 1 \end{bmatrix} \tag{8}$$

$$\mathbf{H}(m) = \begin{bmatrix} \mathbf{X}(m) \\ 0 \end{bmatrix} \tag{9}$$

In this regard, a priori error covariance $\mathbf{P}$ and σ_k^2 could be defined by Equations (10) and (11), respectively.

$$\mathbf{P}_{k|k\ -1}\ =\ \mathbf{F}_k\mathbf{P}_{k-1|k-1}\ \mathbf{F}_k^{'}\ +\mathbf{C}_w\mathbf{C}_w^{'} \tag{10}$$

$$\sigma_k^2\ =\ \frac{1}{M-1}\sum_{m=1}^{M}\left[\mathbf{H}'(m)\left(\check{\mathbf{x}}_{k|k}\ \check{\mathbf{x}}_{k|k}^{'}\ +\mathbf{P}_{k|k}\ \right)\mathbf{H}(m)-\check{\boldsymbol{t}}_{j,g,k}\,(m)\right] \tag{11}$$

where $\mathbf{H}\ =\ \left[\mathbf{H}\,(1)\,,\mathbf{H}\,(2)\,,\mathbf{H}(3)\ldots\mathbf{H}(M)\right]$ and its size is $(N+1)\times M$. Kalman gain ($\mathbf{K}$) is defined with Equation (12).

$$\mathbf{K}_k(m)=\mathbf{P}_{k|k-1}\ \mathbf{H}(m)\left(\mathbf{H}^{'}\,(m)\,\mathbf{P}_{k|k-1}\ \mathbf{H}\,(m)+\sigma_k^2\right)^{-1} \tag{12}$$

We then update gCKC algorithm and error covariance as:

$$\check{\mathbf{x}}_{k|k}\ \ =\check{\mathbf{x}}_{k|k-1}\ \ +\mathbf{K}_k\left(\check{\boldsymbol{t}}_{j,g,k}-\check{\boldsymbol{t}}_{j,g,k-1}\right) \tag{13}$$

$$\mathbf{P}_{k|k}\ \ =\mathbf{P}_{k|k-1}\ \ -\mathbf{K}_k\,\mathbf{H}^{'}\mathbf{P}_{k|k-1} \tag{14}$$

When considering the state-space system matrix, h_m is the output matrix in the state-space model and $\mathbf{K}_k\ =\ [\mathbf{K}_k(1),\mathbf{K}_k(2),\ldots,\mathbf{K}_k(M)]$ of size *(N+1)* $\times$ M is the Kalman gain in the k-th iteration. $\check{\boldsymbol{t}}_{j,g,k}$ in the first gCKC state is generated with the fixed point algorithm and one iteration of gCKC (Holobar and Zazula, 2007) and $\mathbf{P}_{k-1|k--1}$ is initialized as $\mathbf{I}_{N+1}$ i.e., the identity matrix. The estimation of the observation noise power σ_k^2 is achieved by maximizing the Log-Likelihood function $\log(P(\check{\boldsymbol{t}}_j\,|\check{\mathbf{x}}\ \ ,\sigma_k^2))$ relative to σ_k^2 (Bensaid et al., 2010).

In a nutshell, we added additional noise factor in gCKC formulation and estimated it during each iteration since, practically, colored noise exists instead of white noise and gCKC algorithm, which works by correlation procedure, enhances colored noise (Li, 1990). This approach is used in other fields for finding noise parameters (Schwartz et al., 2015, 2016; Ge and Kerrigan, 2017) and also neuro spike decoding (Xue et al., 2017). In this light, EM-Kalman filter could prevent enhancing noise in $\check{\mathbf{x}}_{k|k}$ in each iteration of gCKC and decrease the FP errors. This hypothesis will be, at least partially, confirmed by the results presented in section Results as the newly presented method detects substantially more MUs than CKC method.

Motor Unit Identification

The non-linear soft thresholding method mentioned in section Modified CKC, was applied to the reconstructed MU spike trains to identify MU firing times. Moreover, extracted MUs were monitored in terms of regularity of the firing times and their distinguishability from the background noise. MUs with the PNR <20 dB were excluded from further validations (Holobar et al., 2014). The following firing time parameters were extracted and used for MU exclusion: the number of inconsistent firings times n_I(number of pulses that are fired more than once in every 40 ms), the detection probability p_d(the probability of MU inter-spike interval having normal distribution as assessed by the method proposed by McGill Kevin, 1984) and the mean discharge rate (MDR). A MU was excluded when n_I was higher than 50, or p_d was lower than 50% or MDR was higher than 35 Hz. The underlying assumptions were the following: during constant force isometric contractions, MU firing rates lie between 5 and 25 Hz. Moreover, the inter-spike intervals have a unimodal distribution which is approximately Gaussian (Clamann, 1969; Andreassen and Rosenfalck, 1980). The pseudocode of the proposed HDsEMG decomposition algorithm is available as the **Supplementary Material**.

TABLE 1 | Decomposition validation parameters.

Parameter	Definition
RoA	$\frac{TP}{TP+FN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
DI_{ki}	$DI_{ki}=\frac{\min\{\|m_{ki}\|,\|m_{ki}-m_{k^*i}\|\}}{V_i^{RMS}}$
SIR(i)	$\left(1-\frac{E\left[\left(x_i(n)-\sum_j z_{ij}(n)\right)^2\right]}{E\left[x_i^2(n)\right]}\right)\times 100$

RoA, Rate of agreement; SIR(i), signal-to-interference ratio of the i-th channel; $DI_{k,I}$ is decomposability index and defines as the average MUAP shape of the k-th motor unit in the i-th channel, $\mathbf{m_{k^*i}}$ is the MUAP most similar to $\mathbf{m_{ki}}$, among the other MUAPs in the same channel, $\mathbf{V_i^{RMS}}$ is the RMS value of the channel i, and $\|.\|$ stands for the Euclidean norm; TP (True Positive): the number of MU spikes correctly identified, FN (False negative): the number of missed MU spikes; FP (False Positive): the number incorrectly identified MU spikes; $x_i\,(n)$ denotes the time samples of the i-th surface EMG channel and $\mathbf{z_{ij}}$ stands for the time samples of the action potential train of the j-th motor unit reconstructed from the i-th EMG channel.

Performance Assessment

The identified MU firings were compared with the simulated firings. When at least 30% of the firings were time-locked to within $\pm$ 0.5 ms, a MUs was considered to be identified by our algorithm (Marateb et al., 2011a). For each identified MU, the accuracy of our new decomposition algorithm was assessed using the parameters of signal detection theory (e.g., TP (True Positive; correct firings), FN (False Negative; missed firings), and FP (False Positive; erroneous firings)). TP was the number of firings matching the simulated firings within $\pm$ 0.5 ms. FP was the number of firings not matching any simulated firing to within $\pm$ 0.5 ms. FN was the number of firings of the simulated MU that did not match any firings of the identified MU. Then, the performance indices Sensitivity (Se), Precision (Pr), and the Rate of agreement (RoA) were calculated (Holobar et al., 2010, 2014) (**Table 1**).

Moreover, the signal-to-interference ratio (SIR) was reported as the overall quality of the sEMG decomposition (**Table 1**). In fact, SIR estimates the percentage of the variance of the energy of the single-differential sEMG signals explained by the decomposition (Holobar et al., 2010). Also, the overall decomposability of a MU in the entire recording signals (cDI) was measured by the sum of norm of individual decomposability indices (DIs) (Holobar et al., 2010) (**Table 1**), calculated over all the EMG channels, normalized by the number of channels. The

program was run on an Intel Core i7-8700 3.2 GHz CPU with 32 GB of RAM.

Statistical Analysis

Continuous variables were reported as mean ± std. The level of statistical significance was set to $P = 0.05$. The normality of the variables was assessed using the Shapiro-Wilk test. We used generalized estimating equation (GEE) method (Hardin, 2005) to model factors associated with the repeated responses (RoA and the running time of the algorithm). The paired *t*-test was used to identify if there is a significant difference between the estimated MDR and CoV in the decomposed signals, compared with those of the simulated MU firings. The Pearson correlation coefficient (*r*) was used to assess the association between two normal variables. The statistical analysis was performed using SPSS version 16 (SPSS for Windows, Released 2007, Chicago, IL, USA, SPSS Inc.).

RESULTS

The average number of MUs, identified per contraction and SNR and also the accuracy of the decompositions of the assessed MUs are listed in **Table 2**. The accuracy of the clustering algorithm is also shown in this table as the average number of identified clusters. The cluster's representatives as well as their 20 nearby spikes in the RD plots, were compared with the simulated MU firings. When at least 75% of the firings in a cluster was time locked within 0.5 ms with firings of a simulated MU firings, the cluster was marked as correctly identified. Therefore, the reported number of identified clusters shows how many initial points were related to correct MU clusters. For comparison, the number of identified MUs and their decomposition accuracy of the gCKC (Holobar and Zazula, 2007) are also shown. Overall, our algorithm identified 16.41 ± 4.18 MUs with RoA of 84.00 ± 0.06%. The average number of modified gCKC iterations required to identify individual MU was 54.11 ± 3.06.

The firing characteristics of the MUs, identified by the decomposition program and simulated MUs were very similar (**Table 3**), regardless the tested level of excitation and SNR.

The sensitivity, RoA and precision of the proposed algorithm vs. PNR for different excitation levels and SNR values are depicted in **Figure 2**. Also, the distribution of the precision of the proposed decomposition algorithm compared with that of the gCKC for 30 dB SNR at 30% excitation and 20 dB SNR at 50% excitation is shown in **Figure 3**.

Figure 4 shows an example of the MUAP trains identified during seven seconds of a contraction at 10% MVC and 30 dB SNR. In this case, 14 MUs were identified with the average RoA of 89.00 ± 0.89 (%). **Figure 5** shows the spatial distribution of Single-differential sEMG MUAP of MU 14 form **Figure 4** as estimated by the spike-trigger averaging of the HDsEMG signals over the estimated MU firing times.

TABLE 2 | Decomposition accuracy.

SNR (dB)	MVC (%)	Average number of identified clusters	RoA	Sensitivity	Precision	Number of Identified MUs	SIR (%)	The Average Number of Iterations	Running time (sec) With the Kalman filter	Running time (sec) Without the Kalman filter	RoA (gCKC)	Sensitivity (gCKC)	Precision (gCKC)	Number of Identified MUs (gCKC)
10	10	30	0.75 ± 0.17	0.80 ± 0.14	0.92 ± 0.09	11.60 ± 1.67	13.38 ± 3.76	51.2 ± 3.3	1739 ± 23 [704 ± 11]	1817 ± 7 [791 ± 5]	0.89 ± 0.12	0.95 ± 0.04	0.93 ± 0.09	9.6 ± 1.5
10	30	28	0.76 ± 0.15	0.80 ± 0.13	0.93 ± 0.08	9.62 ± 1.81	14.04 ± 5.06	52.1 ± 2.1	1778 ± 20 [738 ± 6]	1862 ± 13 [811 ± 1]	0.89 ± 0.08	0.94 ± 0.04	0.93 ± 0.05	7.6 ± 2.7
10	50	16	0.77 ± 0.16	0.81 ± 0.13	0.94 ± 0.08	11.00 ± 2.54	12.28 ± 5.49	49.3 ± 2.5	1846 ± 60 [740 ± 13]	1919 ± 22 [830 ± 10]	0.87 ± 0.09	0.93 ± 0.05	0.92 ± 0.06	7.3 ± 1.5
20	10	27	0.89 ± 0.01	0.92 ± 0.02	0.96 ± 0.01	17.20 ± 2.62	15.05 ± 1.95	53.1 ± 3.7	1390 ± 14 [613 ± 2]	1647 ± 11 [680 ± 1]	0.94 ± 0.05	0.97 ± 0.02	0.96 ± 0.03	14.00 ± 1.87
20	30	28	0.88 ± 0.01	0.92 ± 0.01	0.95 ± 0.01	16.51 ± 2.20	15.11 ± 1.10	53.4 ± 3.3	1423 ± 15 [661 ± 23]	1687 ± 6 [704 ± 10]	0.91 ± 0.09	0.95 ± 0.03	0.95 ± 0.05	11.40 ± 3.58
20	50	20	0.88 ± 0.01	0.92 ± 0.01	0.95 ± 0.02	17.01 ± 2.71	15.42 ± 2.37	56.1 ± 2.5	1519 ± 11 [679 ± 18]	1721 ± 12 [718 ± 8]	0.90 ± 0.08	0.95 ± 0.04	0.94 ± 0.05	8.20 ± 1.92
30	10	28	0.90 ± 0.09	0.92 ± 0.07	0.98 ± 0.02	22.00 ± 3.01	18.12 ± 3.22	55.6 ± 1.3	1101 ± 29 [545 ± 47]	1539 ± 36 [670 ± 57]	0.94 ± 0.06	0.96 ± 0.03	0.96 ± 0.04	19.00 ± 0.71
30	30	29	0.90 ± 0.01	0.92 ± 0.02	0.98 ± 0.01	23.00 ± 1.10	17.39 ± 2.91	57.8 ± 2.1	1230 ± 28 [640 ± 18]	1652 ± 40 [703 ± 25]	0.92 ± 0.07	0.96 ± 0.04	0.94 ± 0.05	17.40 ± 2.51
30	50	21	0.90 ± 0.01	0.92 ± 0.01	0.98 ± 0.01	19.80 ± 1.00	15.01 ± 3.38	58.4 ± 2.4	1301 ± 28 [660 ± 27]	1822 ± 16 [718 ± 13]	0.91 ± 0.07	0.95 ± 0.04	0.94 ± 0.05	14.20 ± 3.11

RoA, Rate of agreement; SIR: average signal-to-interference ratio. In gCKC, the number of iterations was fixed to 100. The total number of simulated MUs was 500. The program was run on an Intel Core i7-8700 3.2 GHz CPU with 32 GB of RAM. The reported running time was measured when analyzing 16 s of HDsEMG signals. For clarity reasons, the running time required to analyze 8 s long signals is also provided in squared brackets.

DISCUSSION

Neural decoding is used to identify how the electrical activity of neurons generates responses in the brain (Jacobs et al., 2009).

TABLE 3 | The firing statistics of the decomposed MUs.

SNR (dB)	MVC (%)	MDR (Hz)	MD̆R (Hz)	CoV	C̆oV	MDR bias (Hz)	CoV bias	cDI (%)
10	10	13.17 ± 2.15	13.20 ± 2.19	0.14 ± 0.00	0.14 ± 0.00	0.03 ± 0.08	0.01 ± 0.00	15.40 ± 4.61
10	30	26.40 ± 7.27	26.74 ± 6.83	0.15 ± 0.00	0.14 ± 0.00	0.34 ± 0.24	0.01 ± 0.00	16.44 ± 8.37
10	50	30.19 ± 7.02	31.64 ± 6.02	0.15 ± 0.00	0.14 ± 0.00	1.45 ± 0.26	0.01 ± 0.00	18.19 ± 8.91
20	10	13.12 ± 2.10	13.20 ± 2.19	0.13 ± 0.02	0.14 ± 0.00	0.08 ± 0.08	0.01 ± 0.00	20.08 ± 4.44
20	30	26.77 ± 6.38	26.74 ± 6.83	0.13 ± 0.01	0.14 ± 0.00	0.03 ± 0.08	0.01 ± 0.00	19.87 ± 4.10
20	50	32.02 ± 4.26	31.64 ± 6.02	0.13 ± 0.02	0.14 ± 0.00	0.38 ± 0.14	0.01 ± 0.00	17.64 ± 4.53
30	10	13.18 ± 2.10	13.20 ± 2.19	0.13 ± 0.03	0.14 ± 0.00	0.02 ± 0.03	0.01 ± 0.00	22.63 ± 4.55
30	30	26.75 ± 6.33	26.74 ± 6.83	0.13 ± 0.04	0.14 ± 0.00	0.01 ± 0.01	0.01 ± 0.01	22.61 ± 4.80
30	50	31.88 ± 4.01	31.64 ± 6.02	0.13 ± 0.02	0.14 ± 0.00	0.24 ± 0.05	0.01 ± 0.00	21.89 ± 3.92

MDR, Simulated mean discharge rates; MD̆R: the MDR estimated by the proposed algorithm; CoV, Simulated coefficient of variability for the firing rate; C̆oV, CoV estimated by the proposed algorithm; SIR: average signal-to-interference ratio; cDI: cumulative Decomposability Index. MDR and CoV biases were calculated as the absolute mean value of the pair-wise difference between estimated minus gold standard MDR and CoV, respectively.

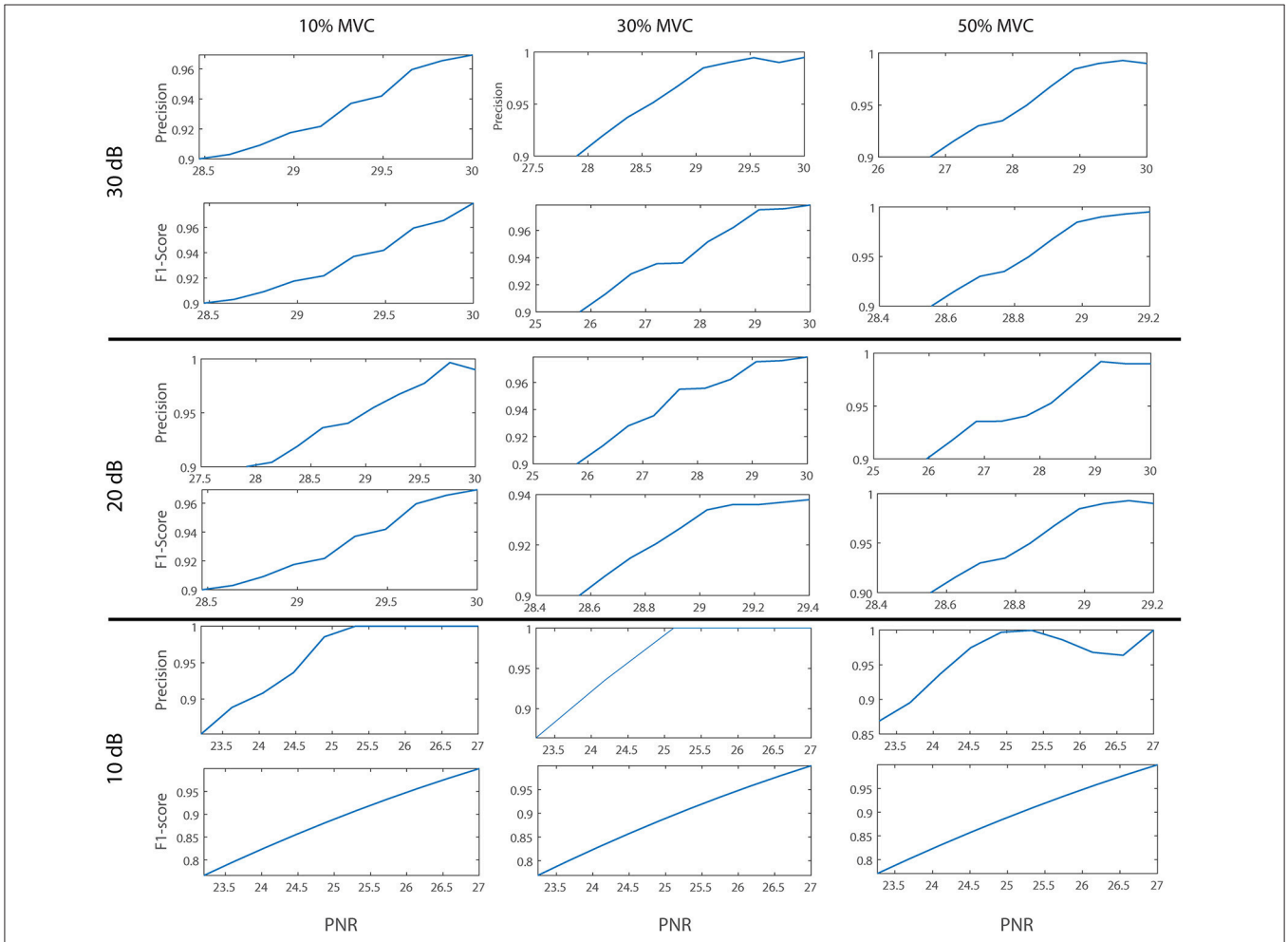


FIGURE 2 | The sensitivity, precision, and Rate of Agreement (RoA) of the proposed algorithm vs. PNR (in dB). Representative plot is provided for each SNR level (10, 20, and 30 dB) at each simulated level of muscle excitation (10, 30, and 50% MVC).

In the case of motor system, this can be performed either at the level of motor nerves or motor units. Although, the methods are very similar, the former is called "spike sorting" in computational neuroscience while the latter is known as "EMG decomposition," or "decoding of the neural drive to the muscles" (Rey et al., 2015; Webster et al., 2016; Karimimehr et al., 2017). Such a

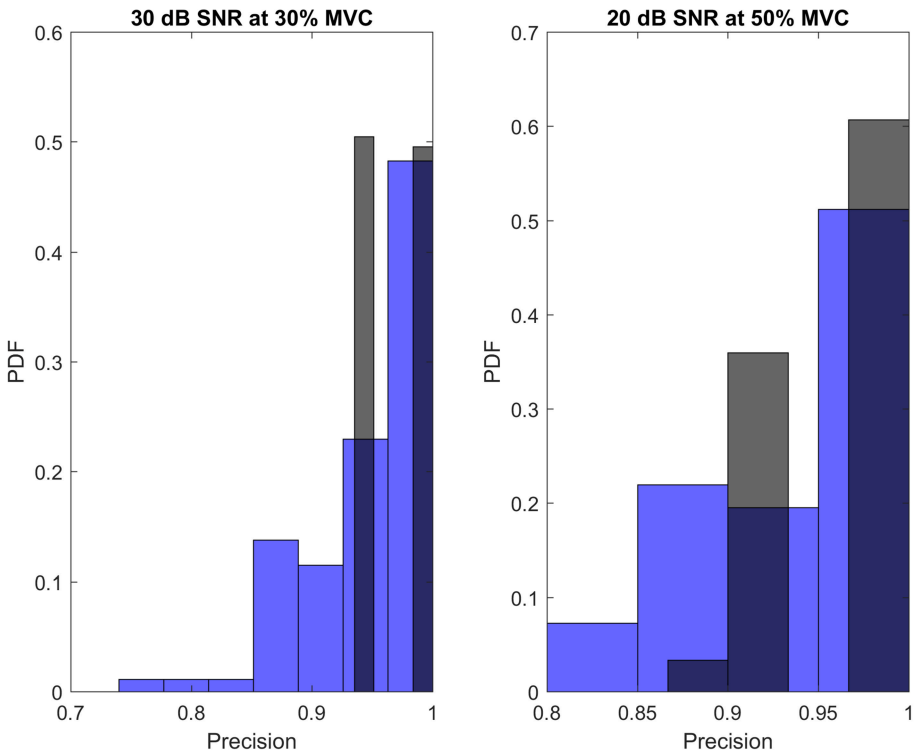


FIGURE 3 | The histogram of the precision of the proposed decomposition algorithm (black), compared to the one form gCKC (blue) for 30 dB SNR at 30% excitation level (left), and 20 dB SNR at 50% excitation level (right).

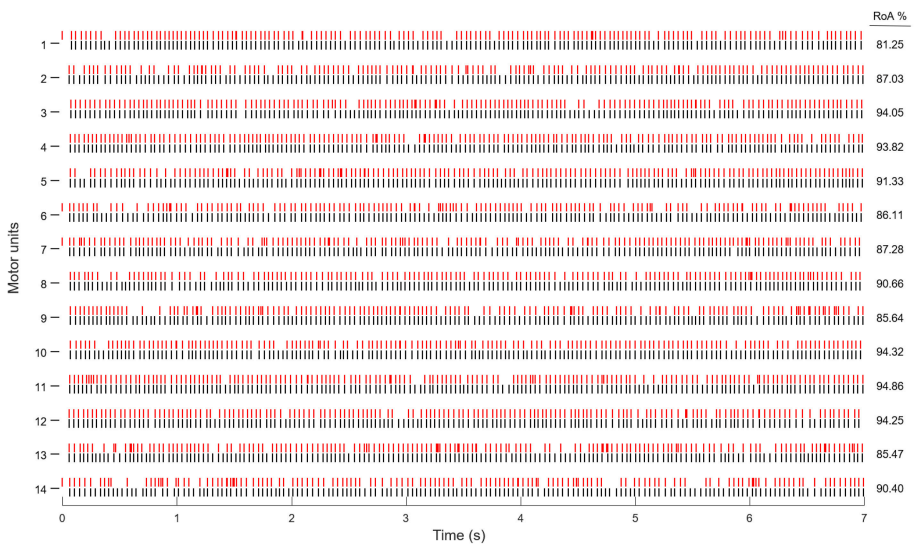


FIGURE 4 | MU spike trains, identified from the simulated sEMG signal with 30 dB SNR and 10% excitation level (red), and the simulated firings (black). Each vertical line indicates one MU firing.

decomposition algorithm could be used in variety of applications, including prosthesis control (Farina et al., 2014a) or robot-assisted neurorehabilitation (Savc et al., 2018). The structure of the spike sorting algorithms is, in principle, similar to that of iEMG decomposition methods, where the recording electrodes are close to the MU fibers. However, this is not the case for sEMG decomposition, where electrodes are located over the skin at a relatively large distance from the muscle fibers.

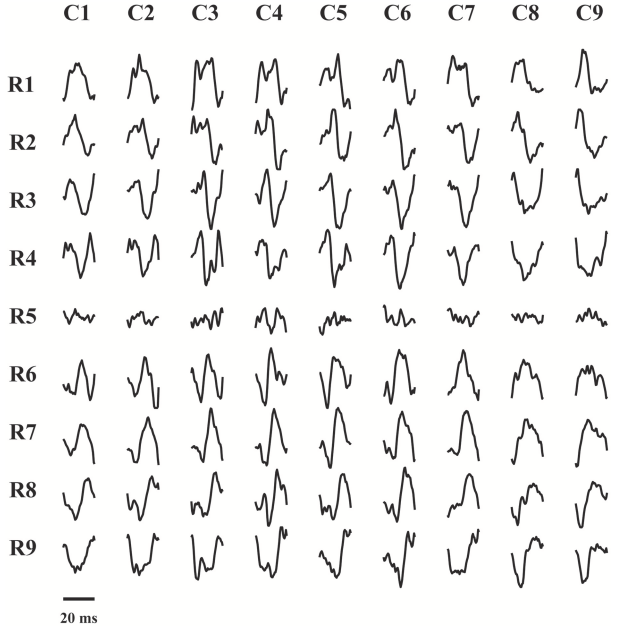


FIGURE 5 | Single-differential sEMG MUAP waveforms of MUs from different sEMG channels. The corresponding MU was identified with accuracy >90%.

sEMG decomposition is considered a very difficult task when low selectivity of traditional sEMG electrodes, low-pass filtering of MUAPs, their overlapping, and shape similarities are considered (Farina et al., 2004). In this study, we proposed a new algorithm that combines the advantages of two different approaches to sEMG decomposition, namely the template-based and the BSS algorithms. Classification procedures are critical steps of the template-based algorithms. sEMG MUAPs are rather similar in shape and their classification is challenging (De Luca et al., 2006; Nawab et al., 2010). BSS algorithms, on the other hand, combines MU activities in the space of MU spikes in order to identify all the MU firings (Holobar and Farina, 2014). Many BSS algorithms cannot guarantee the identification of the same set of sources in multiple runs, because they rely (at least partially) on stochastic algorithms, for example on random initialization of vector $\mathbf{R}_{\mathbf{X}\check{\imath}_j}$ in Equation (5).

Our algorithm, is fully deterministic since it identifies the initial points using a new clustering algorithm (**Figure 1**) rather than random initialization. Other CKC-based algorithms, use the time samples of the peaks of the sEMG signal as the possible starting point of the gradient-based optimization. However, sEMG contains highly overlapped MUAPs, and selected peaks of the signal usually correspond to the overlapped MU activity. This overlapped MU activity is then separated in several iterations of gCKC algorithm (Equation 5). On the other hand, if the initial $\mathbf{R}_{\mathbf{X}\check{\imath}_j}$ estimate is not related to any MU firings, but rather to the movement artifact or line interference, the CKC algorithm cannot amplify the MU spikes and the noise is amplified instead. Thus, being fully deterministic in our sense means that we know where a MU firing exists and we start the CKC gradient optimization from the initial point with a much higher MU identification potential than gCKC algorithm.

Similar aggregative clustering methods have been proposed in the past for sEMG decomposition (Ning et al., 2015). However, aggregative clustering methods, such as K-means or Fuzzy C-means do not have satisfactory "rerun" stability due to poor reproducibility based on random initialization (Jayaram and Klawonn, 2013). Furthermore, they may get stuck in local minimum owing to a two-stage iterative algorithm used to minimize the sum of point-to-center distances over all clusters. Finally, they have high computational complexity due to the need of running the algorithm several times with the pre-defined number of clusters (Duda et al., 2001). Therefore, we used a density-based OPTICS algorithm, which is a single-pass algorithm designed particularly for high-dimensional data (Ankerst et al., 1999; Daszykowski et al., 2002). This algorithm can automatically identify the optimal number of clusters when a proper automatic valley detection algorithm is used (Marateb et al., 2011b). Moreover, it adaptively identifies clusters of different sizes and variable dispersions and densities (Loh and Park, 2014).

In sEMG, there are many active MUs. This causes the overlap between MUAPs and thus the number of isolated segments reduces significantly, especially at higher contraction levels. Although each recording sEMG channel provides a different perspective of the MUAPs, the performance of the initial point detection generally drops at higher contraction levels (**Table 2**; The number of identified clusters variable). This is one of the limitations of the proposed algorithm. Using different spatial filters could improve the performance of this step of the algorithm which will be investigated in our future work.

Moreover, several cluster representatives identified by the OPTICS algorithm may be related to the same MU. This is why the average number of clusters identified by the OPTICS algorithm is higher than the number of identified MUs by the entire algorithm (**Table 2**). Since the MUAP shapes are very similar over the skin and largely submerged into the physiological noise, it is not possible to merge them unless their firing times are identified by the CKC gradient loops. On the other hand, due to the high overlapping between different MUAPs and the similarity between the shapes of different MUAPs in the sEMG signal, it is not optimal to use traditional classifiers (e.g., the minimum-distance classifier) based on the clusters obtained by the OPTICS algorithm. Thus, the combination of the OPTICS and gCKC algorithm is required, as proposed in our manuscript.

In our study, results of clustering were used as initial points in a modified g-CKC algorithm (**Figure 1**). In fact, our algorithm is a Guided-Source Separation (GSS) method with the following modifications introduced into the gCKC method:

(1) An adaptive soft-thresholding method was designed based on stochastic optimization, correlation analysis and estimation theory on the spike times as to reduce the FP (by suppressing noise spikes) and FN (by amplifying valid marginal spikes) error rates;

(2) The Kalman filtering was used to improve the convergence rate of the algorithm and to reduce the noise enhancement in gCKC iterations.

The performance of our algorithm was compared with that of the previously published gCKC algorithm (**Table 3**) since gCKC was shown to be accurate with different muscle architectures (Holobar et al., 2010; Marateb et al., 2011a; Webster et al., 2016) and different pathological conditions including essential tremor (Gallego et al., 2015a), Parkinson's disease (Holobar et al., 2012), diabetes (Watanabe et al., 2013), after stroke (Li et al., 2015), targeted muscle reinnervation (Farina et al., 2014c), and cleft lip surgery (Radeke et al., 2014). The previously introduced PNR, an efficient objective signal-based metric of gCKC decomposition accuracy (Holobar et al., 2014), was also reported for our decomposition results (**Table 2**, **Figure 2**). Although the sensitivity and RoA of the proposed algorithm is less than that of the gCKC algorithm, we identified more MUs at comparable Precision (**Table 2**; The average precision of 0.96 ± 0.05 compared with 0.95 ± 0.05 for gCKC). In fact, as mentioned in section Kalman filter, the Kalman filter decreases the FP errors and such errors have a direct effect on the Precision (**Table 1**). Therefore, FP error reduction goal was achieved by Kalman filter and the adaptive soft thresholding method. In our future work, we intend to focus on the reduction of FN errors. Moreover, we tested our algorithm with and without the Kalman filter. With the Kalman filter, the number of iterations was reduced in average down to 48% compared to the case without the Kalman filter. Moreover, when the Kalman filter was added, two more MUs were correctly identified, on average, whereas the running time of our algorithm was significantly reduced ($P < 0.001$; GEE).

The results of our statistical analysis showed that RoA was significantly associated with the SNR parameter i.e., the signal quality ($P < 0.001$; GEE) but not with the muscle excitation level i.e., signal complexity ($P > 0.05$; GEE). RoA was significantly and highly correlated with cDI at 10 dB and 20 dB SNR ($P < 0.001$; $r > 0.7$), but not at 30 dB SNR. This suggests that the proposed algorithm would work even if MUAPs were relatively similar in shapes when the quality of the signal was high. Moreover, there was no significant differences between the MDR and CoV values estimated from the decomposed sEMG signals and the corresponding simulated MU firings ($P > 0.05$; paired t-test). Thus, the decoded neural information can be considered reliable. Although measuring the MU identification accuracy indirectly, the PNR parameter was highly correlated with all the performance indices, namely with RoA, Sensitivity, and Precision (**Figure 2**; $P < 0.001$, $r > 0.8$). Thus, MUs with high PNR values can be considered as accurately identified. This is in agreement with the results in Holobar et al. (2014); Martinez-Valdes et al. (2016), and Watanabe et al. (2016).

In our study, simulated data was used for validation. Although, the used volume conductor model has been widely discussed in the literature (e.g., in McGill, 2004; Zazula and Holobar, 2005; Holobar et al., 2009; Zalewska, 2009), it could not completely resemble the experimental sEMG signals (De Luca and Nawab, 2011). Thus, further tests on experimental signals from different skeletal muscles in different experimental conditions (i.e., different contraction levels, force profiles, muscle geometries etc.) are required before the reported results can be generalized. This is a common challenge to all the reported HDsEMG decomposition algorithms as the results from one experimental setup cannot easily be generalized to all experimental conditions. In this study, we introduced the novel methodological steps and tested their efficiency against previously introduced CKC method that has been extensively validated over the past decade and applied to many (but not all) experimental conditions. Further generalization of the results reported herein exceeds the scope of this study and is left for the future work.

Indeed, the validation of the decomposition on experimental sEMG signals is, in principle, controversial. Using cross-checking, also known as two-source method, between the decomposition of concurrently recorded iEMG and sEMG signals, could only assess the accuracy of common MUs. Partitioning the sEMG channels into two groups and cross-checking them against each other, on the other hand, is also problematic as the information they convey is highly correlated (Webster et al., 2016). However, the objective measure PNR could be reported and used as the MU reliability index by the decomposition program (Holobar et al., 2014). Moreover, a novel validation approach proposed by Chen et al. could serve as a supplement to the conventional two-source methods (Chen et al., 2018b).

Finally, the proposed GSS algorithm has relatively low sensitivity to MUs, although it identifies more MUs than gCKC (**Table 2**). It identifies only a small portion of active MUs. This is a common limitation to all EMG decomposition approached introduced by now. The proposed algorithm is also limited to off-line analysis and cannot easily be converted to the online neural decoding algorithms (Glaser et al., 2007; Karimimehr et al., 2017). Its online implementation requires improvements in the rate of convergence of gradient-based optimization (Equation 5). For this purpose, the condition number of the Hessian matrix could be analyzed and proper diagonal scaling could be used to reduce the number of iterations (Beck, 2014), but this additional step has not been thoroughly investigated yet. Moreover, detection of firing-time inconsistency could be added to the OPTICS clustering algorithm in order to increase the number of MUs identified by the entire algorithm. We will address these methodological improvements in our future work.

In conclusion, we proposed a new framework for sEMG decomposition. The proposed algorithm is promising and a new offline tool in clinical neurophysiology.

AUTHOR CONTRIBUTIONS

MRM, HM, SK, and AH participated in conceptualization, data curation, and validation. MRM, HM, SK, MAM, JK, and AH participated in formal analysis, investigation, and methodology. HM, MAM, and AH participated in project administration and resources. MRM, HM participated in software, and visualization. HM and AH supervised the project. MAM and AH acquired

funding. MRM, HM, JK, and AH contributed to writing the original draft and SK, MAM revised the manuscript. All authors read and approved the final version of the manuscript and agreed for all aspects of the work.

FUNDING

This work was supported by the Spanish Ministry of Economy and Competitiveness- Spain project DPI2017-83989-R, by the University of Isfahan and by the Slovenian Research Agency (projects J2-7357 and L7-9421 and Programme funding P2-0041).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00014/full#supplementary-material

REFERENCES

Andreassen, S., and Rosenfalck, A. (1980). Regulation of the firing pattern of single motor units. *J. Neurol. Neurosurg. Psychiatr.* 43, 897–906. doi: 10.1136/jnnp.43.10.897

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). "OPTICS: ordering points to identify the clustering structure", in *ACM Sigmod Record: ACM*, eds A. Deshpandeh, Z. Ives, and V. Raman (New York, NY: Association for Computing Machinery), 49–60. doi: 10.1145/304181.304187

Artoni, F., Fanciullacci, C., Bertolucci, F., Panarese, A., Makeig, S., Micera, S., et al. (2017). Unidirectional brain to muscle connectivity reveals motor cortex control of leg muscles during stereotyped walking. *Neuroimage* 159, 403–416. doi: 10.1016/j.neuroimage.2017.07.013

Balasubramanian, K., and Obeid, I. (2011). Fuzzy logic-based spike sorting system. *J. Neurosci. Methods* 198, 125–134. doi: 10.1016/j.jneumeth.2011.03.016

Beck, A (2014). *Introduction to Nonlinear Optimization : Theory, Algorithms, and Applications with MATLAB*. Philadelphia, PA: Society for Industrial and Applied Mathematics; Mathematical Optimization Society. doi: 10.1137/1.9781611973655

Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., and Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* 45, 434–444. doi: 10.1109/78.554307

Bensaid, S., Schutz, A., and Slock, D. T. M. (2010). "Single microphone blind audio source separation using EM-Kalman filter and short+long term AR modeling," in *Latent Variable Analysis and Signal Separation, LVA/ICA 2010*. Lecture Notes in Computer Science, Vol. 6365, eds V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-15995-4_14

Chen, C. C., Hsieh, J. C., Wu, Y. Z., Lee, P. L., Chen, S. S., Niddam, D. M., et al. (2008). Mutual-information-based approach for neural connectivity during self-paced finger lifting task. *Hum. Brain Mapp.* 29, 265–280. doi: 10.1002/hbm.20386

Chen, M., Holobar, A., Zhang, X., and Zhou, P. (2016). Progressive fastICA peel-off and convolution kernel compensation demonstrate high agreement for high density surface EMG decomposition. *Neural Plast.* 2016:3489540. doi: 10.1155/2016/3489540

Chen, M., Zhang, X., Chen, X., and Zhou, P. (2018a). Automatic implementation of progressive fastICA peel-off for high density surface EMG decomposition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 144–152. doi: 10.1109/TNSRE.2017.2759664

Chen, M., Zhang, X., and Zhou, P. (2018b). A novel validation approach for high-density surface EMG decomposition in motor neuron disease. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 1161–1168. doi: 10.1109/TNSRE.2018.2836859

Chen, M., and Zhou, P. (2016). A novel framework based on fastICA for high density surface EMG decomposition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 117–127. doi: 10.1109/TNSRE.2015.2412038

Clamann, H. P. (1969). Statistical analysis of motor unit firing patterns in a human skeletal muscle. *Biophys. J.* 9, 1233–1251. doi: 10.1016/S0006-3495(69)86448-9

Crochiere, R. E. (1979). "A general program to perform sampling rate conversion of data by rational ratios," in *Programs for Digital Signal Processing (Digital Signal Processing Committee of the IEEE Acoustics, Speech, and Signal Processing Society*, ed C. J. Weinstein (New York, NY: IEEE Press, Programs), 8.2-1–8.2-7.

Daszykowski, M., Walczak, B., and Massart, D. L. (2002). Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS. *J. Chem. Inf. Comput. Sci.* 42, 500–507. doi: 10.1021/ci010384s

De Luca, C. J., Adam, A., Wotiz, R., Gilmore, L. D., and Nawab, S. H. (2006). Decomposition of surface EMG signals. *J. Neurophysiol.* 96, 1646–1657. doi: 10.1152/jn.00009.2006

De Luca, C. J., and Nawab, S. H. (2011). Reply to farina and enoka: the reconstruct-and-test approach is the most appropriate validation for surface EMG signal decomposition to date. *J. Neurophysiol.* 105, 983–984. doi: 10.1152/jn.01060.2010

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. New York, NY: Wiley.

Durrant-Whyte, H. F. (1988). Sensor models and multisensor integration. *Int. J. Rob. Res.* 7, 97–113. doi: 10.1177/027836498800700608

Farina, D., and Holobar, A. (2015). Human- machine interfacing by decoding the surface electromyogram [Life Sciences]. *IEEE Signal Process. Mag.* 32, 115–120. doi: 10.1109/MSP.2014.2359242

Farina, D., and Holobar, A. (2016). Characterization of human motor units from surface EMG decomposition. *Proc. IEEE* 104, 353–373. doi: 10.1109/JPROC.2015.2498665

Farina, D., Holobar, A., Merletti, R., and Enoka, R. M. (2010). Decoding the neural drive to muscles from the surface electromyogram. *Clin. Neurophysiol.* 121, 1616–1623. doi: 10.1016/j.clinph.2009.10.040

Farina, D., Jiang, N., Rehbaum, H., Holobar, A., Graimann, B., Dietl, H., et al. (2014a). The extraction of neural information from the surface EMG for the control of upper-limb prostheses: emerging avenues and challenges. *IEEE Trans. Neural Syst. Rehabil. Eng.* 22, 797–809. doi: 10.1109/TNSRE.2014.2305111

Farina, D., and Merletti, R. (2001). A novel approach for precise simulation of the EMG signal detected by surface electrodes. *IEEE Trans. Biomed. Eng.* 48, 637–646. doi: 10.1109/10.923782

Farina, D., Merletti, R., and Enoka, R. M. (2004). The extraction of neural strategies from the surface EMG. *J. Appl. Physiol.* 96, 1486–1495. doi: 10.1152/japplphysiol.01070.2003

Farina, D., Merletti, R., and Enoka, R. M. (2014b). The extraction of neural strategies from the surface EMG: an update. *J. Appl. Physiol.* 117, 1215–1230. doi: 10.1152/japplphysiol.00162.2014

Farina, D., Rehbaum, H., Holobar, A., Vujaklija, I., Jiang, N., Hofer, C., et al. (2014c). Noninvasive, accurate assessment of the behavior of representative populations of motor units in targeted reinnervated muscles. *IEEE Trans. Neural Syst. Rehabil. Eng.* 22, 810–819. doi: 10.1109/TNSRE.2014.2306000

Feo, T. A., and Resende, M. G. (1995). Greedy randomized adaptive search procedures. *J. Glob. Optim.* 6, 109–133. doi: 10.1007/BF01096763

Fuglevand, J., Winter, D. A., and Patla, A. E. (1993). Models of recruitment and rate coding organization in motor-unit pools. *J. Neurophysiol.* 70, 2470–2488. doi: 10.1152/jn.1993.70.6.2470

Gallego, J. A., Dideriksen, J. L., Holobar, A., Ibanez, J., Glaser, V., Romero, J. P., et al. (2015a). The phase difference between neural drives to antagonist muscles in essential tremor is associated with the relative strength of supraspinal and afferent input. *J. Neurosci.* 35, 8925–8937. doi: 10.1523/JNEUROSCI.0106-15.2015

Gallego, J. A., Dideriksen, J. L., Holobar, A., Ibanez, J., Pons, J. L., Louis, E. D., et al. (2015b). Influence of common synaptic input to motor neurons on

the neural drive to muscle in essential tremor. *J. Neurophysiol.* 113, 182–191. doi: 10.1152/jn.00531.2014

Gallego, J. A., Dideriksen, J. L., Holobar, A., Rocon, E., Pons, J. L., and Farina, D. (2017). "Neural control of muscles in tremor patients," in *Converging Clinical and Engineering Research on Neurorehabilitation II: Proceedings of the 3rd International Conference on NeuroRehabilitation (ICNR2016), October 18-21, 2016, Segovia, Spain,* eds. J. Ibáñez, J. González-Vargas, J. M. Azorín, M. Akay and J. L. Pons (Cham: Springer International Publishing), 129–134. doi: 10.1007/978-3-319-46669-9_24

Garcia, G. A., Okuno, R., and Azakawa, K. (2005). A decomposition algorithm for surface electrode-array electromyogram. *IEEE Eng. Med. Biol. Magaz.* 24, 63–72. doi: 10.1109/MEMB.2005.1463398

Gazzoni, M., Farina, D., and Merletti, R. (2004). A new method for the extraction and classification of single motor unit action potentials from surface EMG signals. *J. Neurosci. Methods* 136, 165–177. doi: 10.1016/j.jneumeth.2004.01.002

Ge, M., and Kerrigan, E. C. (2017). Noise covariance identification for non-linear systems using expectation maximization and moving horizon estimation. *Automatica* 77, 336–343. doi: 10.1016/j.automatica.2016.11.011

Ghaderi, P., and Marateb, H. R. (2017). Muscle activity map reconstruction from high density surface EMG signals with missing channels using image inpainting and surface reconstruction methods. *IEEE Trans. Biomed. Eng.* 64, 1513–1523. doi: 10.1109/TBME.2016.2603463

Glaser, V., Holobar, A., and Zazula, D. (2007). "An approach to the real-time surface electromyogram decomposition," in *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007*, eds T. Jarm, P. Kramar, and A. Zupanic (Heidelberg: Springer), 105–108. doi: 10.1007/978-3-540-73044-6_27

Glaser, V., Holobar, A., and Zazula, D. (2013). Real-Time motor unit identification from high-density surface EMG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 21, 949–958. doi: 10.1109/TNSRE.2013.2247631

Gross, J., Kujala, J., Hamalainen, M., Timmermann, L., Schnitzler, A., and Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 98, 694–699. doi: 10.1073/pnas.98.2.694

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* 9, 934–946. doi: 10.1038/nrn2497

Hardin, J. W. (2005). "Generalized Estimating Equations (GEE)," in *Encyclopedia of Statistics in Behavioral Science*, eds B. S. Everitt and D. C. Howell (Hoboken, NJ: John Wiley & Sons Inc.). doi: 10.1002/0470013192.bsa250

Hashimoto, Y., Ushiba, J., Kimura, A., Liu, M., and Tomita, Y. (2010). Correlation between EEG-EMG coherence during isometric contraction and its imaginary execution. *Acta Neurobiol. Exp.* 70, 76–85.

Holobar, A., and Farina, D. (2014). Blind source identification from the multichannel surface electromyogram. *Physiol Meas.* 35, R143–R165. doi: 10.1088/0967-3334/35/7/R143

Holobar, A., Farina, D., Gazzoni, M., Merletti, R., and Zazula, D. (2009). Estimating motor unit discharge patterns from high-density surface electromyogram. *Clin. Neurophysiol.* 120, 551–562. doi: 10.1016/j.clinph.2008.10.160

Holobar, A., Glaser, V., Gallego, J. A., Dideriksen, J. L., and Farina, D. (2012). Non-invasive characterization of motor unit behaviour in pathological tremor. *J. Neural Eng.* 9:056011. doi: 10.1088/1741-2560/9/5/056011

Holobar, A., Minetto, M. A., Botter, A., Negro, F., and Farina, D. (2010). Experimental analysis of accuracy in the identification of motor unit spike trains from high-density surface EMG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 18, 221–229. doi: 10.1109/TNSRE.2010.2041593

Holobar, A., Minetto, M. A., and Farina, D. (2014). Accurate identification of motor unit discharge patterns from high-density surface EMG and validation with a novel signal-based performance metric. *J. Neural Eng.* 11:016008. doi: 10.1088/1741-2560/11/1/016008

Holobar, A., and Zazula, D. (2004). Multichannel Blind source separation using convolution kernel compensation. *IEEE Trans. Signal Process.* 55, 4487–4496. doi: 10.1109/TSP.2007.896108

Holobar, A., and Zazula, D. (2007). "Gradient convolution kernel compensation applied to surface electromyograms," in *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, London, UK, September 9-12, 2007. Proceedings*, eds. M. E. Davies,. C. J. James, S. A. Abdallah, and M. D. Plumbley (Heidelberg: Springer), 617–624. doi: 10.1007/978-3-540-74494-8_77

Holobar, A., and Zazula, D. (2008). "On the selection of the cost function for gradient-baseddecomposition of surface electromyograms," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE: Vancouver), 4668–4671.

Ioannides, A. A., and Mitsis, G. D. (2010). Do we need to consider non-linear information flow in corticomuscular interaction? *Clin. Neurophysiol.* 121, 272–273. doi: 10.1016/j.clinph.2009.11.005

Jacobs, A. L., Fridman, G., Douglas, R. M., Alam, N. M., Latham, P. E., Prusky, G. T., et al. (2009). Ruling out and ruling in neural codes. *Proc. Natl Acad. Sci. U.S.A.* 106, 5936–5941. doi: 10.1073/pnas.0900573106

Jayaram, B., and Klawonn, F. (2013). "Can fuzzy clustering avoid local minima and undesired partitions?," in *Computational Intelligence in Intelligent Data Analysis*, eds. C. Moewes and A. Nürnberger (Heidelberg: Springer), 31–44. doi: 10.1007/978-3-642-32378-2_3

Karimimehr, S., Marateb, H. R., Muceli, S., Mansourian, M., Mañanas, M. A., and Farina, D. (2017). A real-time method for decoding the neural drive to muscles using single-channel intra-muscular EMG recordings. *Int. J. Neural Syst.* 27:1750025. doi: 10.1142/S0129065717500253

Kevin, C. M. (1984). *A Method for Quantitating the Clinical Electromyography*. Ph.D., Stanford University.

Kim, B., Kim, L., Kim, Y.-H., and Yoo, S. K. (2017). Cross-association analysis of EEG and EMG signals according to movement intention state. *Cogn. Syst. Res.* 44, 1–9. doi: 10.1016/j.cogsys.2017.02.001

Kleine, B. U., Van Dijk, J. P., Lapatki, B. G., Zwarts, M. J., and Stegeman, D. F. (2007). Using two-dimensional spatial information in decomposition of surface EMG signals. *J. Electromyogr. Kinesiol.* 17, 535–548. doi: 10.1016/j.jelekin.2006.05.003

Krzakala, F., Ricci-Tersenghi, F., Zdeborova, L., Zecchina, R., Tramel, E. W., and Cugliandolo, L. F. (2016). *Statistical Physics, Optimization, Inference, and Message-Passing Algorithms : École de Physique des Houches : Special Issue, 30 September-11 October 2013*. Oxford: Oxford University Press.

Larose, D. T. (2005). "Hierarchical and k-means clustering," in *Discovering Knowledge in Data*, ed D. T. Larose (Hoboken, NJ: John Wiley & Sons Inc.), 147–162. doi: 10.1002/0471687545.ch8

Li, W. (1990). Mutual information functions versus correlation functions. *J. Stat. Phys.* 60, 823–837. doi: 10.1007/BF01025996

Li, X., Holobar, A., Gazzoni, M., Merletti, R., Rymer, W. Z., and Zhou, P. (2015). Examination of poststroke alteration in motor unit firing behavior using high-density surface EMG decomposition. *IEEE Trans. Biomed. Eng.* 62, 1242–1252. doi: 10.1109/TBME.2014.2368514

Loh, W.-K., and Park, Y.-H. (2014). "A survey on density-based clustering algorithms," in *Ubiquitous Information Technologies and Applications: CUTE 2013*, eds Y-S. Jeong, Y-H. Park, C-H. Hsu and J. J. Park (Heidelberg: Springer), 775–780. doi: 10.1007/978-3-642-41671-2_98

Marateb, H. R., McGill, K. C., Holobar, A., Lateva, Z. C., Mansourian, M., and Merletti, R. (2011a). Accuracy assessment of CKC high-density surface EMG decomposition in biceps femoris muscle. *J. Neural Eng.* 8:066002. doi: 10.1088/1741-2560/8/6/066002

Marateb, H. R., Muceli, S., McGill, K. C., Merletti, R., and Farina, D. (2011b). Robust decomposition of single-channel intramuscular EMG signals at low force levels. *J. Neural Eng.* 8:066015. doi: 10.1088/1741-2560/8/6/066015

Marateb, H. R., Rojas-Martínez, M., Mansourian, M., Merletti, R., and Villanueva, M. A. M. (2012). Outlier detection in high-density surface electromyographic signals. *Med. Biol. Eng. Comput.* 50, 79–89. doi: 10.1007/s11517-011-0790-7

Martinez-Valdes, E., Laine, C. M., Falla, D., Mayer, F., and Farina, D. (2016). High-density surface electromyography provides reliable estimates of motor unit behavior. *Clin. Neurophysiol.* 127, 2534–2541. doi: 10.1016/j.clinph.2015.10.065

McGill, K. C. (2004). Surface electromyogram signal modelling. *Med. Biol. Eng. Comput.* 42, 446–454. doi: 10.1007/BF02350985

McGill, K. C., and Dorfman, L. J. (1984). High-resolution alignment of sampled waveforms. *IEEE Trans. Biomed. Eng.* 462–468. doi: 10.1109/TBME.1984.325413

McGill, K. C., Lateva, Z. C., and Marateb, H. R. (2005). EMGLAB: An interactive EMG decomposition program. *J. Neurosci. Methods* 149, 121–133. doi: 10.1016/j.jneumeth.2005.05.015

Meng, F., Tong, K. Y., Chan, S. T., Wong, W. W., Lui, K. H., Tang, K. W., et al. (2008). Study on connectivity between coherent central

rhythm and electromyographic activities. *J. Neural Eng.* 5, 324–332. doi: 10.1088/1741-2560/5/3/005

Merletti, R., and Farina, D. (2016). *Surface Electromyography : Physiology, Engineering and Applications.* Hoboken, NJ: IEEE Press; Wiley. doi: 10.1002/9781119082934

Merletti, R., and Parker, P. A. (2004). *Electromyography: Physiology, Engineering, and Non-invasive Applications.* Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0471678384

Nawab, S. H., Chang, S.-S., and De Luca, C. J. (2010). High-yield decomposition of surface EMG signals. *Clin. Neurophysiol.* 121, 1602–1615. doi: 10.1016/j.clinph.2009.11.092

Nawab, S. H., Wotiz, R. P., and De Luca, C. J. (2008). Decomposition of indwelling EMG signals. *J. Appl. Physiol.* 105, 700–710. doi: 10.1152/japplphysiol.00170.2007

Negro, F., and Farina, D. (2011). Linear transmission of cortical oscillations to the neural drive to muscles is mediated by common projections to populations of motoneurons in humans. *J. Physiol.* 589, 629–637. doi: 10.1113/jphysiol.2010.202473

Negro, F., Muceli, S., Castronovo, A. M., Holobar, A., and Farina, D. (2016). Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. *J. Neural Eng.* 13:026027. doi: 10.1088/1741-2560/13/2/026027

Ning, Y., Zhu, S., Zhu, X., and Zhang, Y. (2013). "A hybrid multi-channel surface EMG decomposition approach by combining CKC and FCM," in: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER).* (San Diego, CA), 335–338. doi: 10.1109/NER.2013.6695940

Ning, Y., Zhu, X., Zhu, S., and Zhang, Y. (2015). Surface EMG decomposition based on K-means clustering and convolution kernel compensation. *IEEE J. Biomed. Health Inform.* 19, 471–477. doi: 10.1109/JBHI.2014.2328497

Pani, D., Barabino, G., Citi, L., Meloni, P., Raspopovic, S., Micera, S., et al. (2016). Real-time neural signals decoding onto off-the-shelf DSP processors for neuroprosthetic applications. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 993–1002. doi: 10.1109/TNSRE.2016.2527696

Povalej BrŽan, P., Gallego, J., Romero, J., Glaser, V., Rocon, E., Benito-León, J., et al. (2017). New perspectives for computer-aided discrimination of parkinson's disease and essential tremor. *Complexity* 2017:4327175. doi: 10.1155/2017/4327175

Radeke, J., Van Dijk, J. P., Holobar, A., and Lapatki, B. G. (2014). Electrophysiological method to examine muscle fiber architecture in the upper lip in cleft-lip patients. *J. Orofac. Orthop.* 75, 51–61. doi: 10.1007/s00056-013-0193-5

Ren, X., Hu, X., Wang, Z., and Yan, Z. (2006). MUAP extraction and classification based on wavelet transform and ICA for EMG decomposition. *Med. Biol. Eng. Comput.* 44, 371–382. doi: 10.1007/s11517-006-0051-3

Rey, H. G., Pedreira, C., and Quian Quiroga, R. (2015). Past, present and future of spike sorting techniques. *Brain Res. Bull.* 119, 106–117. doi: 10.1016/j.brainresbull.2015.04.007

Rubino, D., Robbins, K. A., and Hatsopoulos, N. G. (2006). Propagating waves mediate information transfer in the motor cortex. *Nat. Neurosci.* 9, 1549–1557. doi: 10.1038/nn1802

Savc, M., Glaser, V., Kranjec, J., Cikajlo, I., Matjačić, Z., and Holobar, A. (2018). Comparison of convolutive kernel compensation and non-negative matrix factorization of surface electromyograms. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 1935–1944. doi: 10.1109/TNSRE.2018.2869426

Schelter, B., Timmer, J., and Eichler, M. (2009). Assessing the strength of directed influences among neural signals using renormalized partial directed coherence. *J. Neurosci. Methods* 179, 121–130. doi: 10.1016/j.jneumeth.2009.01.006

Schwartz, B., Gannot, S., and Habets, E. A. (2015). Online speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Trans. Speech Lang. Process.* 23, 394–406. doi: 10.1109/TASLP.2014.2372342

Schwartz, O., Gannot, S., and Habets, E. A. (2016). An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation. *IEEE/ACM Trans. Speech Lang. Process.* 24, 1495–1510. doi: 10.1109/TASLP.2016.2553457

Siqueira Júnior, A. L. D., and Soares, A. B. (2015). A novel method for EMG decomposition based on matched filters. *Res. Biomed. Eng.* 31, 44–55. doi: 10.1590/2446-4740.0643

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends Cogn. Sci.* 8, 418–425. doi: 10.1016/j.tics.2004.07.008

Stearns, S. D., and Hush, D. R. (2011). *Digital Signal Processing With Examples in MATLAB®.* Boca Raton, FL: CRC Press.

Thomas, J., Deville, Y., and Hosseini, S. (2006). Time-domain fast fixed-point algorithms for convolutive ICA. *IEEE Signal Process. Lett.* 13, 228–231. doi: 10.1109/LSP.2005.863638

Úbeda, A., Del Vecchio, A., Sartori, M., Yavuz, U. S., Negro, F., Felici, F., et al. (2017). "Corticospinal coherence during frequency-modulated isometric ankle dorsiflexion," in *Converging Clinical and Engineering Research on Neurorehabilitation II: Proceedings of the 3rd International Conference on NeuroRehabilitation (ICNR2016), October 18-21, 2016, Segovia, Spain,* eds J. Ibáñez, J. González-Vargas, J. M. Azorín, M. Akay, and J. L. Pons (Cham: Springer International Publishing), 135–140.

Watanabe, K., Gazzoni, M., Holobar, A., Miyamoto, T., Fukuda, K., Merletti, R., et al. (2013). Motor unit firing pattern of vastus lateralis muscle in type 2 diabetes mellitus patients. *Muscle Nerve* 48, 806–813. doi: 10.1002/mus.23828

Watanabe, K., Holobar, A., Kouzaki, M., Ogawa, M., Akima, H., and Moritani, T. (2016). Age-related changes in motor unit firing pattern of vastus lateralis muscle during low-moderate contraction. *Age* 38:48. doi: 10.1007/s11357-016-9915-0

Webster, J. G., Marateb, H. R., and McGill, K. C. (2016). "Electromyographic (EMG) decomposition," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, ed J. G. Webster (Hoboken, NJ: John Wiley & Sons Inc.). doi: 10.1002/047134608X.W8296

Werner, T., Vianello, E., Bichler, O., Garbin, D., Cattaert, D., Yvert, B., et al. (2016). Spiking neural networks based on OxRAM synapses for real-time unsupervised spike sorting. *Front. Neurosci.* 10:474. doi: 10.3389/fnins.2016.00474

Wheeler, K. R., Chang, M. H., and Knuth, K. H. (2006). Gesture-based control and EMG decomposition. *IEEE Trans. Syst. Man Cyber.* 36, 503–514. doi: 10.1109/TSMCC.2006.875418

Winslow, J., Dididze, M., and Thomas, C. K. (2009). Automatic classification of motor unit potentials in surface EMG recorded from thenar muscles paralyzed by spinal cord injury. *J. Neurosci. Methods* 185, 165–177. doi: 10.1016/j.jneumeth.2009.09.012

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement.* Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470549148

Xu, Z., Xiao, S., and Chi, Z. (2001). ART2 neural network for surface EMG decomposition. *Neural Comput. Appl.* 10, 29–38. doi: 10.1007/s005210170015

Xue, M., Wu, H., Zeng, Y., and Li, Y. (2017). "Unsupervised neuron spike decoding for macaque's finger position via EM algorithm," in *Information, Cybernetics and Computational Social Systems (ICCSS), 2017 4th International Conference* (IEEE: Dalian), 599–604. doi: 10.1109/ICCSS.2017.8091485

Yoshida, K., Farina, D., Akay, M., and Jensen, W. (2010). Multichannel intraneural and intramuscular techniques for multiunit recording and use in active prostheses. *Proc. IEEE* 98, 432–449. doi: 10.1109/JPROC.2009.2038613

Zalewska, E. (2009). Insight into the motor unit activation and structure properties gained from EMG signal analysis. *Clin. Neurophysiol.* 120, 449–450. doi: 10.1016/j.clinph.2008.12.037

Zazula, D., and Holobar, A. (2005). An approach to surface EMG decomposition based on higher-order cumulants. *Comput. Methods Programs Biomed.* 80(Suppl. 1), S51–60. doi: 10.1016/S0169-2607(05)80006-9

Zhou, P., and Rymer, W. Z. (2004). MUAP number estimates in surface EMG: template-matching methods and their performance boundaries. *Ann. Biomed. Eng.* 32, 1007–1015. doi: 10.1023/B:ABME.0000032463.26331.b3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer PZ declared a past co-authorship with one of the authors, AH, to the handling Editor.

Copyright © 2019 Mohebian, Marateb, Karimimehr, Mañanas, Kranjec and Holobar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 03 May 2019
doi: 10.3389/fnins.2019.00398

A Regression-Based Framework for Quantitative Assessment of Muscle Spasticity Using Combined EMG and Inertial Data From Wearable Sensors

Xu Zhang[1†], Xiao Tang[1†], Xiaofei Zhu[1], Xiaoping Gao[2], Xiang Chen[1] and Xun Chen[1*]*

[1] *Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, China,*
[2] *Department of Rehabilitation Medicine, First Affiliated Hospital of Anhui Medical University, Hefei, China*

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Chris A. McGibbon,
University of New Brunswick
Fredericton, Canada
Zhong Yin,
University of Shanghai for Science and Technology, China

***Correspondence:**
Xiaoping Gao
gxp678@163.com
Xun Chen
xunchen@ustc.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *19 December 2018*
Accepted: *08 April 2019*
Published: *03 May 2019*

Citation:
Zhang X, Tang X, Zhu X, Gao X, Chen X and Chen X (2019) A Regression-Based Framework for Quantitative Assessment of Muscle Spasticity Using Combined EMG and Inertial Data From Wearable Sensors. Front. Neurosci. 13:398. doi: 10.3389/fnins.2019.00398

There have always been practical demands for objective and accurate assessment of muscle spasticity beyond its clinical routine. A novel regression-based framework for quantitative assessment of muscle spasticity is proposed in this paper using wearable surface electromyogram (EMG) and inertial sensors combined with a simple examination procedure. Sixteen subjects with elbow flexor or extensor (i.e., biceps brachii muscle or triceps brachii muscle) spasticity and eight healthy subjects were recruited for the study. The EMG and inertial data were recorded from each subject when a series of passive elbow stretches with different stretch velocities were conducted. In the proposed framework, both lambda model and kinematic model were constructed from the recorded data, and biomarkers were extracted respectively from the two models to describe the neurogenic component and biomechanical component of the muscle spasticity, respectively. Subsequently, three evaluation methods using supervised machine learning algorithms including single-/multi-variable linear regression and support vector regression (SVR) were applied to calibrate biomarkers from each single model or combination of two models into evaluation scores. Each of these evaluation scores can be regarded as a prediction of the modified Ashworth scale (MAS) grade for spasticity assessment with the same meaning and clinical interpretation. In order to validate performance of three proposed methods within the framework, a 24-fold leave-one-out cross validation was conducted for all subjects. Both methods with each individual model achieved satisfactory performance, with low mean square error (MSE, 0.14 and 0.47) between the resultant evaluation score and the MAS. By contrast, the method using SVR to fuse biomarkers from both models outperformed other two methods with the lowest MSE at 0.059. The experimental results demonstrated the usability and feasibility of the proposed framework, and it provides an objective, quantitative and convenient solution to spasticity assessment, suitable for clinical, community, and home-based rehabilitation.

Keywords: spasticity assessment, machine learning, surface electromyogram, stroke, home-based rehabilitation

INTRODUCTION

Spasticity is a clinical symptom of neural disorders and was prevalent in stroke, spinal cord injury, cerebral palsy and multiple sclerosis patients (Nielsen et al., 2007; Naghdi et al., 2014). It was considered to be caused by the increased excitability of stretch reflex (Nielsen et al., 2007). The primary manifestation of spasticity is an abnormal muscle tone in the involved muscle, which is generally shown as a resistance when the muscle is stretched passively (Trompetto et al., 2014). The resistance is also found to be velocity-dependent (Lee et al., 2002; Mullick et al., 2013). The spasticity may cause abnormal physical appearance and even muscle contracture (Ada et al., 2006; Naghdi et al., 2014), consequently weakening the ability of physical activity and seriously affecting the daily life of patients. Therefore, quantitative assessment of spasticity is vital for early interventions during rehabilitation treatment.

For quantitative assessment of spasticity, routine clinical scales such as modified Ashworth scale (MAS) and the modified tardieu scale (MTS) are most frequently used (Mehrholz et al., 2005). The MAS measures the resistance during passive soft-tissue stretching without consideration of stretch velocity. According to the specified regulations pre-defined in MAS, clinicians are able to grade the spasticity. Although limitations in performance exist due to its rough measurement procedures and relative subjectivity (Pandyan et al., 1999, 2003; Fleuren et al., 2010; Phadke et al., 2015), standardized user guidelines have been developed and the inter-rater reliability for testing of upper limb spasticity has been proven to be acceptable (Blackburn et al., 2002). In addition, due to its convenience without requirement of instrument, the MAS is a primary clinical measure of spasticity and is even considered as a gold standard of being the referential yardstick for other new spasticity measurement methods (Pisano et al., 2000). The measurement of MTS involves two passive movements at a very slow speed and a high speed as fast as possible, respectively (Mehrholz et al., 2005). It has been regarded as a more reliable scale for spasticity assessment with higher inter-rater reliability demonstrated in several studies (Singh et al., 2011), but the validity and reliability of MTS are still doubtable in some investigations (Mehrholz et al., 2005; Naghdi et al., 2014). With this semi-quantitative manner of measurement, the use of clinical scales just meets basic needs of clinical applications. More objective and quantifiable tools are demanded for muscle spasticity assessment.

Electrophysiological techniques can be used to examine altered neurotransmission in the spinal neuronal pathway for understanding underlying pathophysiology as well as assessing muscle spasticity (Pisano et al., 2000; Cooper et al., 2005; Biering-Sørensen et al., 2006; Kim et al., 2011; Bar-On et al., 2013; Naghdi et al., 2014). Both Hoffmann-reflex (H-reflex) test and electromyography (EMG) measurement are always involved in these techniques. Several investigations reported that biomarkers extracted from the H-reflex test increased in patients with spasticity compared to healthy controls but such an increase exhibited no or poor correlation to the clinical scores (Naghdi et al., 2014). An increase in EMG biomarker during instrumented passive stretch was found to moderately correlate with the MAS grades in lower limb spasticity of children with cerebral palsy (Jobin and Levin, 2000; Bar-On et al., 2013). Despite limited contribution to the quantitative assessment of spasticity, the electrophysiological examinations can provide an easy and the most reliable way of determining the stretch reflex threshold and information concerning neural transmission. Thus their potential role in quantifying spasticity is significant.

Alternatively, biomechanical approaches are frequently used to quantify the velocity-dependent resistance during passive movement by measuring joint position, angular velocity, and torque (reactive-resistance) (Pisano et al., 2000; Condliffe et al., 2005; Ardabili et al., 2011; Bar-On et al., 2013; Li et al., 2017). These investigations focus on assessing the muscular mechanics of spasticity, and have achieved highly reliable assessment for spasticity. For example, McGibbon et al. (2016) presented a kinematic model and addressed the issue of identifying categorical degrees of muscle spasticity by classification learning using a linear discriminant analysis algorithm. However, specific devices such as the isokinetic and torque sensors and well-controlled laboratory environments are generally required to ensure precise measurement and accurate quantification (Condliffe et al., 2005; Ardabili et al., 2011). With sophisticated protocols, biomechanical approaches are usually involved in scientific research rather than clinical practice.

Furthermore, various investigations about spasticity have mentioned that the spasticity resistance consists of both neurogenic and non-neurogenic (primarily mechanical) components (Huang et al., 2004; Lam et al., 2005). The former mainly represents abnormal neural regulation and induced hyperactivity of motoneuron (Nielsen et al., 2007). The latter may result from altered mechanical properties of muscle soft tissues (the muscle fibers and tendons) and limb joints, which form the secondary contributor to the spasticity resistance (Stecco et al., 2014). Thus it is reasonable that a precise spasticity assessment would involve both electrophysiological and biomechanical aspects. Recently, many efforts have been made toward assessing spasticity from both aspects with simultaneous use of electrophysiological and biomechanical measurements. For example, NeuroFlexor has been developed by Lindberg et al. (2011) as a fine-tuned device with capability of quantitative discrimination between the neural and mechanical components of the total resistance as a result of muscle spasticity. Due to its capability, this device was reported to be used for spasticity assessment with satisfactory precision and reliability.

Besides the precision, there is a great demand for wearability and portability of the devices and fast and convenient implementation of the measurement. Spasticity is always regarded to vary across individuals and to last a long term over multiple stages of motor recovery. Its assessment is required to be repeated for continuous management in multiple sites from inpatient clinic through community to home. Specifically, considering the prevalence of muscle spasticity after a variety of neurological injuries, there is an increasing demands for implementing spasticity assessment in community and home environments, where it is impractical to apply high-cost specialized instruments and complex experimental setups. Therefore, it is of great significance to develop a convenient

and effective approach for quantitative assessment of muscle spasticity suitable for community and home rehabilitation.

With the above-mentioned considerations, this paper presents a novel regression-based framework for quantitative assessment of muscle spasticity using a wearable sensing system including surface EMG (sEMG) and inertial sensors along with a simple and practical procedure. Corresponding to both the neural and biomechanical components contributing to muscle spasticity, two models (the lambda model and the kinematic model) are built respectively to derive biomarkers from sensory data, which are discriminable across different degrees of spasticity. These biomarkers, from one model or the combination of both models need to be processed to predict an evaluation score that is the same meaningful as the routine clinical assessment score (i.e., the MAS) but has purely objective and quantitative characters. Supervised regression learning is a nature solution as it is able to establish an expert system (also known as numerical model) for automatically and accurately producing numerical evaluation scores. Therefore, supervised regression algorithms such as single-/multi-variable linear regression and support vector regression (SVR) were employed within the framework. The framework proposed in this study provides a practical solution to spasticity assessment in a convenient, objective and quantitative way. With the demonstrated effectiveness, such convenience can especially extend the sites for assessing the muscle spasticity and its intervention outcome from inpatient hospital to home, toward advanced spasticity management and rehabilitation.

MATERIALS AND METHODS

Subjects

Sixteen subjects with spasticity (14 males and 2 females; aged from 33 to 71 years, averaged 54 ± 10 years; time duration since diagnosis from 19 to 86 days, averaged 52 ± 10 days) and eight healthy subjects (six males and two females; aged from 22 to 45 years, averaged 29 ± 9 years) were recruited in the study. This study was approved by the Ethics Review Committee for Clinical Medical Research of First Affiliated Hospital of Anhui Medical University. Inclusion criteria for participants with spasticity include: (a) currently experiencing one of the following disease: stroke (cerebral ischemia, brain hemorrhage or brainstem infarction and etc.), acquired brain trauma or incomplete spinal cord injury and accompanied by spasticity in flexor and extension muscles of the elbow; (b) the spasticity of elbow extensor or flexor was assessed within 1–3 grades using MAS; (c) the range of elbow joint during passive stretch was at least 120 degrees; (d) medically stable with clearance to participate; (e) without any historical musculoskeletal injuries or cognition problems; (f) able to offer informed signed consent prior to any procedure of the experiment. For all subjects, clinical assessment of spasticity using the MAS was performed just half an hour prior to their participation into the experiment, by one qualified, experienced physical therapist. The demographic information of the participants with spasticity is presented in **Table 1**, and characteristics of healthy subjects is provided in **Table 2**.

Experiments

A home-made multi-channel signal recording system supporting up to 16 sEMG recording channels and 9-axis inertial measurement units (IMUs) was used in this study. In this system, each individual EMG sensor consists of two parallel bar-shaped electrodes with a width of 1 mm, a length of 10 mm, and a between-electrode distance of 10 mm, thus constituting one single-differential sEMG recording channel. Each IMU (MPU9250, InvenSense, San Jose, CA, United States) consisted of a 3-axis accelerometer, a 3-axis gyroscope and a 3-axis magnetometer for recording inertial data. In the experiments, one sEMG sensor was used to target at the belly of the biceps brachii muscle or the triceps brachii muscle (depending on appearance of the muscle spasticity) with its electrode bar perpendicular to the muscle fibers and the IMU was placed on the ipsilateral medial wrist for recording stretch angular velocity (see **Figure 1A**). The surface EMG sensor and the IMU were embedded in a stretchable armband and wristband respectively to ensure secured placement and to enhance portability of both sensors. A round reference electrode (Dermatrode; American Imex, Irvine, CA, United States) was placed on arm fossa cubitalis on the same side. In this system, each sEMG channel was amplified with a gain of 600 in total and further digitized by a 16 bit anolog-to-digital converter (ADS1198, Texas Instruments, Dallas, TX, United States) with a sampling rate of 1 kHz. The IMU was designed to produce digitalized data at 100 Hz per axis. The digitalized data from both sensors were simultaneously recorded.

After skin preparation with medical alcohol, the subjects were asked to participate into two sessions of test. Each session consisted of 15 to 20 trials of passive stretch with different velocities. In a single trial, the subject was instructed to fully extend his tested elbow at 180 degrees with palm upward. Then, the tested elbow was passively pulled by an experimenter to elbow flexion at 40–60 degrees with a stretch range of 120–140 degrees. After a 2-s pause, it was passively stretched back to 180 degrees (see **Figure 1B**). In each trial, the stretch velocity was determined subjectively by the experimenter and kept to be almost constant during the stretch. Varying stretch velocities were applied among the 15–20 trials, thus resulting in a varying duration of 3–7 s for each trial. Sufficiently long resting periods were allowed between any two consecutive trials and sessions to avoid mental or muscular fatigue and to reduce stretch-induced temporary muscle tones. No specialized equipment was required to control and record the velocity or stretch angle directly in this experiment. Meanwhile, all data were recorded by the system, transferred and restored to a portable computer via a USB cable for further analysis.

Data Analysis

The data analysis procedure of the proposed regression learning framework for spasticity evaluation is shown in **Figure 2**, with details described as follows.

TABLE 1 | Demographic information for each of 16 subjects with muscle spasticity.

No.	Age range	Sex	Brain lesion	Involved muscle	Duration (days)	Spastic side	MAS grade
1	46–50	Male	Brain hemorrhage	BB	56	Left	1+
2	41–45	Male	Brainstem hemorrhage	TB	69	Left	2
3	46–50	Male	Brain hemorrhage	TB	38	Right	1
4	71–75	Male	Brain hemorrhage	BB	29	Right	1+
5	56–60	Male	Brain infarction	BB	52	Right	1+
6	56–60	Female	Brain infarction	BB	76	Left	2
7	56–60	Male	Brain infarction	BB	73	Left	1+
8	56–60	Male	Left thalamic hemorrhage	TB	37	Right	2
9	46–50	Male	Brain hemorrhage	TB	62	Right	3
10	41–45	Male	Spinal cord injury	BB	86	Left	1
11	61–65	Female	Brain hemorrhage	BB	19	Right	1
12	56–60	Male	Brain hemorrhage	BB	52	Left	2
13	31–35	Male	Brain hemorrhage	BB	47	Right	1+
14	61–65	Male	Acquired brain trauma	BB	38	Right	1
15	46–50	Male	Brain hemorrhage	BB	60	Right	1+
16	36–40	Male	Spinal cord injury	BB	35	Right	3

BB, biceps brachii muscle; TB, triceps brachii muscle; MAS, modified Ashworth scale.

TABLE 2 | Demographic information for each of eight healthy subjects.

No.	Age range	Sex	Tested side	MAS grade
1	25–30	Female	Right	0
2	41–45	Male	Left	0
3	21–25	Male	Left	0
4	21–25	Male	Left	0
5	21–25	Male	Right	0
6	21–25	Female	Left	0
7	21–25	Male	Left	0
8	36–40	Male	Left	0

MAS, modified Ashworth scale.

Data Pre-processing and Segmentation

All the data were processed in the Matlab (version, 2014a, The Mathworks, Inc., Natick, MA, United States). The sEMG signals were filtered by a zero-lag fourth-order Butterworth band pass filter set at 20–450 Hz. Since only the angular velocity data were used in this study, the gyroscope signals were selected and then filtered by a zero-lag second-order Butterworth low pass filter set at 10 Hz. Representative recorded signals of a passive stretch trial were showed in **Figure 3**, including a three-axis gyroscope signal and an EMG signal. The stretch angular velocity was the vector superposition of the three-axis gyroscope signal. An amplitude thresholding method was employed to detect the onset of sEMG response as well as the onset and offset of stretch movement (Silva et al., 2017), as the vertical dashed lines marked in **Figure 3**, and its threshold was set at three times standard deviation of the baseline (Zhang et al., 2011). After auto-detection of onset, those segments with obvious noise contamination were discarded by visual examination.

Lambda Model

The lambda model is a classic measure to quantitatively evaluate spasticity using surface EMG and motion data (Levin et al., 2000). This model evaluates the stretch reflex threshold which refers to the joint angle when the EMG signal is exactly induced during a stretch movement of a targeted muscle (Levin et al., 2000). In addition, it presents that the regulation of the central nervous system (CNS) for the stretch reflex threshold is an essential mechanism for joint motion control, and it is related to the control of joint position. Thus the lesion of CNS after stroke may interferes with the regulation process then result in a deficits of the stretch reflex threshold regulation (Levin et al., 2000; Calota et al., 2008). As a result, the induced timing of EMG could vary a lot due to different degrees of CNS lesions. This could be reflected in the model in terms of decreased stretch reflex threshold. A velocity-dependent character of the stretch reflex threshold is also demonstrated in this model, for it is shown by the fact that a higher stretch velocity leads to a shorter latency of the stretch reflex. The shorter latency can be represented by a smaller angle range for joint motion.

Considering that the increase of stretch angular velocity generally gives a decreasing trend of stretch reflex threshold, a linear regression model was therefore built between the stretch angular velocity and dynamic stretch reflex threshold (DSRT, refers to the mutative stretch reflex threshold, the joint angular when EMG is induced, at different stretch angular velocity which is greater than zero) in these investigations. Please note that the original proposer accurately quantified the angular velocity with specialized equipment. With wearable IMU sensors to measure elbow joint movements, however, such velocity needed to be estimated approximately. The mean stretch angular velocity was calculated over the entire duration of elbow stretch as a rough estimation of the elbow joint stretch velocity. The DSRT

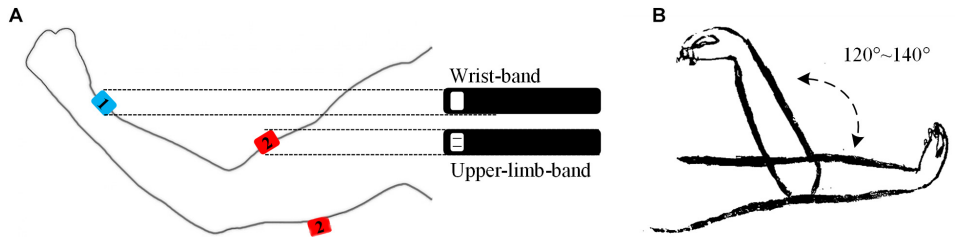


FIGURE 1 | Illustrations of the band-like sensor placement **(A)** and a passive elbow stretch **(B)** in the experiment. The sensor marked in number 1 stands for the IMU and the sensor marked in 2 a sEMG sensor, which was embedded in a wrist band and an upper-limb-band, respectively. The sEMG sensor can be used to target at the biceps brachii muscle or the triceps brachii muscle respectively.

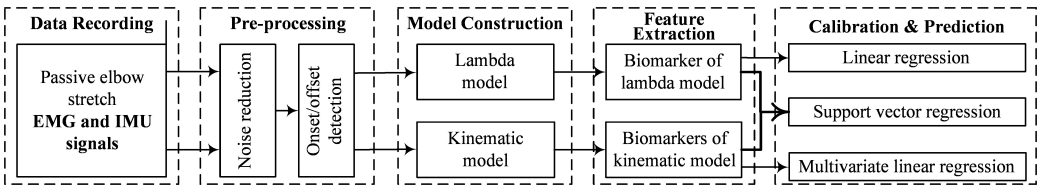


FIGURE 2 | Block diagram of the proposed framework for spasticity evaluation.

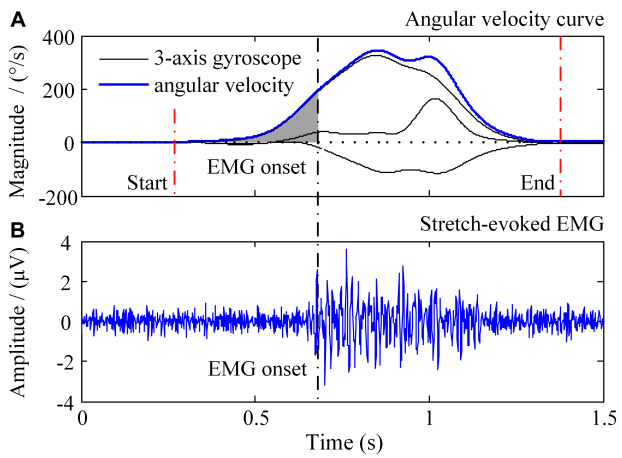


FIGURE 3 | Illustration of the recorded signals in a trial of passive elbow stretch. In the **(A)**, the black solid lines represent 3-axis gyroscope data and the blue solid line is the angular velocity, respectively. The **(B)** shows the sEMG signals evoked by passive elbow stretch.

was calculated by integrating the stretch angular velocity from the beginning to the moment when the EMG was evoked, which was marked as shaded area in **Figure 3**. Thus, the mean stretch angular velocity (independent variable) and the DSRT (dependent variable) were applied to construct the linear lambda model using a linear regression, as showed in **Figure 4A**. Due to the special design of approximation estimation using the averaging approach, multiple trials were required to enhance reliability of the regression analysis with a sufficient amount of data points. This explains the reason for conducting multiple trials in the experimental protocol. On this basis, exclusion of outliers was further implemented using a 95% confidence interval (Montgomery et al., 2012), when the linear regression was conducted for each subject. With the determined linear regression line, the intercept was calculated to be a special DSRT when the stretch velocity approximated into zero. It was termed tonic stretch reflex threshold (TSRT) in this study, indicating the stretch reflex threshold under quasi-static stretching. According to previous studies, the TSRT in the lambda model represents the physiological range of stretch reflex regulation (Levin et al., 2000). In general, the TSRT may exceed the biomechanical range of joint movement for healthy subjects (Musampa et al., 2007), as a reflection of normal stretch reflex regulation. For patients with spasticity, the lesion of CNS may result in a disorder of stretch reflex regulation, which would reduce the stretch reflex threshold as well as limit the range of joint movement. Therefore, TSRT was regarded as a biomarker derived from the lambda model for spasticity assessment. Please also note that sEMG would not be evoked until the stretch velocity exceeded a large threshold for healthy subjects. For any subject without typically evoked EMG in most of the trials (across multiple velocities), the TSRT was conformably set to 120 degrees, according to the physiological relevance of the lambda model.

In order to evaluate the degree of spasticity for each patient, the biomarker TSRT needs to be calibrated with a meaningful scale, which referred to the routine MAS in this study. Thus, the calibration procedure was implemented by the means of performing a supervised linear regression analysis between the TSRT biomarkers and MAS grades. In this analysis, MAS grades 0, 1, 1+, 2, and 3 were numerically represented by 0, 1, 1.5, 2, and 3, respectively. Given experimental data from 24 subjects in total (16 subjects with spasticity and 8 healthy controls), a 24-fold leave-one-out cross-validation was performed. Data from 23 subjects were used as the training dataset to establish a regression

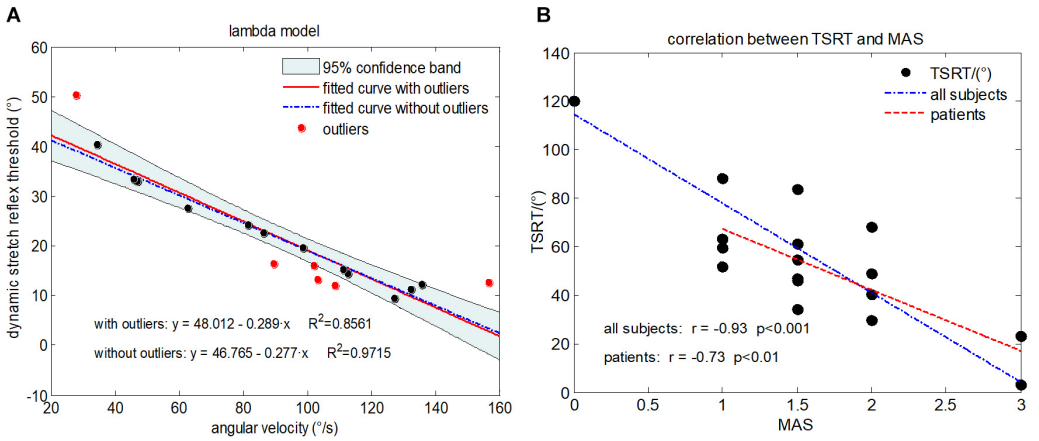


FIGURE 4 | (A) Illustration of the lambda model constructed for a representative subject (Subject 1) with MAS grade of 1+, where each data point is derived from a trial. The regression line derived from all data points is shown in red. The data points in red represent outliers. The final regression line was represented by a blue dashed line when outliners were excluded. **(B)** The scatter plot showing relationship between the TSRT value and MAS grades for all subjects. Each subject is represented by a data point located by its TSRT value and MAS grade. The blue dashed line is derived from regression analysis based on data of all subjects including healthy controls and subjects with spasticity. The red line denotes regression analysis performed on only subjects with spasticity.

line as the best fit between their biomarkers and MAS grades (i.e., labels for supervised learning), whereas data from the remaining subject were used as the testing dataset to predict an evaluation score from the input biomarker (i.e., the input feature) based on the learnt regression line. Thus, after 24 rounds of testing, each subject had a predicted evaluation score. Such a procedure was designed also with the purpose to verify the performance of the biomarker TSRT in assessing muscle spasticity.

Kinematic Model

The kinematic model was proposed for spasticity evaluation by McGibbon et al. (2013), which was based on a constant-jerk assumption. According to the assumption, the intended motion curve of elbow stretch could be constructed for each subject as a reference motion pattern (Musampa et al., 2007). A consistent pattern is assumed between the actual and reconstructed (reference) motion curve across healthy control muscles, whereas this consistency cannot be found in spastic muscles. Significant deviations from the reference pattern were observed on the actual motion curve due to interference from abnormal muscle tension. Consequently, the similarity between the actual motion curve and reference pattern was used to evaluate the degree of spasticity resistance.

The construction of kinematic model is illustrated in **Figure 5**, with an actual example of angular velocity curve recorded from the gyroscope (marked as a black solid line in **Figure 5** II) during a stretch in a single trial. Firstly, its entire time duration (t) was divided into two periods t1 and t2 by the moment of its maximal value (tmax), representing an accelerating stage and a decelerating stage respectively. Next, both integral operation and differential operation were performed on the angular velocity curve to obtain an angle curve (black solid line in I) and an angular acceleration curve (black solid line in III), respectively. Then, the angular acceleration curve (III) was reconstructed to form a subtriangular curve according to constant jerk assumption. Considering the fact that the vectorial summation of the angular velocity should be zero, the peak of angular acceleration in the second stage (denoted as acc2) was determined while this peak in the first stage (denoted as acc1) was optionally set at 4000 degree/s^2 here. Once peaks of angular acceleration in both stages were determined, the reconstructed angular acceleration curve (blue dashed line in III) was obtained following a subtriangular rule. Afterward, two consecutive integrations were performed on the reconstructed angular acceleration curve to calculate a reconstructed angle curve (the blue dashed line in I), followed by a scaling procedure for matching the range of motion. Consequently, a reconstructed angle curve (the black dashed line in IV) was obtained. Finally, the reconstructed angle curve in IV was differentiated once and twice to produce a reconstructed angular velocity curve (V) and a reconstructed acceleration curve (VI), respectively.

It was found in this study (see section "Results") that the reconstructed kinematic curves presented deviations from their corresponding recorded curves, and such deviations become greater in the cases of higher MAS grades. Especially, the angular acceleration curves exhibited more fluctuations with respect to its reconstructed ones for subjects with spasticity than healthy controls. In the kinematic model, therefore, three correlation coefficients between the actual curves and reconstructed curves of the angle, angular velocity and angular acceleration were calculated as three biomarkers. In addition, the median frequency (MDF) of actual angular acceleration curve (Szeto et al., 2005) was also calculated. Consequently, there are four biomarkers to measure the spasticity resistance for each subject. In order to minimize the effect of angular velocity variation as a result of manual stretch, we only chose the trails with time duration of 1–2 s for extracting biomarkers.

Similar to the quantitative evaluation method based on lambda model, the four biomarkers derived from the kinematic

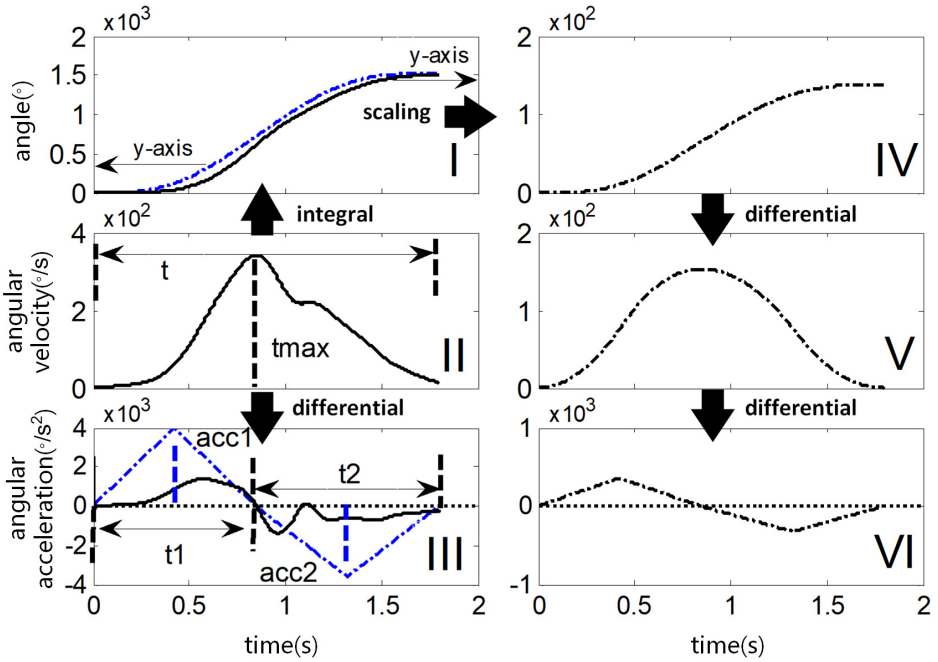


FIGURE 5 | Illustration of constructing kinematic variables in the kinematic model. The black solid line in I, II, and III represents the actual angle curve, the angular velocity curve, and the angular acceleration curve respectively during passive elbow stretch. Their reconstructed and scaled curves are shown in the black dotted line in IV, V, VI on the right panel. The blue dotted line in I and III illustrates the scaling and reconstruction procedure respectively, in detail for comparison.

model also needed to be calibrated to produce a meaningful evaluation score by means of performing a supervised multivariate regression analysis (Berndt and Savin, 1977). A 24-fold leave-one-out cross-validation was applied to data from 24 subjects. Biomarkers and corresponding MAS grades from any 23 subjects formed the training dataset to learn a regression line by supervised multivariate regression analysis, and the remaining sample was used as the testing dataset to produce an evaluation score based on the input biomarkers and the learnt regression line. Each subject produced an evaluation score after 24 round of testing.

Final Assessment by Fusion of Both Models

Given different but complementary information from both models in terms of spasticity assessment, their fusion may ensure comprehensive and improved performance. Therefore, fusion of both models was implemented within the proposed framework, by combining their biomarkers through a supervised regression learning procedure. For each tested muscle of one subject, the TSRT from the lambda model, the four biomarkers (i.e., correlation coefficients of the angle, the angular velocity and the angular acceleration between the recorded curves and the re-constructed ones, and the MDF of the recorded angular acceleration) from the kinematic model were concatenated to form a five-dimensional feature vector. Similarly, a calibration procedure should be designed to interpret each feature vector as a meaningful evaluation score. We chose a SVR analysis for the calibration. The SVR is a version of support vector machine (SVM) for regression analysis which is based on the assumption of a linear regression function in a high dimensional feature space where the input data are mapped via a non-linear function (Basak et al., 2007). Compared with the ordinary least square regression, SVR can be used to fix both linear regression and non-linear regression tasks. Besides, it avoids multicollinearity problem between independent variables and reduce the impact of abnormal samples on model training thus offering more accurate model parameters. In this paper, we just briefly review the SVR analysis here as described in the literature (Basak et al., 2007; Ameri et al., 2014).

Assuming a set of training samples $D = \{(x_1, y_1), (x_2, y_2), ..., (x_L, y_L)\}$, where $x_i \in R^n$ represents input feature vectors and $y_i \in R$ represents the true MAS grades for all subjects. A non-linear projection $\phi(x)$ was applied for mapping the input data x to a higher dimensional space. Then the case of non-linear function f is described in the form as:

$$f(x) = w^T \phi(x) + b \tag{1}$$

the SVR can be formulated as:

$$\begin{aligned} &\min_{w,b,\xi_i^-,\xi_i^+} \tfrac{1}{2}||w||^2 + C\sum_{i=1}^{L}(\xi_i^- + \xi_i^+) \\ &s.t. f(x_i) - y_i \le \varepsilon + \xi_i^- \\ &\quad y_i - f(x_i) \ge \varepsilon + \xi_i^+ \\ &\quad \xi_i^- \ge 0,\ \xi_i^+ \ge 0 \end{aligned} \tag{2}$$

where C is a properly chosen penalty factor, and ξ_i^- and ξ_i^+ are slack variables representing upper and lower constraints on the outputs of the system. The Lagrangian multiplier method is used to solve the constrained optimization problem, and then the $f(x)$ can be written in the following form:

$$f(x) = \sum_{i=1}^{L} (\widehat{\alpha}_i - \alpha_i)\kappa(x, x_i) + b \tag{3}$$

where $\kappa(x_i, x) = \phi(x_i)^T \phi(x)$ is a radial basis function (RBF) kernel as follows:

$$\kappa(x_i, x) = \exp(-\frac{||x_i - x||^2}{2\sigma^2}) \tag{4}$$

The coefficients α_i and $\widehat{\alpha}_i$ in (3) are Lagrangian multiplier, and b is the bias constant determined by Karush-Kuhn-Tucker (KKT) conditions (Shevade et al., 2000).

In order to evaluate the spasticity for each subject, we similarly adopted a 24-fold leave-one-out cross validation. In this case, the input feature vectors $\boldsymbol{x}_i$ and clinical MAS grades y_i of any 23 subjects consisted of the training set, the remaining subject was used for testing to predict an evaluation score. During the training process, three important SVR parameters including penalty factor C, radius ε and a free parameter σ of RBF kernel needed to be determined toward optimal performance. Through some pretests, these parameters were empirically set at 30 for C, 0.09 for ε and 0.007 for σ. Finally, the well-trained SVR can be used as a tool to predict an evaluation score based on the input feature vector of any subject.

Performance Evaluation and Statistical Analysis

For each of 24 subjects, three evaluation scores were obtained for muscle spasticity assessment from the lambda model, the kinematic model and their combination, respectively. In fact, each of those scores can be regarded as an estimate of the MAS because they were calibrated with the same scale. Therefore, mean square error (MSE) between the evaluation scores and the corresponding clinical MAS grades was calculated to evaluate the performance of each of the three methods for spasticity assessment.

$$MSE = \frac{1}{L}\sum_{i=1}^{L}(Y_i - \widehat{Y}_i) \tag{5}$$

where L is the sample number, Y_i is the clinical MAS grades and $\widehat{Y}_i$ represents the assessment indicators given by the three assessment model. There were five subgroups of subjects with five MAS grades rating 0, 1, 1+ (1.5), 2, and 3 in this study.

In order to examine capability of three assessment methods in discriminating subjects/muscles with different MAS grades, a two-way ANOVA was performed on the evaluation score with the assessment method considered as within-subject factor (three levels) and the MAS grade as the between-subject factor (five levels). *Post hoc* pairwise multiple comparisons with Bonferroni correction were used. The significance level was set to 0.05 for all analyses. All statistical analyses were completed using SPSS software (ver. 16.0, SPSS, Inc., Chicago, IL, United States).

RESULTS

Figure 4 reports the experimental results from the lambda model. A representative example of the lambda model constructed for subject 1 was showed in **Figure 4A**. After excluding the outliers using 95% confidence region, the coefficient of determination R^2 between angular velocity and the DSRT in linear regression analysis increased from 0.86 to 0.97. The DSRT was approximately linearly decreased with the increase in stretch angular velocity, which was verified by the regression line $y = -0.277x + 46.765$. Therefore, the obtained TSRT for this subject was 46.765. **Figure 4B** reports the distribution of TSRT values for all subjects categorized by MAS grades, where the healthy subjects had the same TSRT value of 120 degrees. It can be found that the TSRT was generally declined with increased grades of spasticity, showing a negative correlation between the MAS and the TSRT with a correlation coefficient of −0.93 for all subjects and −0.73 for subjects with spasticity. Subsequently, **Figure 6** shows scatter plot of evaluation scores yielded by the lambda model versus the true MAS grades for all subjects. The evaluation score was regarded as a predicted MAS grades, and the MSE between those scores and corresponding MAS grades was 0.14. The linear regression analysis reported a high coefficient of determination R^2 up to 0.84 between the evaluation scores and MAS grades.

Figure 7 showed the constructed kinematic curves for a healthy subject and four subjects with spasticity at different clinical MAS grades. From the subject 5 with a MAS grade of 1+, as a representative example, the kinematic curve showed a different pattern with noticeable deviations from the reference curve from healthy controls. This is always the case for all subjects with spasticity. A weak or strong correlation was presented

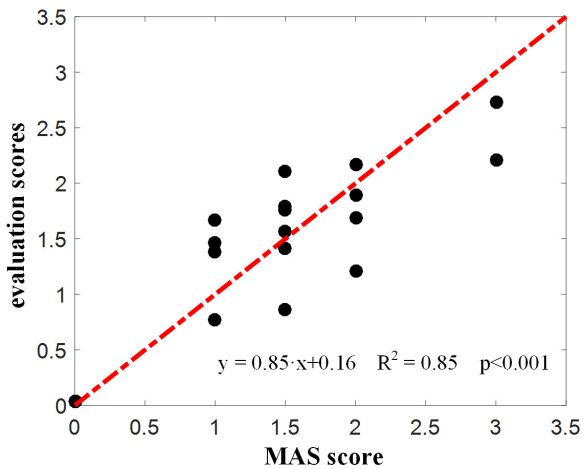


FIGURE 6 | Evaluation scores derived from the lambda model for all subjects.

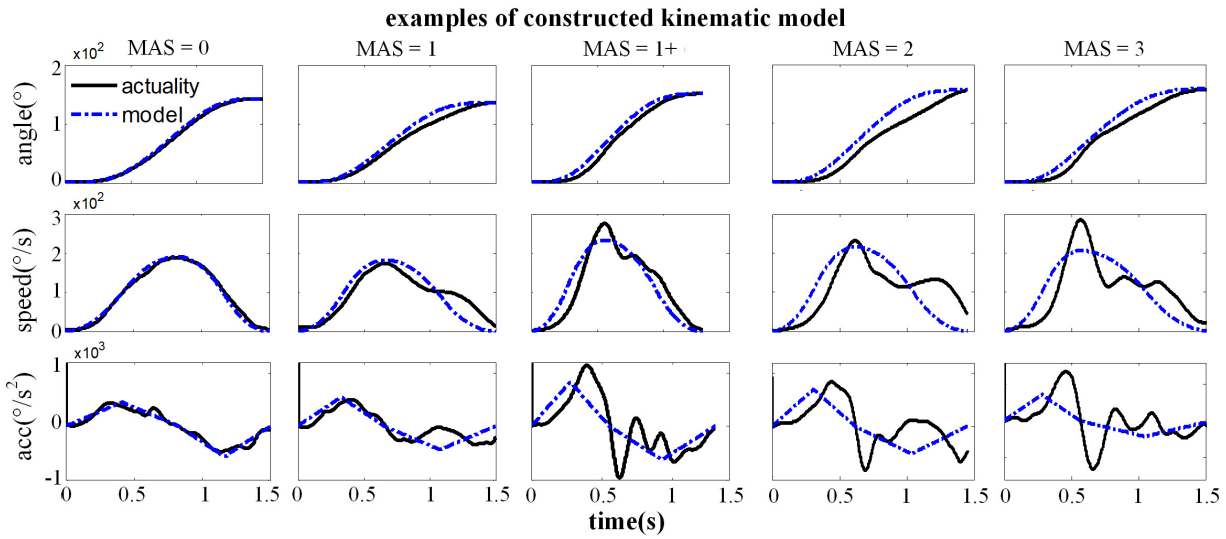


FIGURE 7 | Representative examples for the constructed kinematic variables of a healthy subject (MAS = 0) and four subjects with spasticity (MAS = 1∼3), including the angle, speed and acceleration curve. The black solid line was the actual kinematic curve and the blue dashed line was the constructed intended curve.

TABLE 3 | Correlation coefficients between each of biomarkers from the kenimatic model and the MAS grade.

Biomarkers		Correlation coefficient	*p*-value
Correlation coefficient	Acc*	−0.79	$p<0.001$
	Velocity*	−0.56	$p<0.001$
	Angle*	−0.26	$p>0.05$
MDF	Acc	0.64	$p<0.001$

Acc, velocity*, angle*: three biomarkers represent the correlation coefficients between the actual and reconstructed motion curves of angular acceleration, angular velocity and angle, respectively.*

between each of four biomarkers from the kinematic model and the MAS, with an absolute correlation coefficient ranging from 0.26 to 0.79, as exhibited in **Table 3**. Using the multivariate linear regression to calibrate the biomarkers from kinematic model, the final evaluation scores were reported in **Figure 8** for all subjects. This method predicted greatly varied evaluation scores for subjects with the same MAS grade. Specifically, it failed to discriminate the subjects with a MAS grade of 3 by its predicted evaluation score. As a result, the MSE was 0.47 for the kinematic model, and the evaluation score derived from this model showed moderate goodness of fit ($R^2 = 0.4990$, $p < 0.001$) to the MAS grades.

Figure 9 shows the evaluation scores obtained from fusion of both lambda and kinematic models using SVR for all subjects. The resultant evaluation scores appeared to positively correlate with the true MAS grades, with a regression line $y = 0.91x + 0.10$ whose goodness of fit was strong ($R^2 = 0.93$, $p < 0.001$), and a MSE of 0.059 was achieved. Furthermore, as compared with the single model, the combination of two models yielded the evaluation scores which are more distinguishable between any two MAS grades.

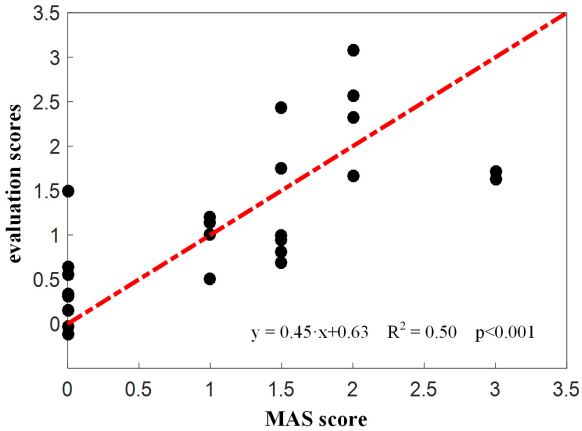


FIGURE 8 | Evaluation scores derived from the kinematic model for all subjects.

The ANOVA report no significant main effect of the evaluation method ($F = 0.42$, $p > 0.05$), but the spasticity degree (expressed as MAS grade) had a significant main effect on the evaluation scores ($F = 53.6$, $p < 0.001$). In addition, interaction between both factors was reported to be significant ($F = 3.13$, $p < 0.01$). Multiple pairwise comparisons further revealed that significant difference between any two MAS grades can be found except the comparisons between 1 and 1+, and between 1+ and 2, regardless of any method/model used.

DISCUSSION

This paper presents a novel framework for quantitative muscle spasticity assessment using supervised regression learning

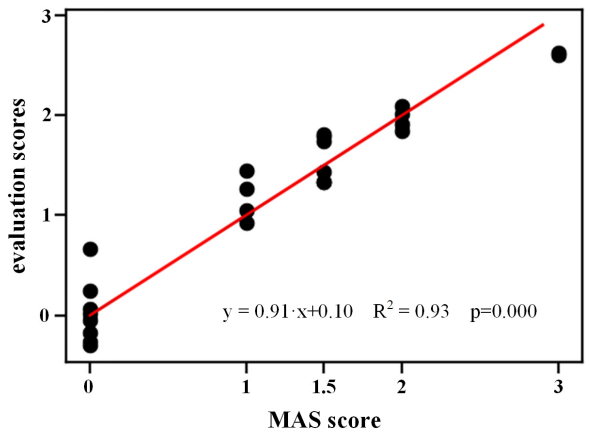


FIGURE 9 | Evaluation scores derived from combination of both models using SVR for all subjects.

methods to mine combined sEMG and inertial data. The evaluation procedure consisted of a series of passive elbow stretch with different (subjectively determined) angular velocities, during which no specialized instrument was required for constant speed control. This simplified and convenient procedure is suitable for daily manipulation for it is truly in accordance with clinical routine and acceptable for both medical professionals and patients. The wearable design of the arm-bands embedded small, low-cost sEMG sensors and IMUs can especially ensure the acquisition of both neural and kinematic data in a user-friendly way. In addition, the use of regression learning, as a typical group of machine learning, enables establishment of an expert system without assistance of any clinician or other professional to make assessment decisions based on input sensory data. All these features ensure a great potential of the proposed framework in clinical, community and even home-based rehabilitation.

Within the proposed framework, two different models namely the lambda model and the kinematic model were tested, respectively. The performance of any single model was re-evaluated given sEMG and IMU data from low-cost, wearable sensors. The biomarkers extracted from the two models showed a weak to significant correlation with the MAS grades. After the calibration procedure, the evaluation scores obtained from each model had the same meanings as the corresponding clinical MAS grades, with significant correlation for lambda model (R^2 = 0.84, p < 0.001) and strong correlation for kinematic model (R^2 = 0.50, p < 0.001) between them as well. All these findings were exactly consistent with those from previous studies (Feldman and Levin, 1995; Musampa et al., 2007; Calota et al., 2008). Such consistency not only confirmed the diagnostic efficiency of both models within the proposed framework in assessing muscle spasticity degrees, but also demonstrated that both models can work well under a specially designed protocol with simple testing procedures and essentially the same easily obtainable sensor apparatus for data recording. Therefore, under the proposed framework, the neurogenic component and non-neurogenic (mechanical) component of spasticity resistance can still be well-extracted and interpreted as potential quantitative evaluation descriptions using both models, respectively.

Please note that there are contradictory findings regarding the correlation between the lambda model and the MAS grade from the literature. Several studies have reported that this correlation was none or very poor (Levin et al., 2000; McGibbon et al., 2013), while others reported a good correlation (Kim et al., 2011). Consistent with the later report, our work exhibited a really good correlation with a coefficient of −0.93. The inconsistency between our study and others reporting no correlation can be mainly attributed into different technical details for data processing. For example, the TSRT of healthy individuals were conformably set at 120 due to the absence of typically evoked EMG associated with the spasticity. Exclusion of outliers was further implemented using a 95% confidence interval for the regression analysis, in order to overcome possible measurement errors given the portable sensors and manually controlled strength reflex. Both settings were specific to the lambda model conducted in our study. Another reason for explaining uncertain correspondence of the lambda model with the MAS is different way of their spasticity measurement. As discussed previously, the lambda model merely measured a certain aspect (i.e., neurogenic component) of the spasticity whereas the MAS emphasizes the spasticity-induced resistance to passive stretch of the muscle. If the neurogenic component becomes the primary contributor to muscle spasticity for a certain group of subjects, the lambda model is able to be highly correlated with the MAS, and otherwise, they may not agree very well. Both above-discussed reasons can also be used to explain result differences between the lambda model and kinematic model. It is suggested from the scatter plots (**Figures 6**, **7**) that they may not correspond well with each other. In detail, the lambda model worked well in assessing muscles/subjects with MAS grades of 0 and 3, but gives indistinguishable judgments among MAS grades 1, 1+, and 2. By contrast, the kinematic model presents better performance in discriminating MAS grades 1, 1+ and 2, but fails to discriminate MAS grades of 0 and 3. All these findings are regarded as evidences suggesting advanced performance and comprehensive description of spasticity by information fusion of both models.

Given the complementary capabilities of both models in assessing spasticity, their fusion is also supported by the proposed framework. Beside two single-model evaluation methods, the third method employed the SVR working with all biomarkers derived from both models to produce a comprehensive evaluation score. By comparing the three evaluation methods, the method with two-model combination was found to outperform other two methods, in terms of the lowest MSE value. This confirms the necessity of information fusion toward improved performance. In fact, the stretch reflex resistance has been regarded to be formed by both the neurogenic and mechanical components of the spasticity, which can be well-described by the lambda model and the kinematic model, respectively. This provides a very straightforward reason for explaining improved performance when both models were considered by the

proposed framework. Although each single model can be used alone for spasticity assessment with satisfactory performance, their combination further achieved improved performance with decreased MSE. Such a finding demonstrated feasibility and effectiveness of the proposed framework in combining both electrophysiological and kinematic information toward an advanced assessment of muscle spasticity. However, the ANOVA reported no significant difference between the three methods. The reason for explaining this could be attributed into the fact that all three evaluation methods within the proposed framework were calibrated into the same scale of the MAS. Despite of this, the ANOVA reported significant effect on the evaluation scores derived from these methods, indicating its capability of discriminating and quantifying different degrees of spasticity. In addition, comparing to the graded MAS grades, the evaluation method presented a more continuous distribution ranging from 0 to 3 (it was the maximum MAS for subjects recruited), which helps to enhance the accuracy and sensitivity of spasticity evaluation.

It is worth noticing that the proposed framework was applied to wearable low-precision sensors, and that an experiment protocol with free-hand manipulation was adopted in this study. As a result, the stretch velocities cannot be controlled and measured precisely. This would be expected to impact the efficacy of the proposed framework. In order to overcome these limitations, a series of trials with repeated passive elbow stretches were performed. It was also designed for ensuring sufficient data. Therefore, the proposed framework did not rely on data from any single trial, but mined these pooled low-precision data using machine learning algorithms to enhance the reliability and accuracy of spasticity evaluation. A representative example can be found in the lambda model. When quantifying a relationship between the dynamic stretch reflex threshold and angular velocity, a 95% confidence interval was designed in order to exclude the outliers. Finally, we found that the proposed framework for spasticity assessment works well on the low-precision data, yielding a high correlation to MAS grades, indicating the feasibility of using the simplified protocol in assessing muscle spasticity quantitatively. This also confirms the usability of the proposed framework.

However, there are still several limitations exist in this study. For all subjects, their MAS grades were taken as the regression target to calibrate and evaluate the computerized methods in this study. In another word, the MAS was regarded as the ground truth. However, the clinical MAS was a subjective evaluation with some well-known limitations, and thus it might not be considered to be absolutely accurate and trustable. For example, it is easy to misjudge the grade 1+ in the MAS. Therefore, the limitation in accuracy of the MAS might be the explanation of the failure in distinguishing the grade 1 and 1+, 1+ and 2 when implementing the proposed computerized method. In addition, due to the demand of sufficient data to ensure reliability of the evaluation method, a large number of repetitive trials are required. This protocol is a little time-assuming. The proposed framework employed relatively classic regression learning algorithms, many sophisticated learning algorithms such as deep learning may be beneficial to performance improvement, given recent development of machine learning techniques. Another key problem is the limited sample size used in this study. No subject with the maximal MAS grade of 4 was recruited due to inability of their tested muscles to complete the required experiment as a result of high stiffness. More samples would be benefit to enhance accuracy and reliability of the proposed framework according to supervised machine learning theory. Therefore, it is our expectation to enlarge more samples and to evolve the current work toward an expert system for quantitative spasticity assessment, suitable for clinical and family rehabilitation.

CONCLUSION

The proposed evaluation framework is in line with clinical measurement practice, thus providing convenient, objective and promotable spasticity assessment method in clinics. Therefore, it provides feedback about therapy effect of patients, which helps in the design and adjustment of effective and individualized rehabilitation protocols and plays an important role in spasticity management and intervention for clinical or family rehabilitation.

ETHICS STATEMENT

This study was approved by the Ethics Review Committee for Clinical Medical Research of First Affiliated Hospital of Anhui Medical University. All subjects offered informed and signed consent prior to any procedure of the experiment.

AUTHOR CONTRIBUTIONS

XuZ conceived the study and participated in the entire procedure of the study including experiments, data analyses, result interpretation, manuscript drafting and revisions. XT participated in the data analyses and interpretation, and performed substantial revisions of the manuscript. XiZ conducted experiments and participated in manuscript drafting. XG conceived the study and participated in result interpretation and manuscript revisions. XiC and XuC participated in data analysis and interpretation, and manuscript revisions. XuC also coordinated the study. All authors approved the final version of the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61771444).

REFERENCES

Ada, L., O'Dwyer, N., Ada, L., O'Dwyer, N., and O'Neill, E. (2006). Relation between spasticity, weakness and contracture of the elbow flexors and upper limb activity after stroke: an observational study. *Disabil. Rehabil.* 28, 891–897.

Ameri, A., Kamavuako, E. N., Scheme, E. J., Englehart, K. B., and Parker, P. A. (2014). Support vector regression for improved real-time, simultaneous myoelectric control. *IEEE Trans. Neural Syst. Rehabil. Eng.* 22, 1198–1209. doi: 10.1109/TNSRE.2014.2323576

Ardabili, N. S., Abdollahi, I., Khorramymehr, S., Shirzad, H., and Bahadorany, H. (2011). "Quantitative evaluation of spasticity at the elbow of stroke patients," in *Proceedings of the 18th Iranian Conference of Biomedical Engineering (ICBME)*, (Piscataway, NJ: IEEE), 131–136.

Bar-On, L., Aertbeliën, E., Wambacq, H., Severijns, D., Lambrecht, K., Dan, B., et al. (2013). A clinical measurement to quantify spasticity in children with cerebral palsy by integration of multidimensional signals. *Gait Posture* 38, 141–147. doi: 10.1016/j.gaitpost.2012.11.003

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Inf. Proc. Lett. Rev.* 11, 203–224.

Berndt, E. R., and Savin, N. E. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrica* 45, 1263–1277.

Biering-Sørensen, F., Nielsen, J. B., and Klinge, K. (2006). Spasticity-assessment: a review. *Spin. Cord* 44:708.

Blackburn, M., Van Vliet, P., and Mockett, S. P. (2002). Reliability of measurements obtained with the modified Ashworth scale in the lower extremities of people with stroke. *Phys. Ther.* 82, 25–34.

Calota, A., Feldman, A. G., and Levin, M. F. (2008). Spasticity measurement based on tonic stretch reflex threshold in stroke using a portable device. *Clin. Neurophysiol.* 119, 2329–2337. doi: 10.1016/j.clinph.2008.07.215

Condliffe, E. G., Clark, D. J., and Patten, C. (2005). Reliability of elbow stretch reflex assessment in chronic post-stroke hemiparesis. *Clin. Neurophysiol.* 116, 1870–1878.

Cooper, A., Musa, I. M., Van Deursen, R., and Wiles, C. M. (2005). Electromyography characterization of stretch responses in hemiparetic stroke patients and their relationship with the Modified Ashworth scale. *Clin. Rehabil.* 19, 760–766.

Feldman, A. G., and Levin, M. F. (1995). The origin and use of positional frames of reference in motor control. *Behav. Brain Sci.* 18, 723–744. doi: 10.1002/wcs.108

Fleuren, J. F., Voerman, G. E., Erren-Wolters, C. V., Snoek, G. J., Rietman, J. S., Hermens, H. J., et al. (2010). Stop using the Ashworth Scale for the assessment of spasticity. *J. Neurol. Neurosurg. Psychiatry* 81, 46–52.

Huang, S. C., Chen, J. J. J., Wang, Y. L., and Piotrkiewicz, M. (2004). Spasticity evaluation of hemiparetic limbs in stroke patients before intervention by using portable stretching device and EMG. *J. Med. Biol. Eng.* 24, 29–35.

Jobin, A., and Levin, M. F. (2000). Regulation of stretch reflex threshold in elbow flexors in children with cerebral palsy: a new measure of spasticity. *Dev. Med. Child Neurol.* 42, 531–540.

Kim, K. S., Seo, J. H., and Song, C. G. (2011). Portable measurement system for the objective evaluation of the spasticity of hemiplegic patients based on the tonic stretch reflex threshold. *Med. Eng. Phys.* 33, 62–69. doi: 10.1016/j.medengphy.2010.09.002

Lam, W. K., Leong, J. C. Y., Li, Y. H., Hu, Y., and Lu, W. W. (2005). Biomechanical and electromyographic evaluation of ankle foot orthosis and dynamic ankle foot orthosis in spastic cerebral palsy. *Gait Posture* 22, 189–197.

Lee, H. M., Huang, Y. Z., Chen, J. J., and Hwang, I. S. (2002). Quantitative analysis of the velocity related pathophysiology of spasticity and rigidity in the elbow flexors. *J. Neurol. Neurosurg. Psychiatry* 72, 621–629.

Levin, M. F., Selles, R. W., Verheul, M. H., and Meijer, O. G. (2000). Deficits in the coordination of agonist and antagonist muscles in stroke patients: implications for normal motor control. *Brain Res.* 853, 352–369.

Li, S., Shin, H., Zhou, P., and Li, X. (2017). Different effects of cold stimulation on reflex and non-reflex components of Poststroke spastic hypertonia. *Front. Neurol.* 8:169. doi: 10.3389/fneur.2017.00169

Lindberg, P. G., Gäverth, J., Islam, M., Fagergren, A., Borg, J., and Forssberg, H. (2011). Validation of a new biomechanical model to measure muscle tone in spastic muscles. *Neurorehabil. Neural Repair* 25, 617–625. doi: 10.1177/1545968311403494

McGibbon, C. A., Sexton, A., Jones, M., and O'Connell, C. (2013). Elbow spasticity during passive stretch-reflex: clinical evaluation using a wearable sensor system. *J. Neuroeng. Rehabil.* 10:61. doi: 10.1186/1743-0003-10-61

McGibbon, C. A., Sexton, A., Jones, M., and O'Connell, C. (2016). Quantification of elbow muscle tone from an instrumented manual stretch-reflex test. *Phys. Med. Rehabil. Res.* 1, 1–11.

Mehrholz, J., Wagner, K., Meissner, D., Grundmann, K., Zange, C., Koch, R., et al. (2005). Reliability of the modified tardieu scale and the modified ashworth scale in adult patients with severe brain injury: a comparison study. *Clin. Rehabil.* 19, 751–759.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, Vol. 821. Hoboken, NJ: John Wiley & Sons.

Mullick, A. A., Musampa, N. K., Feldman, A. G., and Levin, M. F. (2013). Stretch reflex spatial threshold measure discriminates between spasticity and rigidity. *Clin. Neurophysiol.* 124, 740–751. doi: 10.1016/j.clinph.2012.10.008

Musampa, N. K., Mathieu, P. A., and Levin, M. F. (2007). Relationship between stretch reflex thresholds and voluntary arm muscle activation in patients with spasticity. *Exp. Brain Res.* 181, 579–593.

Naghdi, S., Ansari, N. N., Abolhasani, H., Mansouri, K., Ghotbi, N., and Hasson, S. (2014). Electrophysiological evaluation of the Modified Tardieu Scale (MTS) in assessing poststroke wrist flexor spasticity. *NeuroRehabilitation* 34, 177–184. doi: 10.3233/NRE-131016

Nielsen, J. B., Crone, C., and Hultborn, H. (2007). The spinal pathophysiology of spasticity–from a basic science point of view. *Acta physiol.* 189, 171–180.

Pandyan, A. D., Johnson, G. R., Price, C. I. M., Curless, R. H., Barnes, M. P., and Rodgers, H. (1999). A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. *Clin. Rehabil.* 13, 373–383.

Pandyan, A. D., Price, C. I., Barnes, M. P., and Johnson, G. R. (2003). A biomechanical investigation into the validity of the modified Ashworth Scale as a measure of elbow spasticity. *Clin. Rehabil.* 17, 290–294.

Phadke, C. P., Ismail, F., and Boulias, C. (2015). Current challenges to clinical assessment of spasticity. *Int. J. Neurol. Res.* 1, 1–4. doi: 10.1586/ern.13.10

Pisano, F., Miscio, G., Del Conte, C., Pianca, D., Candeloro, E., and Colombo, R. (2000). Quantitative measures of spasticity in post-stroke patients. *Clin. Neurophysiol.* 111, 1015–1022.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Netw.* 11, 1188–1193. doi: 10.1109/72.870050

Silva, M. B., Silva, A. N., Naves, E. L. M., Palomari, E. T., and Soares, A. B. (2017). An improved approach for measuring the tonic stretch reflex response of spastic muscles. *Comput. Biol. Med.* 80, 166–174. doi: 10.1016/j.compbiomed.2016.12.001

Singh, P., Joshua, A. M., Ganeshan, S., and Suresh, S. (2011). Intra-rater reliability of the modified Tardieu scale to quantify spasticity in elbow flexors and ankle plantar flexors in adult stroke subjects. *Ann. Indian Acad. Neurol.* 14:23. doi: 10.4103/0972-2327.78045

Stecco, A., Stecco, C., and Raghavan, P. (2014). Peripheral mechanisms contributing to spasticity and implications for treatment. *Curr. Phys. Med. Rehabil. Rep.* 2, 121–127.

Szeto, G. P. Y., Straker, L. M., and O'Sullivan, P. B. (2005). EMG median frequency changes in the neck–shoulder stabilizers of symptomatic office workers when challenged by different physical stressors. *J. Electromyogr. Kinesiol.* 15, 544–555.

Trompetto, C., Marinelli, L., Mori, L., Pelosin, E., Currà, A., Molfetta, L., et al. (2014). Pathophysiology of spasticity: implications for neurorehabilitation. *BioMed Res. Int.* 2014:354906.

Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., and Yang, J. (2011). A framework for hand gesture recognition based on accelerometer and EMG sensors. *IEEE Trans. Syst. Man Cybernet. Part A Syst. Hum.* 41, 1064–1076.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Zhang, Tang, Zhu, Gao, Chen and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

TECHNOLOGY REPORT
published: 28 June 2019
doi: 10.3389/fnins.2019.00626

STEFTR: A Hybrid Versatile Method for State Estimation and Feature Extraction From the Trajectory of Animal Behavior

*Shuhei J. Yamazaki[1,2], Kazuya Ohara[3], Kentaro Ito[4], Nobuo Kokubun[4,5], Takuma Kitanishi[6,7,8], Daisuke Takaichi[9], Yasufumi Yamada[10], Yosuke Ikejiri[1,2], Fumie Hiramatsu[1], Kosuke Fujita[1†], Yuki Tanimoto[1†], Akiko Yamazoe-Umemoto[1], Koichi Hashimoto[11], Katsufumi Sato[12], Ken Yoda[13], Akinori Takahashi[4,5], Yuki Ishikawa[9], Azusa Kamikouchi[9], Shizuko Hiryu[10], Takuya Maekawa[3] and Koutarou D. Kimura[1,2]**

[1] Graduate School of Science, Osaka University, Toyonaka, Japan, [2] Graduate School of Natural Sciences, Nagoya City University, Nagoya, Japan, [3] Graduate School of Information Science and Technology, Osaka University, Suita, Japan, [4] Department of Polar Science, The Graduate University for Advanced Studies, Tachikawa, Japan, [5] National Institute of Polar Research, Tachikawa, Japan, [6] Department of Physiology, Osaka City University Graduate School of Medicine, Osaka, Japan, [7] Center for Brain Science, Osaka City University Graduate School of Medicine, Osaka, Japan, [8] Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Kawaguchi, Japan, [9] Graduate School of Science, Nagoya University, Nagoya, Japan, [10] Faculty of Life and Medical Sciences, Doshisha University, Kyotanabe, Japan, [11] Graduate School of Information Sciences, Tohoku University, Sendai, Japan, [12] Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan, [13] Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

Animal behavior is the final and integrated output of brain activity. Thus, recording and analyzing behavior is critical to understand the underlying brain function. While recording animal behavior has become easier than ever with the development of compact and inexpensive devices, detailed behavioral data analysis requires sufficient prior knowledge and/or high content data such as video images of animal postures, which makes it difficult for most of the animal behavioral data to be efficiently analyzed. Here, we report a versatile method using a hybrid supervised/unsupervised machine learning approach for behavioral *st*ate *e*stimation and *f*eature ex*tr*action (STEFTR) only from low-content animal trajectory data. To demonstrate the effectiveness of the proposed method, we analyzed trajectory data of worms, fruit flies, rats, and bats in the laboratories, and penguins and flying seabirds in the wild, which were recorded with various methods and span a wide range of spatiotemporal scales—from mm to 1,000 km in space and from sub-seconds to days in time. We successfully estimated several states during behavior and comprehensively extracted characteristic features from a behavioral state and/or a specific experimental condition. Physiological and genetic experiments in worms revealed that the extracted behavioral features reflected specific neural or gene activities. Thus, our method provides a versatile and unbiased way to extract behavioral features from simple trajectory data to understand brain function.

Keywords: quantitative behavioral analysis, behavioral states, feature extraction, navigation, calcium imaging, genetic screening

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Mark Shein-Idelson,
Tel Aviv University, Israel
Danny Eytan,
Technion Israel Institute of Technology, Israel

****Correspondence:***
Koutarou D. Kimura
kokimura-lab@umin.ac.jp;
kokimura@nsc.nagoya-cu.ac.jp

†Present address:
Kosuke Fujita,
Department of Ophthalmology, Graduate School of Medicine, Tohoku University, Sendai, Japan
Yuki Tanimoto,
RIKEN Center for Brain Science, Wako, Japan

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *18 December 2018*
Accepted: *31 May 2019*
Published: *28 June 2019*

Citation:
Yamazaki SJ, Ohara K, Ito K, Kokubun N, Kitanishi T, Takaichi D, Yamada Y, Ikejiri Y, Hiramatsu F, Fujita K, Tanimoto Y, Yamazoe-Umemoto A, Hashimoto K, Sato K, Yoda K, Takahashi A, Ishikawa Y, Kamikouchi A, Hiryu S, Maekawa T and Kimura KD (2019) STEFTR: A Hybrid Versatile Method for State Estimation and Feature Extraction From the Trajectory of Animal Behavior. Front. Neurosci. 13:626. doi: 10.3389/fnins.2019.00626

INTRODUCTION

The brain receives, integrates, and processes a range of ever-changing environmental information to produce relevant behavioral outputs. Therefore, understanding salient behavioral features can augment our understanding of important aspects of environmental information as well as of brain activity, which links the environmental information to behavior. Recent technological development of compact and inexpensive cameras and/or global positioning system (GPS) devices has facilitated convenient monitoring and recording of animal behavior (Dell et al., 2014; Egnor and Branson, 2016; Brown and de Bivort, 2018). However, the behavioral data generated through these approaches are frequently represented as a few simple measures, such as velocity, migratory distance, or the probability of reaching a particular goal, due to the challenges related to identification of specific aspects of behavior to be analyzed; in other words, it is still difficult to figure out how we can describe an animal behavior meaningfully (Berman, 2018). Owing to poor description of behavior, dynamic neural activity, for example, is not sufficiently interpreted even though simultaneous optical monitoring can measure a large number of time-series neural activities (Alivisatos et al., 2012; Landhuis, 2017). This large asymmetry in data richness between neural activity and behavior has emerged as one of the most significant issues in modern neuroscience (Anderson and Perona, 2014; Gomez-Marin et al., 2014; Krakauer et al., 2017).

One way to overcome the challenges in the appropriate descriptions of behavior is to describe its salient features via comprehensive analysis through an approach such as machine learning. Machine learning involves extracting latent patterns and uncovering knowledge from a large amount of data (Bishop, 2006). In fact, multiple behavioral analysis methods based on machine learning have been reported in the last decade (Baek et al., 2002; Stephens et al., 2008; Branson et al., 2009; Dankert et al., 2009; Brown et al., 2013; Kabra et al., 2013; Vogelstein et al., 2014; Wiltschko et al., 2015; Robie et al., 2017; Mathis et al., 2018). Most of these studies have classified behavioral states based on detailed analyses of animal postures as observed in video images (Dell et al., 2014); the classification of behavioral states into classes, such as foraging, sleeping, chasing, or fighting, is considered to be critical for efficient behavioral analysis, as each of the behavioral feature varies differently across different behavioral states (Patterson et al., 2008; Jonsen et al., 2013; Egnor and Branson, 2016). Although these methods have worked successfully for the analysis of behavioral videos of worms, fruit flies, and rodents in laboratories, they have some limitations. First, these methods are not suitable for analyzing relatively long-distance navigation given their requirement of recording reasonably large and detailed images of animals in the video frame. Second, the extraction of behavioral features from a state, as opposed to just state classification, is more critical in understanding how environmental information and/or brain activities trigger transitions among states for a behavioral response.

To analyze relatively long-distance navigation behavior comprehensively, we developed a method for the estimation of behavioral states and extraction of relevant behavioral features based only on the trajectories of animals. For estimating behavioral states, we used an unsupervised learning method involving the expectation maximization (EM) algorithm (Dempster et al., 1977) because it is difficult for the human eye to classify behavior into distinct states without using posture images. For extracting salient behavioral features, we used information gain, an index used for a supervised learning method (the decision tree analysis) (Quinlan, 1986), and compared the features between two different experimental conditions (e.g., with or without certain stimulus). It is because supervised learning is considered advantageous in the extraction of characteristic behavioral features and comparing them among multiple conditions. We named this hybrid supervised/unsupervised machine learning approach as the *st*ate *e*stimation and *f* eature ex*tr*action (STEFTR) method (**Figure 1**).

Because the STEFTR method only uses trajectory information for the analysis, it becomes possible to analyze the movement behavior of various animals regardless of the spatiotemporal scale of movement. As proof-of-principle experiments, we analyzed the trajectories of worms, flies, rats, and bats in laboratories and those of penguins and flying seabirds in the wild; these experiments involved a spatiotemporal scale ranging from mm to 1,000 km in space and from sub-seconds to days in time. The behavioral states of worms and penguins estimated by the STEFTR method were in reasonable conformation with the ones described in previous literature, supporting the reliability of our method. We further extracted learning-dependent behavioral features from a behavioral state of worms, in which one of the behavioral features is correlated with learning-dependent changes in neural activities. We also analyzed the behavioral features of mutant strains of worms and found that the patterns of features are correlated with gene function, suggesting that comprehensive feature extraction may enable us to estimate unknown functions of a gene product. We were also able to extract learning-dependent features from bats and pheromone-dependent features from fruit flies. Taken together, our findings indicate that the STEFTR method allows us to estimate internal state, neural activity, and gene function related to animal behavior only from movement trajectories, regardless of the recording method or the spatiotemporal scales.

MATERIALS AND METHODS

Overview of Behavioral State Estimation

For the analysis of trajectory information of an animal obtained from video images or from the GPS device attached to an animal, approximately 1/1,000 and 1/100 of the median recording time across animals were used as a unit for time frame and the time window for moving average, respectively (**Table 1**). These values were used to draw the eight histograms of the averages (*Ave*) and the variances (*Var*) of velocity (V), bearing (B), time-differential of V (dV) and B (dB) as the basic behavioral features. The time window for moving average was critical to reduce noise and to detect relatively long trends of behavior

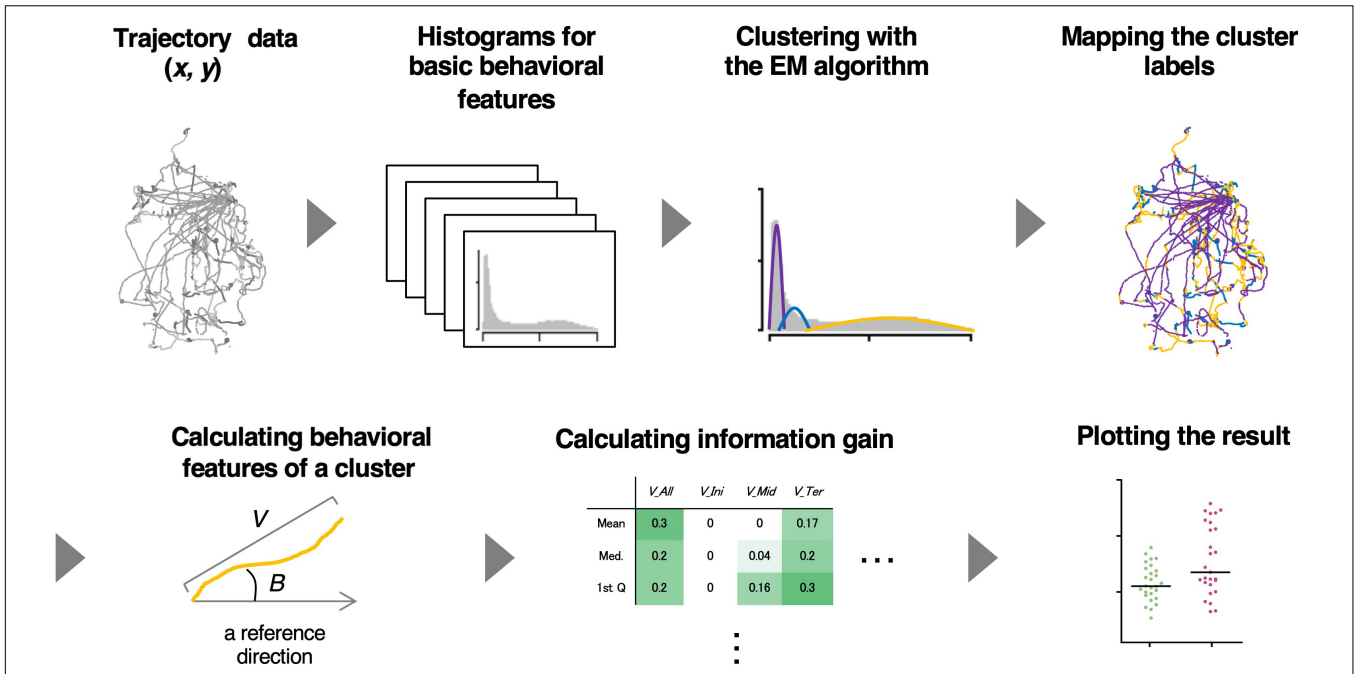


FIGURE 1 | A workflow of the STEFTR method. Trajectory data of animals are used to calculate eight basic behavioral features, and one of them is analyzed by the EM algorithm to estimate behavioral states (upper panels). From a behavioral state, behavioral features are comprehensively evaluated by using information gain (lower panels).

TABLE 1 | Summary of recording and analysis conditions of animal behavior.

Animal	Worms	Penguins	Flying seabirds	Rats	Flies	Bats
Condition	Naive, mock, preexposed	1	1	1	With or without pheromone	With or without learning
Animal #/condition	50 each	11	54	1	27, 30	3 each
Trajectory #/condition	50 each	11	54	45	27, 30	9 each
Recording time (min–max)	600 s	1,626–2,787 min	484–8,293 min	1,031–3,194 s	16 s	1.67–2.53 s
Recording time (median)	600 s	2,279 min	972 min	1,822 min	16 s	1.95 s
Time unit	1 s	1 min	1 min	1 s	0.25 s	0.008 s
Time window for moving average	12 s	20 min	10 min	20 s	–	–
Feature used for clustering	*dB_Var*	*V_Ave*	*B_Var*	*V_Var*	–	–
Length unit	mm	m	km	cm	a.u.	m
Approx. distance (min–max)	5–30 mm	1–10 km	100–2,000 km	180 × 180 cm	1–500 mm	6 m

(see section "Discussion"). Each histogram was regarded as a mixture of normal distributions, and EM algorithm was used to estimate the number of clusters (i.e., normal distributions) and their boundaries. We then calculated how many individual clusters in a histogram were well-separated using the separation index, and a histogram with the best separation index was chosen for further analysis (for details, see the following sections). The individual clusters in the chosen histogram are considered to be corresponding to different behavioral states. In the case of worms, the EM analysis was performed with specifying the maximum cluster number of 20. In other cases, maximum cluster number 5 was predetermined based on the knowledge that the number of basic behavioral states are several in general (Patterson et al., 2008).

The EM algorithm assigns a cluster label to each time frame although the clustering results should be smooth in time because each cluster should reflect a behavioral state of an animal. To smooth out the clustering results and removing outlying results, moving average was again applied to the cluster labels, which resulted in clusters resemble to the human-labeled behavioral states.

When the value of a basic behavioral feature changes suddenly and largely, the influence of the change may extend over a wide range. For example, if an animal moving straightly initiates local search suddenly, *dB* value will be 0°, 0°, 0°, 0°, 0°, 0°, 180°, 0°, 90°, 0°, 270°, etc. If moving average with ±5 time frame is applied, the value change occurs from −5 time frame of the sudden value change, which should be compensated.

Because worm's clusters 0 and 1 corresponded to this case, the beginning and the end of each cluster 0 was extended by the half of time window.

The cluster labels obtained as described above were mapped to the corresponding trajectory position with colors. We used a custom-made python program for calculating basic behavioral features, Weka data mining software (the University of Waikato, New Zealand) (Frank et al., 2016) for EM calculation, and Excel (Microsoft) for others.

EM Algorithm for Cluster Analysis

A set of values of the ith basic behavioral feature F_i (e.g., *V_Ave*), which were extracted from trajectories of interest, and the number of clusters N were given. We employed the EM algorithm to cluster F_i into N clusters, i.e., a mixture of N Gaussians. The probability distribution of the Gaussian mixtures M_N is represented as follows:

$$p(f_\mathrm{i}|M_N) = \sum_{n=1}^{N} \pi_n N(f_\mathrm{i}, \mu_n, \sigma_n),$$

where f_i is one of the feature values among F_i, π_n is the mixture weight of the nth Gaussian, μ_n is the mean of the nth Gaussian, and σ_n is the standard deviation of the nth Gaussian. The EM algorithm was used to estimate the cluster parameters: π_n, μ_n, and σ_n.

Determination of Cluster Number Using Log-Likelihood for Model

To find the best cluster number N, we evaluated a set of clusters obtained by the EM algorithm using log-likelihood of a set of feature values F_i under model M_n (n shows the cluster number). That is, (i) we performed the EM algorithm to obtain the cluster parameters for each n, which increased from 1 by 1. (ii) If the log-likelihood of F_i for M_n was increased compared with that for M_{n-1}, n was further increased. (iii) If not, the best N was determined as $n-1$. The Weka software (Frank et al., 2016) was used for this process.

Automatic Selection of Basic Behavioral Feature by Separation Index

To choose a histogram that best represents multiple behavior states, we calculated the "separation index" based on the following two criteria: (i) the distances among clusters were large and (ii) a peak of each cluster was apparent. The first criterion was calculated based on the overlapping area of different clusters. The second criterion was designed based on an idea that, when behavior clusters are not separated, the histogram (mixture of the clusters) seems to have no apparent peaks. Based on the above criteria, the separation index of a histogram was calculated as follows:

$$s(H) = (1 - \mathrm{Ov}(M_N)) + \frac{\min(N, \mathrm{Mx}(H))}{N},$$

where $\mathrm{Ov}(M_N)$ represents the proportion of overlapping area of adjacent Gaussian distributions in M_N, and $\mathrm{Mx}(H)$ represents the number of local maxima in the histogram H. To calculate local maxima, we first estimated the probability density function of H using the kernel density estimation and then calculated the derivative of the function. Note that, to eliminate noise local maxima, we ignored local maxima smaller than 0.1%. A histogram with the largest index was selected for further analysis.

Feature Extraction With Information Gain

We leverage information gain to evaluate the classification ability of each feature, i.e., its ability to identify a characteristic of a state (cluster). Information entropy is used to compute the ambiguity of a set of data points according to the following formula:

$$H = -\sum_{n=1}^{N} p_n \log_2 p_n,$$

where p_n is the proportion of data points belonging to the nth cluster. Given that we classify all the data points into two groups (i.e., two experimental conditions) using a particular threshold related to a specific feature, the feature is considered to be a characteristic feature (in that it classifies the data points well) if the ambiguity within the two groups is lower than that of the original data set. Thus, we evaluated features in terms of their ability in ambiguity reduction upon classification (information gain).

For worm odor avoidance behavior, we extracted behavioral features that have positive information gain in naive versus pre-exposed worms or in mock-treated versus pre-exposed worms (see below). Next, we chose the extracted features that were common for both comparisons; these features were termed as "features modulated in a learning-dependent manner." For flies, behavioral features were compared between with or without pheromone tapping. For bats, behavioral features were compared between unfamiliar flights (1st–3rd) and familiar (10th–12th) flights. The Weka software was used for these calculations.

Behavioral Parameters Included in a Feature Vector

For the machine learning analysis of worm's odor avoidance behavior, the following behavioral features were calculated for each cluster 0 segment from the coordinates of the centroid of the trajectory: velocity (V), bearing (B), odor concentration the worm experienced during the run (C), the time differential values for these (dV, dB, and dC), directedness (*Dir*) (Gorelik and Gautreau, 2014), curvature (called weathervane; *WV*) (Iino and Yoshida, 2009), and durations of cluster 0 and 1 (*Clst0Dur* and *Clst1Dur*, respectively). For V, dV, B, dB, C, dC, and *Dir*, the average (*Ave*) and median (*Med*) values for at the initiation (*Ini*), middle (*Mid*), termination (*Ter*), and all (*All*) periods of a cluster 0 segment were calculated. A total of 333 features was calculated by combining all these features.

For analyzing changes in the flight of bats, the following behavioral features in each flight were calculated from the coordinates of the bats and obstacles: three-dimensional flight velocity (V), horizontal and vertical bearings of the flight (*B_hori*

and *B_vert*, respectively), distance (*R_obs*) and bearing (*B_obs*) of the bat to the nearest edge point of the obstacle chain array, longitudinal directional distance to the frontal chain array (*R_x*), and lateral directional distance to the inside pitch of the chain array (*R_y*). Time-differential values were calculated for *V* (*dV*), *B* (*dB*), *dB* (*ddB*), and the flight height (*dH*), which were calculated with frame units of the high-speed video cameras (1/125 s). All flight trajectories were divided into three segments: early, middle, and late terms. The time window for the analysis of each behavioral feature was 0.1, 0.2, or 0.3 s before or while ($t = 0$) passing through the chain array. A total of 42 features was calculated by combining all these features.

Excel and Visual C# (Microsoft) were used for the calculations, while the Beeswarm package for R (The R Project) was used to obtain a scatter plot of the data.

Worms

The culture and handling of *Caenorhabditis elegans* strains were performed according to techniques described previously (Brenner, 1974). Wild-type Bristol strain RRID:WB-STRAIN:N2_Male and mutant strains RRID:WB-STRAIN:MT1219 *egl-3(n589)*, RRID:WB-STRAIN:VC671 *egl-3(ok979)*, RRID:WB-STRAIN:KP2018 *egl-21(n476)*, RRID:WB-STRAIN:CX4544 *ocr-2(ak47)*, RRID:WB-STRAIN:JC1636 *osm-9(ky10)*, and RRID:WB-STRAIN:FK127 *tax-4(p678)*, RRID:WB-STRAIN:MT6308 *eat-4(ky5)*, and RRID:WB-STRAIN:IK105 *pkc-1(nj1)* were obtained from the Caenorhabditis Genetics Center at the University of Minnesota, United States. The RRID:WB-STRAIN:KDK1 *dop-3(tm1356)* strain was originally obtained from the National BioResource Project (Japan) and back-crossed with the wild-type N2 strain five times.

A 2-nonanone avoidance assay was performed according to the protocol described previously (Kimura et al., 2010; Yamazoe-Umemoto et al., 2015). Briefly, two to three young adult hermaphrodite worms grown synchronously were placed in the center of a 9-cm nematode growth media (NGM) plate. Worm behavior was recorded for 12 min after 2-μL of 30% 2-nonanone (cat. no. 132-04173; Wako, Japan) diluted in 99.5% ethanol (cat. no. 0057-00456; Wako, Japan) were placed at two spots on the surface of the NGM plate. This assay was performed under the following three conditions: (1) Naive—the worms cultivated on 6-cm NGM plates with the RRID:WB-STRAIN:OP-50 bacteria as food were briefly washed with NGM buffer and subjected to the assay; (2) Pre-exposed—the worms were subjected to the assay after being pre-exposed to 0.6 μL of 15% 2-nonanone spotted on the lid of a 6-cm NGM plate for 1 h without food; and (3) Mock—the worms were subjected to the assay after being pre-exposed to ethanol similarly to the pre-exposed condition. We added the mock-treated control group to ensure that the starvation itself did not affect the odor avoidance behavior of worms and to extract behavioral features modulated by odor pre-exposure compared with the naive and mock-treated control groups. Images of worms on the 9-cm NGM plate during the odor avoidance assay were acquired by a high-resolution USB camera (DMK 72AUC02; The Imaging Source, United States) with a lens (LM16JC5MW; Kowa, Japan) at 1 Hz for 12 min. The coordinates of individual animals' centroids were acquired from the recorded images using the Move-tr/2D software (Library Co., Ltd., Tokyo, Japan) and used for the STEFTR analysis.

Similar to the other sensory behaviors of the worms, trajectories in the 2-nonanone avoidance behavior can be divided into two states: (1) "run"—a relatively long period of straight movement, and (2) "pirouette"—a period of short movements interrupted by frequent reversals and turns (Pierce-Shimomura et al., 1999; Kimura et al., 2010). The angular change per second was calculated from the centroid coordinates, and movements of 1 s with angular changes larger than 90° were classified as a turn. The histogram of turn intervals could be fitted to two exponentials, suggesting that the turn intervals are regulated by two probabilistic mechanisms (Pierce-Shimomura et al., 1999; Yamazoe-Umemoto et al., 2015). The time point at which the two exponentials intersected was defined as t_{crit} and turn intervals longer or shorter than the t_{crit} were classified as runs or included in pirouettes, respectively. The t_{crit} was calculated for the control (i.e., naive and mock-treated) condition for wild-type and mutant strains. Excel (Microsoft) was used for the above calculations. The odor concentrations that the worms experienced at specific spatiotemporal points were calculated according to the dynamic odor gradient model based on the measured odor concentration (Tanimoto et al., 2017; Yamazoe-Umemoto et al., 2018).

Statistical analyses were performed with Prism ver. 5.0 for Mac OSX (GraphPad Software, CA, United States) and R (The R Project). The sample size was determined based on the previous report (Yamazoe-Umemoto et al., 2015). A part of the original data used in this study had already been analyzed and published previously (Yamazoe-Umemoto et al., 2015), and re-analyzed with the STEFTR method.

Penguins

Fieldwork was performed on chick-rearing Adélie penguins *Pygoscelis adeliae* at Hukuro Cove colony (69°13′ S, 39°38′ E) in Lützow-Holm Bay, East Antarctica. GPS-depth loggers (GPL380-DT or GPL400-D3GT, weighing 55–85 g; Little Leonardo, Japan) were deployed among 11 penguins during the period between 27 December 2016 and 10 January 2017 and recovered from all the birds after 1–2 days. While the loggers were set to record GPS positions and depth every second, they could not record GPS positions when the penguins were diving. Therefore, we linearly interpolated the data, when necessary, to obtain GPS positions every 1 min before further analysis. See Kokubun et al. (2015) for methodological details. This fieldwork was carried out in accordance with the recommendations of the Law relating to Protection of the Environment in Antarctica. The protocol was approved by the Ministry of the Environment, Government of Japan. The sample size was not predetermined.

Flying Seabirds

Fieldwork was performed on streaked shearwater *Calonectris leucomelas* at Funakoshi-Ohshima Island (39°24′ N, 141°59′ E) between August and September in 2011, 2012, 2013, and 2015. We attached GPS loggers (GiPSy-2, 37 × 16 × 4 mm or GiPSy-4, 37 × 19 × 6 mm; TechnoSmArt, Rome, Italy) to the back feathers of chick-rearing streaked shearwaters with Tesa® tape

(Beiersdorf AG; GmbH, Hamburg, Germany) and cyanoacrylate glue (Loctite®401; Henkel Ltd., Hatfield, United Kingdom). The loggers were housed in waterproof heat-shrink tubing and set to record one fix per minute. The total mass of the unit was 25 g, which was less than 5% of the mean mass of the birds in accordance with the suggested load limit for flying seabirds. After approximately 2 weeks of deployment, we recaptured and retrieved the loggers. See Yoda et al. (2014) for methodological details. The study was carried out in accordance with the recommendations of the guidelines of the Animal Experimental Committee of Nagoya University. The protocol was approved by the Animal Experimental Committee of Nagoya University. Fieldwork was conducted with permission from the Ministry of the Environment and Agency for Cultural Affairs, Japan. The sample size was not predetermined. A part of the original data used in this study had already been analyzed and published previously (Yoda et al., 2014), and re-analyzed with the STEFTR method.

Rats

Locomotion data of an adult male Long Evans rat were obtained from the Collaborative Research in Computational Neuroscience (CRCNS; RRID:SCR_005608) data sharing website[1] (Hippocampus, hc-3 dataset, ec013 rat) (Mizuseki et al., 2014). The rat foraged for randomly dispersed water or foods on an elevated open field (180 cm × 180 cm) for 17–53 min. The rat's position was tracked by monitoring two light-emitting diodes mounted above the head with an overhead video camera at 30 Hz. The 30-Hz tracking data were resampled to 39.0625 Hz for offline processing. This study was carried out in accordance with the recommendations of the Regulations on Animal Experiments at Osaka City University. The protocol was approved by the Animal Care and Ethics Committee of Osaka City University. The sample size was not predetermined. The original data used in this study had already been analyzed and published previously (Diba and Buzsáki, 2008; Mizuseki et al., 2009, 2014), and re-analyzed with the STEFTR method.

Flies

Fruit flies *D. melanogaster* were raised on standard yeast-based media at 25°C and 40–60% relative humidity under a 12-h light/dark cycle. *Canton-S* flies aged between 6 and 8 days after eclosion were used as wild-type. After eclosion, the males were housed singly, while females were housed in groups until the experiment.

The locomotion measurement was performed as described previously with minor modifications (Kohatsu et al., 2011; Kohatsu and Yamamoto, 2015). Briefly, a male fly was tethered with a metal wire on its dorsal thorax and positioned over an air-supported Styrofoam ball (diameter, c. a. 6 mm). The locomotion trajectory of the fly was recorded by monitoring the rotations of the Styrofoam ball using an optical computer mouse sensor (BSMRU21BK; Buffalo Inc., Nagoya, Japan). The sensor detected the movements of the ball in the horizontal (Δx) and vertical (Δy) directions, which correspond to lateral and forward movements of the male fly, respectively. The Δx and Δy values, together with timestamps, were sent to a computer at 60 Hz via an Arduino Due microcontroller (Switch Science, Japan) with a custom sketch program. The 60-Hz data were down-sampled to 4-Hz data for the information gain analysis. The measurements were obtained at 25 ± 1°C and 50 ± 10% relative humidity and within 4 h after light onset.

Female pheromones were applied to the male fly by placing the female's abdomen in contact with the male's foreleg at the onset of the measurement. A manipulator (M-3333, Narishige, Tokyo, Japan) actuated a pipette with a volume of 200 μL (FUKAEKASEI Co., Ltd., China), in which a live female with her abdomen exposed toward a male fly was captured. We manually controlled the position of the manipulator to contact the female's abdomen to the male's foreleg. This contact procedure was omitted in the control experiments.

Visual stimulus was applied directly after pheromone application by starting horizontal movements of the female fly in front of the male fly as described (Kohatsu et al., 2011). The visual stimulus consisted of ten left-right horizontal movements of the female that lasted for 40 s. Each movement started with the female fly in the front of the male fly (i.e., center) and continued as the female fly moved left until it reached the left end of the rail (i.e., 5 mm away from the center), then moved right until it reached the right end of the rail (i.e., 5 mm away from the center), and ended when it came back to the center with a constant velocity of 5 mm/s. This movement was driven by a stepper motor (42BYG Stepper Motor, Makeblock Co., Ltd., Shenzhen, China) controlled by a custom sketch program (processing software version 3.3.7). We defined one round (4 s in total) as the movement related to the female starting to move away from the center, reaching the left end of the rail, passing the center, moving away to reach the right end of the rail, and coming back to the center again. Each of the "moving away" and "coming back" periods lasted for 1 s.

The Δx and Δy values with timestamps obtained in the final eight rounds were used for the analysis. To detect the characteristic parameters for the chasing behavior, we used Δx and Δy values during the period when the female was moving away from the male (2 s/round). We set the angle of the chasing behavior as 0 degree when the male moved forward. Angles between 0 and 90 degrees indicate that the male fly is moving toward the female that was moving away from the male.

As the parameters (velocity, bearing, and their time-differential values) were not normally distributed (Shapiro–Wilk test; see **Supplementary Table 13**), their values were compared between conditions (with/without pheromone) using the Mann–Whitney *U* test followed by Bonferroni correction for multiple comparisons. We used the Steel–Dwass test to compare values of the parameters between rounds. Statistical analyses were conducted using R software version 3.4.4. No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar to those in previous studies (Kohatsu et al., 2011; Kohatsu and Yamamoto, 2015).

Bats

Three adult Japanese horseshoe bats (*Rhinolophus ferrumequinum nippon*, body length: 6–8 cm, body mass: 20–30 g) were captured from natural caves in the Hyogo and Osaka prefectures in Japan as previously described

[1] https://crcns.org/data-sets/hc/hc-3

(Yamada et al., 2016). The bats were housed in a temperature- and humidity-controlled colony room [4 m (L) × 3 m (W) × 2 m (H)] with a 12-h light/dark cycle at Doshisha University in Kyoto, Japan, and were allowed to fly freely and given access to mealworms and water. Captures were conducted under license and in compliance with current Japanese law. This study was carried out in accordance with the recommendations of Principles of Animal Care [publication no. 86-23 (revised 1985) of the National Institutes of Health] and all Japanese laws. The protocol was approved by the Animal Experiment Committee of Doshisha University.

Methods for acoustic navigation measurement in bats have been described elsewhere (Yamada, 2017). Briefly, the experiments were conducted in a flight chamber constructed using steel plates [9 (length) × 4.5 (width) × 2.5 m (height)] under lighting with red filters (>650 nm) to avoid visual effects on the bats. An obstacle environment was constructed using plastic chains (diameter: 4 cm) suspended from the ceiling of the chamber. The chains were arranged at 15-cm intervals along the *x*-axis and at 22-cm intervals along the *y*-axis so that the bat was forced to fly in an S-shaped pattern without passing between chains. Three naive bats were observed for 12 continuous repeated flights so that their echolocation behavior in unfamiliar and familiar spaces could be compared. In this study, the first three flights were defined as unfamiliar flights, while the last three flights were defined as familiar flights.

The flight behavior of the bats was recorded at 125 frames/s using two digital high-speed video cameras (MotionPro X3; IDT Japan, Inc., Japan) placed in the left and right corners of the flight chamber. Based on a direct linear transformation technique, the successive 3D positions of the flying bats, as well as the locations of other objects, were reconstructed using motion analysis software (DIPPMotionPro ver. 2.2.1.0; Ditect Corp., Japan). The statistical calculations were performed with SPSS version 23 (IBM Corp.).

Calcium Imaging of Worm's Neurons

Calcium imaging of the worms' ASH neurons was performed according to the previous method with some modifications (Tanimoto et al., 2017). Briefly, transgenic strains expressing GCaMP3 (Tian et al., 2009) and mCherry (Shaner et al., 2004) in ASH sensory neurons under the *sra-6* promoter (KDK70034 and KDK70072; 20 ng/μl of *sra-6p::GCaMP3*, 20 ng/μl of *sra-6p::mCherry*, 10 ng/μl of *lin-44p::GFP*, 50 ng/μl of PvuII-cut N2 genomic DNA as a carrier in N2 background) were placed on an NGM agar plate on a robotic microscope system, OSB2 (Tanimoto et al., 2017). Although these transgenic worms were immobilized with the acetylcholine receptor agonist levamisole (Lewis et al., 1980) for high-throughput data acquisition through simultaneous imaging of multiple worms, the previous study revealed that the ASH activity is essentially unaffected by levamisole-treatment (Tanimoto et al., 2017). A constant gas flow of 8 mm/min was delivered, in which the mixture rate of 2-nonanone gas with air was changed to create a temporal gradient of odor concentration. The temporal change in odor concentration was measured by a custom-made semiconductor sensor before and after the series of calcium imaging experiments on each day. The fluorescence signals of GCaMP3 and mCherry in ASH neurons were divided into two channels using W-View (Hamamatsu, Japan), an image splitting optic, and captured by an electron multiplying charge-coupled detector (EM-CCD) camera (ImagEM; Hamamatsu, Japan) at 1 Hz. The intensities of fluorescence signals from cell bodies were extracted and quantified by ImageJ (NIH) after background subtraction. The average ratio over 30 s prior to the odor increase was used as a baseline (F_0), and the difference from F_0 (δF) was used to calculate the fluorescence intensities of GCaMP3 and mCherry ($F = \delta F/F_0$). The ratio between florescence intensities of GCaMP and mCherry (GCaMP/mCherry) was used in the figure.

RESULTS

Estimation of Behavioral States

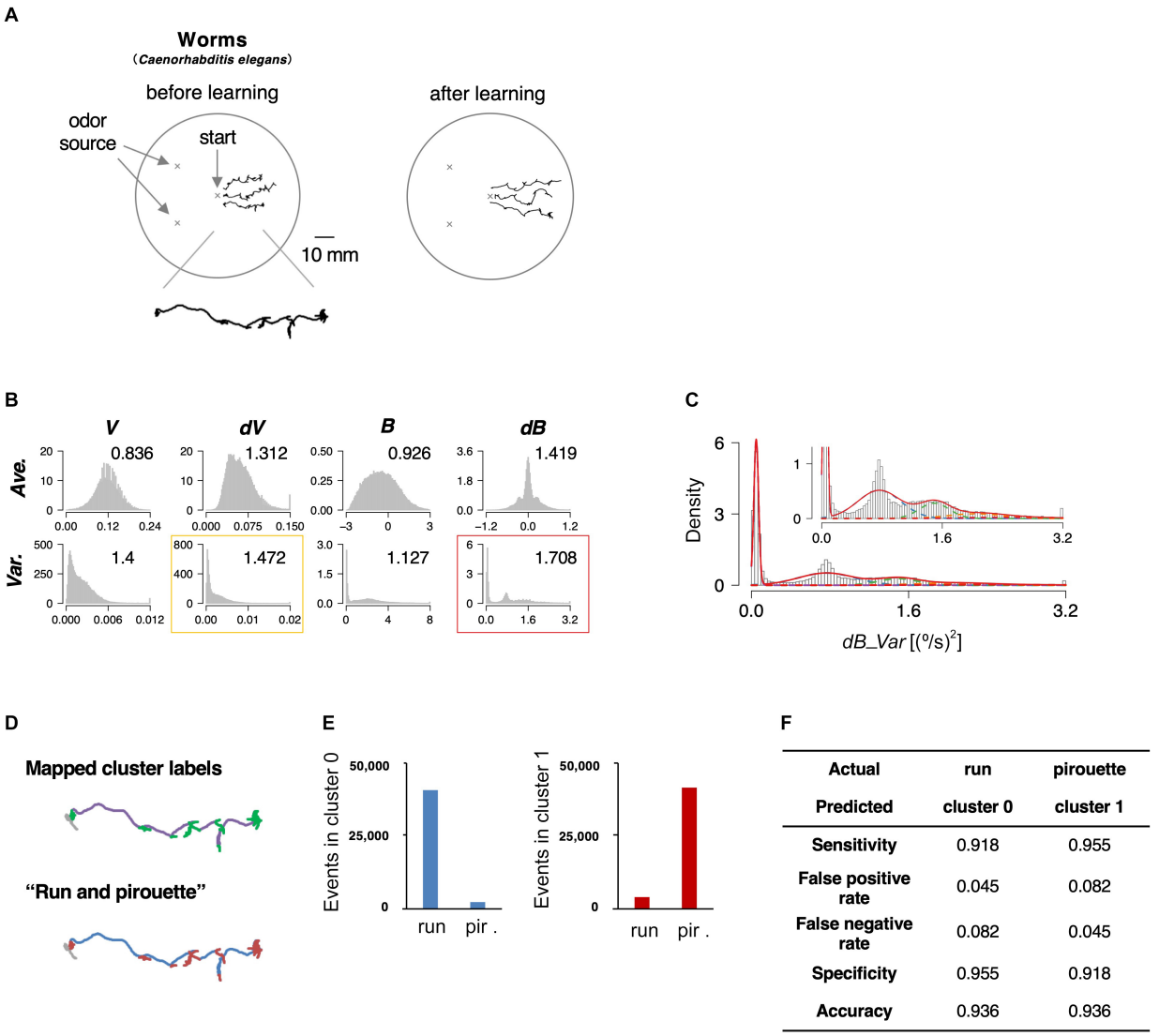
As the first part of the analysis, we classified the trajectory into several behavioral states based on the distribution of a basic behavioral feature. The behavior of animals consists of several states (Patterson et al., 2008; Jonsen et al., 2013), where basic behavioral features such as speed and direction change are likely distributed probabilistically with a center value that is optimal for each state. Thus, behavior can be more easily characterized when the behavioral features are analyzed for each state rather than for the entire behavior as a whole. In fact, classifying the trajectory into several states is one of the essential preprocessing steps in trajectory mining of people and vehicles in data science (Zheng, 2015).

For the state classification, we calculated the averages (*Ave*) and variances (*Var*) of four basic behavioral features: velocity (V), temporal changes in velocity (i.e., acceleration, dV), bearing (B), and temporal changes in bearing (dB). These eight features were represented in the form of histograms. Based on our hypothesis that values of basic behavioral features are likely distributed probabilistically in each state, we considered that histogram peaks may correspond to different behavioral states. One histogram was then regarded as a mixture of Gaussian distributions, and EM algorithm, an iterative method to estimate model parameters that maximize the likelihood of the model (Dempster et al., 1977), was used to estimate the average and variance values of each cluster. The separations of clusters in a histogram were evaluated by the separation index (see section "Materials and Methods"), and the best-separated histogram was chosen for further analysis. The individual clusters in the histogram were considered to correspond to different behavioral states. As proof-of-principle experiments, we analyzed the trajectories of worms and penguins, whose behavioral states have been studied previously using other methods (Pierce-Shimomura et al., 1999; Yoda et al., 2001).

The roundworm *Caenorhabditis elegans* has been used as a model animal for quantitative behavioral analysis owing to the ease of tracking behavior (movement for a few cm on agar surface can be easily recorded with an inexpensive high-resolution camera), optical monitoring neural activities, and genetic analyses and manipulations (De Bono and Maricq, 2005). Further, the neuronal wiring in *C. elegans* has been described in complete detail (White et al., 1986). In this study, we focused

on the avoidance behavior to the repulsive odor of 2-nonanone (**Figure 2A**, left) (Bargmann et al., 1993; Kimura et al., 2010). We chose this behavioral paradigm for the proof-of-principle experiment because the odor avoidance behavior has been quantitatively, although not fully, analyzed previously (Kimura et al., 2010; Yamazoe-Umemoto et al., 2015). The behavior of the worms was recorded with a USB camera for 12 min, and the position of each worm's centroid was extracted every second (**Table 1**).

We calculated the separation index in 8 basic features, and the variance values for bearing change with five clusters turned out to be the best (*dB_Var*, **Figures 2B,C**). Upon mapping the clusters on to the trajectories, we found that cluster 0 corresponds to relatively straight part, while the other clusters correspond to more complex parts (**Figure 2D**, upper panel). Cluster 0 and clusters 1–4 mainly corresponded to "run" and "pirouette," respectively, which are the classic two behavioral states of worms and have been found in multiple types of



F

Actual	run	pirouette
Predicted	cluster 0	cluster 1
Sensitivity	0.918	0.955
False positive rate	0.045	0.082
False negative rate	0.082	0.045
Specificity	0.955	0.918
Accuracy	0.936	0.936

FIGURE 2 | State estimation of worms. **(A)** Examples of the trajectories of three worms before (left) or after odor learning (right) in 12 min of 2-nonanone avoidance assay, overlaid on a schematic drawing of a 9 cm plate. One of the trajectories is magnified below. **(B)** The histograms of eight basic behavioral features. Horizontal and vertical axes indicate the values and the density of each feature. The numbers in each panel indicate the separation indices. The red and yellow rectangles indicate the first and second best separation indices, respectively. **(C)** Clustering *dB_Var* by the EM algorithm. Each cluster distribution (cluster 0, 1, 2, 3, and 4 are indicated by purple, blue, green, orange, and red dashed lines, respectively, similarly to the following figures) and the sum of clusters (red solid line) are shown. Inset is a magnified view. **(D)** Comparison of the cluster 0 and 1 (upper panel; purple and green, respectively) with run and pirouette (lower panel; blue and red, respectively) on a trajectory. The initial 2 min (gray in both panels) were excluded from the analyses because worms do not avoid the odor during the period (Kimura et al., 2010). **(E)** Event numbers of cluster 0 (left) and 1 (right) in run and pirouette. **(F)** Matching matrix of the state estimation shows the followings: sensitivity = TP/(TP + FN), false positive rate = FP/(FP + TN), false negative rate = FN/(FN + TP), specificity = TN/(TN + FP), accuracy = (TP + TN)/(TP + TN + FP + FN), where TP, true positive; TN, true negative; FP, false positive; FN, false negative.

sensory behavior (**Figure 2D**, lower panel) (Pierce-Shimomura et al., 1999; Lockery, 2011). "Run" constitutes a relatively straight movement, while "pirouette" is characterized by short, straight movement divided by frequent large changes in angle (turns and reversals). Runs and pirouettes are usually classified based on a threshold value for the duration between consecutive large angle changes (Pierce-Shimomura et al., 1999), unlike in this method. We collectively regarded clusters 1–4 as "cluster 1" because the positional information of worms does not appropriately reflect their actual locations during pirouettes due to insufficient spatiotemporal resolution of the recording system for relatively long-distance navigation, such as odor avoidance behavior (Yamazoe-Umemoto et al., 2015). We found that more than 90% of the cluster 0 and 1 corresponded with the run and pirouette, respectively (**Figures 2E,F** and **Supplementary Figure 1A**). Therefore, we concluded that the STEFTR method properly classified the odor avoidance behavior into distinct behavioral states. Although some cluster(s) do not appear to be Gaussian distributed (**Figure 2C**), the high matching rate supports the legitimacy of this method (see also the following sections for penguins).

Next, we applied the same process to the trajectories of penguins obtained using GPS devices. Penguins are good model wild animals for studying long-distance navigation given their relatively large body size and their habit of returning to a colony, which make the attachment and the recovery of GPS data easy (Yoda et al., 2001; Yoda, 2018). In this study, GPS and depth sensors were attached to 11 penguins from a colony in the Antarctic Continent; the depth sensors were used to evaluate the accuracy of state estimation (see below). The penguins moved by walking and swimming about 10 km for feeding, and each dataset contained up to 2 days of data recordings (**Table 1** and **Figure 3A**). Like in the case of worms, 8 basic behavioral features were extracted from the penguin trajectory data and represented as histograms. We chose the average velocity (*V_Ave*) as it showed the highest separation index (**Figure 3B**). The EM algorithm classified it into five clusters (**Figures 3C–E**, upper panel).

Interestingly, the clusters exhibited significantly different distributions in multiple behavioral features. For example, the values for the duration of each bout were much longer in cluster 0 than in clusters 1, 2, and 3 (**Figure 3E**, middle panel). In addition, although the clusters were classified only based on the horizontal velocities, the depths for cluster 0 and 1 were significantly closer to zero than those for clusters 2, 3, and 4 (**Figure 3E**, lower panel). These results are consistent with the idea that each cluster reflects a behavioral state that is a complex function of multiple behavioral features.

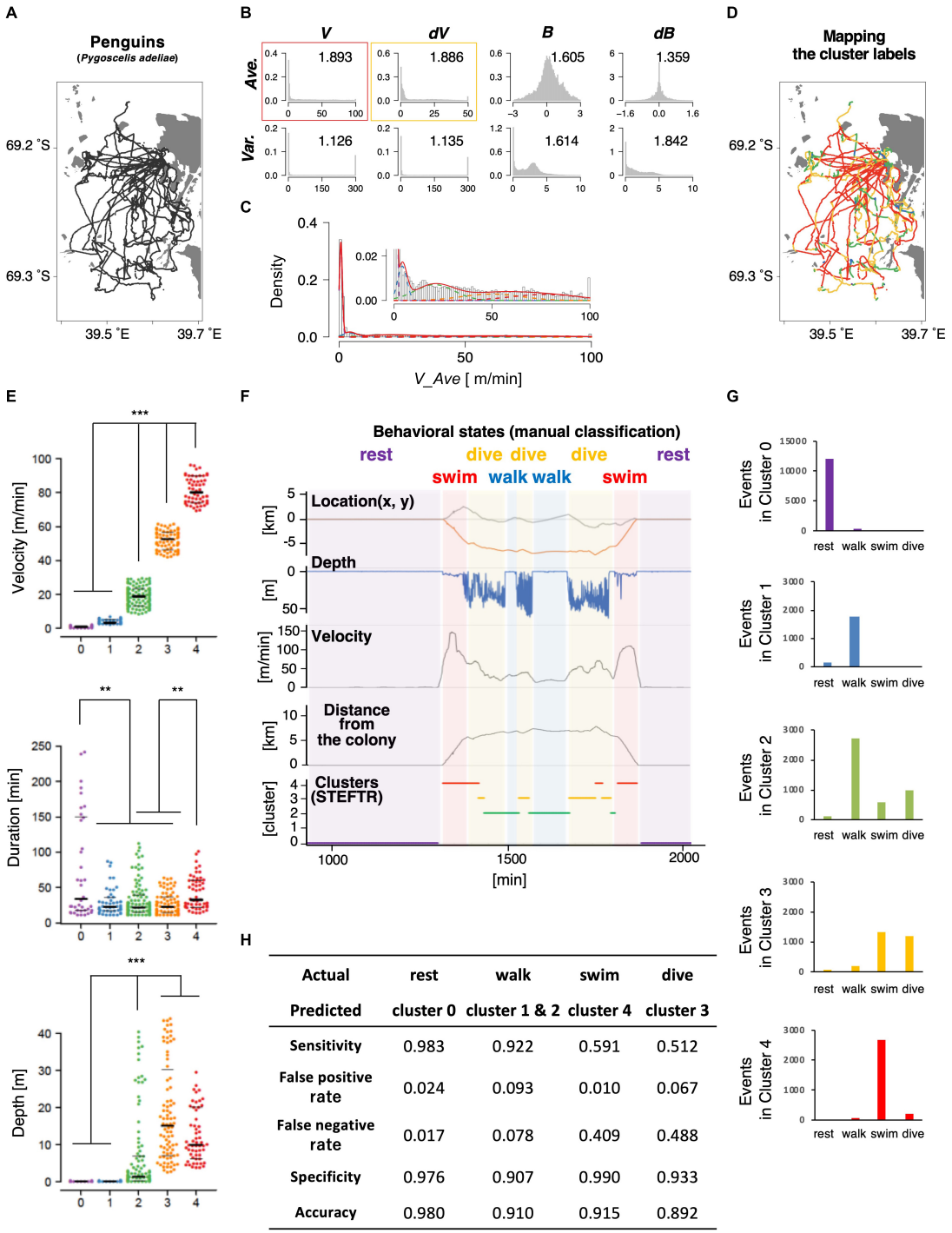
To evaluate whether the clusters actually reflect different behavioral states, we compare the results with the typical manual classification into four states (resting, transit by walking, transit by swimming, and diving), based on diving depth (from depth sensor), movement velocity, and distance from the colony (both calculated from the GPS positional information) (Yoda et al., 2001; Watanabe et al., 2012). Penguins stayed and rested at the colony (location does not change much; depth is zero), moved on land and ice mainly by walking (location changes relatively slowly; depth is near-zero), swim in the sea to go to the foraging area (location changes quickly; depth is relatively shallow in general and sometimes increases when they are moving toward the foraging area), dive deeply at the foraging area (location does not change much; dives occur continuously in bouts), and then come back to the colony by swimming and walking. The resting at the colony and swimming correlated with clusters 0 and 4, respectively (**Figures 3F,G** and **Supplementary Figure 2A**). In addition, most of clusters 1 and 2 correlated with walking, while about 50% of cluster 3 corresponded to diving (**Figure 3G** and **Supplementary Figure 2A**). Thus, when a behavioral state is classified to a cluster other than cluster 3, the penguin is likely to be resting, walking, or swimming. If a behavioral state is classified to cluster 3, which is ∼10% of all the behavior recorded, the penguin is either diving or swimming. Remarkably, although the clustering is only based on the trajectories of 11 penguins for a few days, the false positive rates were less than 10% and the sensitivity of the analysis was greater than 90% in all the cases (**Figure 3H**). Thus, we concluded that the STEFTR method can reasonably estimate different behavioral states only based on trajectory data.

In order to verify the evaluation process by the separation index, we compared the results of the state estimation with the first and second best-separated histograms for the worms and the penguins (indicated by red and yellow rectangles, respectively, in **Figures 2B**, **3B**). In both animal species, the estimated behavioral states based on the second best separation index exhibited a similar tendency to those of the best separation index, although the correlations between clusters and behavioral states were less clear (**Supplementary Figures 1B–E**, **2B–D**). These results strongly support the selection of the basic behavioral feature by the separation index.

The STEFTR method was also applied to the trajectories of flying seabirds in the wild and rats in the laboratory. The seabirds, *Calonectris leucomelas*, traveled ∼100 times longer distances (up to 1,000 km) with ∼10 times the speeds compared to penguins (Yoda et al., 2014; Matsumoto et al., 2017). For the animals, the variance of bearing (*B_Var*) with four clusters were chosen (**Figures 4A–C**). In the case of rats, the variance of velocity change (*dV_Var*) with four clusters was chosen (**Figures 4E–G**). In both animal species, significant differences among clusters were observed in duration and directedness, for example (**Figures 4D,H**). Such information can help ecologists estimate the candidates for feeding areas where fishes may be more densely distributed and discover biologically important marine areas. It can also help neuroscientists in estimating candidate conditions to further explore specific neural activities.

Comprehensive Extraction of Behavioral Features Modulated by Learning

As a second part of the STEFTR method, comprehensive feature extraction was performed by comparing a specific behavioral state in two different conditions, such as cluster 0 of worms before and after learning. Comprehensive semi-automated analysis can be very helpful to compare behavioral features in two conditions. This is because even when the overall result of behavioral responses is different in two conditions, it is still difficult to quantitatively determine which part of the trajectories are



H

Actual	rest	walk	swim	dive
Predicted	cluster 0	cluster 1 & 2	cluster 4	cluster 3
Sensitivity	0.983	0.922	0.591	0.512
False positive rate	0.024	0.093	0.010	0.067
False negative rate	0.017	0.078	0.409	0.488
Specificity	0.976	0.907	0.990	0.933
Accuracy	0.980	0.910	0.915	0.892

FIGURE 3 | State estimation of penguins. **(A)** The trajectories of 11 penguins (black lines) on the Antarctic Continent (gray area; white area is the sea). Horizontal and vertical axes indicate longitude and latitude, respectively. Base topography map for **Figures 3A,D** was downloaded from: http://antarctic.gsi.go.jp/download_01_25000.html#mapLayout. **(B)** The histograms of eight basic behavioral features (upper panels) and the classification by the EM algorithm. The numbers in each panel indicate the separation indices. The red and yellow rectangles indicate the first and second best separation indices, respectively. **(C)** Clustering *V_Ave* into five clusters. **(D)** Mapping of the five clusters on the trajectory. **(E)** Differences in velocity, duration, and depth among the clusters. Each dot represents a cluster bout, and the bars represent the median and the first and third quartiles. Significant differences among clusters suggest that the clusters correspond to different behavioral states. Statistical values were calculated using Kruskal–Wallis test with *post hoc* Dunn's test. $^{**}p < 0.01$, $^{***}p < 0.001$. **(F)** An example of comparison of the clusters from the STEFTR analysis with the behavioral states by manually classified labels, which is based on diving depth, movement speed recorded from GPS data, and distance from the colony. **(G)** Event numbers of each cluster. **(H)** Matching matrix of the state estimation. The statistical details are described in **Supplementary Table 1**.

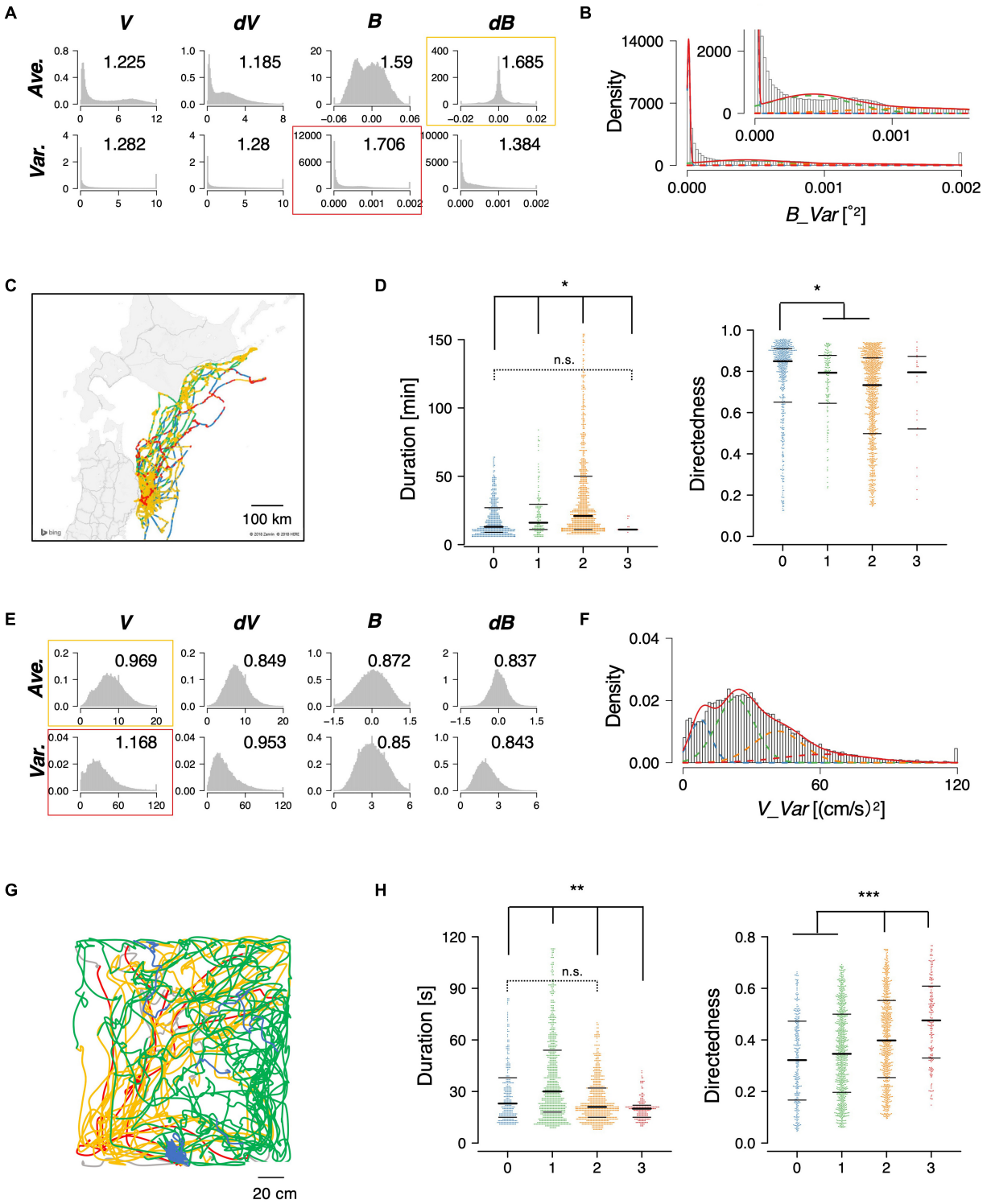


FIGURE 4 | State estimation of flying seabirds in the Pacific Ocean **(A–D)** and rats in the open maze **(E–H)**. **(A)** Eight basic features of flying seabirds. **(B)** Clustering *B_Var* into 4 clusters. **(C)** Mapping of the clusters on the trajectory. The gray region is the northern part of Japan (Tohoku and Hokkaido area), while the white region is the sea. **(D)** Significant differences were observed in duration (left) and directedness (right). **(E)** Eight basic features of rats. **(F)** Clustering *V_Var* into four clusters. **(G)** An example of trajectories of one rat. **(H)** Significant differences were observed in duration (left) and directedness (right). Each dot represents a cluster bout, and the bars represent the median and the first and third quartiles. Statistical values were calculated using Kruskal–Wallis test with *post hoc* Dunn's test. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The statistical details are described in **Supplementary Table 1**.

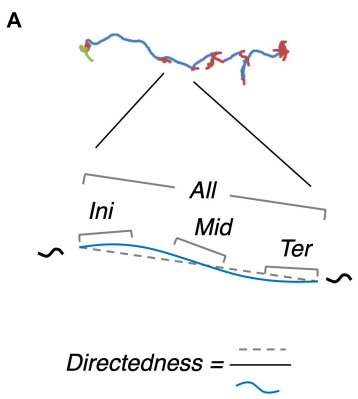
different (**Figure 2A, left** and **right**, for example). Furthermore, even if several behavioral features are found to be different, it is possible that other more prominent feature differences may exist. We considered that learning-dependent changes in behavior should be one of the best models for comprehensive feature extraction because the differences in behavioral features should reflect learning-dependent changes in neural/brain activities.

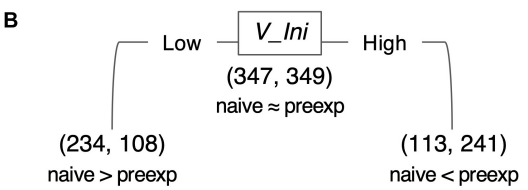
As a useful index for feature extraction, we chose information gain, the index for decision tree analysis (Quinlan, 1986). Binary decision tree analysis is for splitting a dataset into two sub-groups by automatically selecting the best feature and its parameter showing the largest information gain (i.e., difference of uncertainty, or "information entropy," between before and after division). Each data point is then classified into one of the sub-classes based on whether it has a larger or smaller value than an automatically determined threshold. When applied for binary classification, decision tree analysis automatically evaluates the classification performance of a large number of features as designed by the researchers. This analysis results in the extraction of certain features, allowing us to easily understand the utility of particular features in the classification. This approach is substantially different from those that employ support vector machines and/or deep neural networks, wherein the relationships between the classification and the features of the data cannot be easily discerned.

We first analyzed learning-dependent changes in worm odor avoidance behavior. Worm odor avoidance behavior is enhanced by pre-exposure to the odor as a type of non-associative learning, and pre-exposed worms migrate significantly longer distances from the odor source than control worms do during the same period (**Figure 2A**) (Kimura et al., 2010). This phenomenon is interesting because prior exposure to a stimulus generally causes a reduction, instead of enhancement, of the response to the stimulus through adaptation or habituation. Although this is a simple form of learning, this odor learning is modulated by multiple neuromodulators, including dopamine, octopamine (the worm counterpart of mammalian noradrenaline), and neuropeptides (Kimura et al., 2010; Yamazoe-Umemoto et al., 2015). Previous quantitative analyses have shown that the enhanced odor avoidance behavior is not caused by changes in speed, but mostly by increases in run duration (Yamazoe-Umemoto et al., 2015). However, this did not rule out the possibility that other behavioral features play more profound effects.

As an example of comprehensive feature extraction from a behavioral state, we focused on learning-dependent changes in cluster 0 (run) because the values of their centroid migration are quantitatively more reliable than cluster 1 (pirouette) as mentioned above. In addition to the basic behavioral features used for the estimation of behavioral states (*V*, *dV*, *B*, and *dB*), we also calculated directedness (*Dir*) (Gorelik and Gautreau, 2014), and the odor concentration (*C*) and temporal change in odor concentration (*dC*) that each worm experienced during the odor avoidance behavior; *C* and *dC* were calculated based on actual measurements of the dynamic odor gradient (Tanimoto et al., 2017; Yamazoe-Umemoto et al., 2018). For these, we calculated the initiation (*Ini*), middle (*Mid*), termination (*Ter*), and all (*All*) values of a cluster 0 segment (**Figure 5A**). In addition, different time windows (1–6 s in this case) were used to calculate these values because a behavioral feature could be apparent only within a specific temporal window [for example, velocity of run (i.e., cluster 0) starts decreasing 2 s before the end of a run (Pierce-Shimomura et al., 1999)]. We also calculated durations (*Dur*) of cluster 0 and 1, and the weathervane index (*WV*) (Iino and Yoshida, 2009). Information gain for each of these features was compared between naive/mock and pre-exposed conditions (**Figure 5B**, for example). The information gain values for each of the features have been summarized in **Table 2**, and the details are described in **Supplementary Tables 2–11**.

Through this analysis, we were able to find new as well as previously known behavioral features that exhibited learning-dependent changes. First, we found that the duration of each cluster 0 (*Clst0Dur*) exhibited higher information gain (**Table 2**), which corresponded to significantly increased cluster 0 duration (**Figures 6A,B**). This result is consistent with the findings from previous reports (Kimura et al., 2010;





Parameter	naive		preexp		entropy	Info gain
	Low	High	Low	High		
Original	347		349		1[a]	-
V_Ini	234	113	108	241	0.9017[b]	0.0983[c]

($c = a - b$)

FIGURE 5 | Feature extraction of worm behavior. **(A)** Schematic drawing of the behavioral features. **(B)** One example (*V_Ini*; average of time window 2) of calculation of information gain.

TABLE 2 | Learning-dependent features extracted from cluster 0 in odor avoidance behavior of wild-type worms.

Time	Value	*Clst0 Dur*	*Clst1 Dur*	*V*				*dV*				*B*				*dB*				*Dir*				*C*				*dC*				*WV*
		All	All	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All	Ini	Mid	Ter	All
–	Ave	0.098	0.057		–	–	–		–	–	–		–	–	–	0.072	–	–	–		–	–	–	0.163	–	–	–	0.088	–	–	–	
–	Med	–	–	0.040	–	–	–		–	–	–		–	–	–	0.102	–	–	–	–	–	–	–	0.160	–	–	–	0.088	–	–	–	
–	Q1st	–	–		–	–	–		–	–	–		–	–	–	0.065	–	–	–	–	–	–	–	0.164	–	–	–	0.107	–	–	–	
–	Q3rd	–	–		–	–	–		–	–	–		–	–	–	0.083	–	–	–	–	–	–	–	0.161	–	–	–	0.080	–	–	–	
–	Var	–	–		–	–	–		–	–	–		–	–	–		–	–	–	–	–	–	–		–	–	–	0.080	–	–	–	
–	1 %ile	–	–		–	–	–		–	–	–		–	–	–		–	–	–	–	–	–	–	0.163	–	–	–	0.090	–	–	–	
–	5 %ile	–	–		–	–	–		–	–	–		–	–	–		–	–	–	–	–	–	–	0.165	–	–	–	0.100	–	–	–	
–	95 %ile	–	–		–	–	–		–	–	–		–	–	–		–	–	–	–	–	–	–	0.153	–	–	–	0.108	–	–	–	
–	99 %ile	–	–		–	–	–		–	–	–		–	–	–		–	–	–	–	–	–	–	0.148	–	–	–	0.060	–	–	–	
1	Ave	–	–	–	0.114			–				–				–				–	–	–	–	–	0.147	0.162	0.162	–	0.116	0.093	0.104	–
2	Ave	–	–	–	0.120		0.049	–				–				–		0.048		–				–	0.148	0.166	0.163	–	0.111	0.105	0.122	–
3	Ave	–	–	–	0.109			–				–				–		0.049		–				–	0.151	0.163	0.162	–	0.110	0.099	0.128	–
3	Med	–	–	–	0.100		0.039	–				–				–				–	–	–	–	–	0.149	0.162	0.162	–	0.114	0.089	0.134	–
3	Var	–	–	–				–				–				–				–	–	–	–	–	0.111	0.087	0.041	–		0.060	0.050	–
4	Ave	–	–	–	0.099			–				–				–		0.050		–				–	0.147	0.166	0.165	–	0.082	0.098	0.122	–
4	Med	–	–	–	0.100			–				–				–				–	–	–	–	–	0.149	0.166	0.165	–	0.085	0.097	0.145	–
4	Var	–	–	–				–				–				–				–	–	–	–	–	0.102	0.083	0.040	–	0.057	0.082	0.082	–
5	Ave	–	–	–				–				–				–		0.048		–				–	0.146	0.163	0.166	–	0.090	0.112	0.148	–
5	Med	–	–	–	0.078			–				–				–				–	–	–	–	–	0.148	0.162	0.165	–	0.093	0.110	0.125	–
5	Var	–	–	–				–				–				–				–	–	–	–	–	0.097	0.092	0.054	–	0.061	0.112	0.078	–
6	Ave	–	–	–	0.072			–				–				–		0.056		–				–	0.148	0.164	0.146	–	0.083	0.112	0.169	–
6	Med	–	–	–	0.075			–				–				–				–	–	–	–	–	0.148	0.166	0.147	–	0.085	0.111	0.149	–
6	Var	–	–	–				–				–				–				–	–	–	–	–	0.089	0.095	0.060	–		0.116	0.081	–

Numbers are the sum of information gains obtained from the comparisons of naive vs. preexp. and mock. vs. preexp. Darker and lighter colors correspond to larger and smaller values, respectively. Blank cells, no value for information gain was obtained. –, not calculated.

Yamazoe-Umemoto et al., 2015), highlighting the reliability of this method. We also found that the velocity at the beginning of each cluster 0 (*V_Ini*) consistently exhibited higher information gain in the average and median values in multiple time windows (**Table 2**); these values were also significantly different in the pre-exposed worms as compared to the control worms (**Figure 6C**). The previous study has not identified this difference as only average values per run (i.e., cluster 0) have been calculated in the study (Yamazoe-Umemoto et al., 2015). Although the contribution of this behavioral feature to enhanced odor avoidance is unclear at present, our results indicate that the STEFTR method can reveal characteristic feature(s) under specific conditions, which is difficult for human analyses to accomplish.

Odor stimuli during runs, which likely drive the worms' odor avoidance behavior, were also found to be consistently modulated (**Table 2**). In fact, odor concentration (*C*) was significantly lower, and the temporal change in odor concentration (*dC*) was significantly closer to zero (i.e., shallower) in a learning-dependent manner (**Figures 6D,E**). Because the previous study demonstrated that worm odor avoidance behavior depends on *dC* rather than *C* at least in the naive condition (Tanimoto et al., 2017), one possibility is that the changes in the responsiveness of worms to *dC* are the underlying reason for the enhanced odor avoidance. However, it is also possible that the odor-experienced worms were somehow located farther away from the odor source than the unexperienced worms, and hence, sensed lower odor concentrations and shallower odor concentration change than the latter.

Responsiveness of Sensory Neurons to Odor Increase Was Modulated by the Odor Learning

If the change in sensitivity to *dC/dt* is the reason underlying enhanced odor avoidance behavior, it should be associated with changes in neural activity. Thus, we analyzed the responsiveness of a likely candidate, ASH nociceptive neurons (Kaplan, 1996; Bargmann, 2006). Previously, we have established the OSB2 microscope system that allows for *in vivo* calcium imaging of *C. elegans* neurons in the presence of odor stimuli resembling those that the worms experience during the odor avoidance behavior in the plates (**Figure 7A**) (Tanimoto et al., 2017). Using the OSB2 system, we found that ASH neurons are the major sensory neurons to cause pirouettes upon increases in 2-nonanone concentration (Tanimoto et al., 2017). However, whether the ASH response is modulated by 2-nonanone experience has not yet been studied.

We found that ASH responses were indeed modulated by prior odor experience. When the worms were stimulated with a 5 nM/s odor increase rate, which is the lowest rate of change to cause the threshold-level behavioral response in the previous study (Tanimoto et al., 2017), ASH neurons in naive as well as mock-treated worms exhibited robust responses (**Figure 7B**). However, the ASH responses were significantly reduced in the pre-exposed worms (**Figures 7B,C**). This suggests that prior odor experience causes a reduction in the neuronal response to a slight increase in odor concentration, subsequently causing longer run durations and enhanced odor avoidance behavior (**Figure 7D**).

Extracted Behavioral Features of Mutant Strains Correspond to Gene Function

Next, we comprehensively analyzed learning-dependent behavioral changes in the mutant *C. elegans* strains. Many mutant strains of *C. elegans* showing impaired learning have already been isolated and characterized (Bargmann, 2006; Sasakura and Mori, 2013), and the behavioral abnormalities observed in these mutants should reflect the role of the causal genes in neural function. In fact, we have previously shown that two different groups of genes involved in the enhanced odor avoidance behavior cause different abnormalities in behavioral features when mutated (Yamazoe-Umemoto et al., 2015). However, as the behavioral features exhibited by a mutant strain could be different from one another, identification of abnormal behavioral features is often laborious and time-consuming.

In addition to studying the previously described mutants with defective enhanced odor avoidance behavior (*egl-3* and *egl-21* for neuropeptide biosynthesis, and *dop-3* for dopamine receptor) (Kass et al., 2001; Suo et al., 2004; Yamazoe-Umemoto et al., 2015), we also analyzed mutant strains found to be involved in the phenomenon in this study (*ocr-2* and *osm-9* for TRP channels; *tax-4* for CNG channel; *eat-4* for vesicular glutamate transporter; *pkc-1* for protein kinase) (Land et al., 1994; Komatsu et al., 1996; Colbert et al., 1997; Lee et al., 1999; Tobin et al., 2002) (**Table 3**).

Neuropeptide mutant strains did not exhibit learning-dependent changes in behavioral features, except for the velocity of *egl-3(ok979)*. This result is consistent with the previous finding that neuropeptide signaling is required for the acquisition of odor memory (Yamazoe-Umemoto et al., 2015). *egl-3(ok979)* may have exhibited stronger phenotypes than *egl-3(n589)* because they are non-sense and missense mutants, respectively. Also consistent with the previous report (Yamazoe-Umemoto et al., 2015), the *dop-3* mutants exhibited abnormalities in direction-related behavioral features (*B* and *Dir*) while the changes in cluster 0 durations and velocities are similar to those of wild-type worms (**Table 3**). Furthermore, with respect to the newly added mutant strain, similar patterns are observed in *ocr-2* and *osm-9* mutants of the TRP channel involved in sensory perception. On the other hand, *tax-4*, which is also involved in sensory perception but expressed in a different set of sensory neurons (Komatsu et al., 1996; Tobin et al., 2002), and *eat-4* and *pkc-1* mutants showed different patterns of abnormalities. Taken together, our results suggest that the patterns of features extracted from mutant strains may reflect functional groupings of the mutated genes. Thus, profiling and classification of extracted mutant features of unknown genes may be useful in the estimation of their physiological functions.

Feature Extraction of Fly Sexual Behavior

Next, we applied the technique to comprehensive feature extraction of animal behavior under specific situations in two different conditions—heterosexual chasing behavior of *Drosophila melanogaster* with or without pheromone sensation.

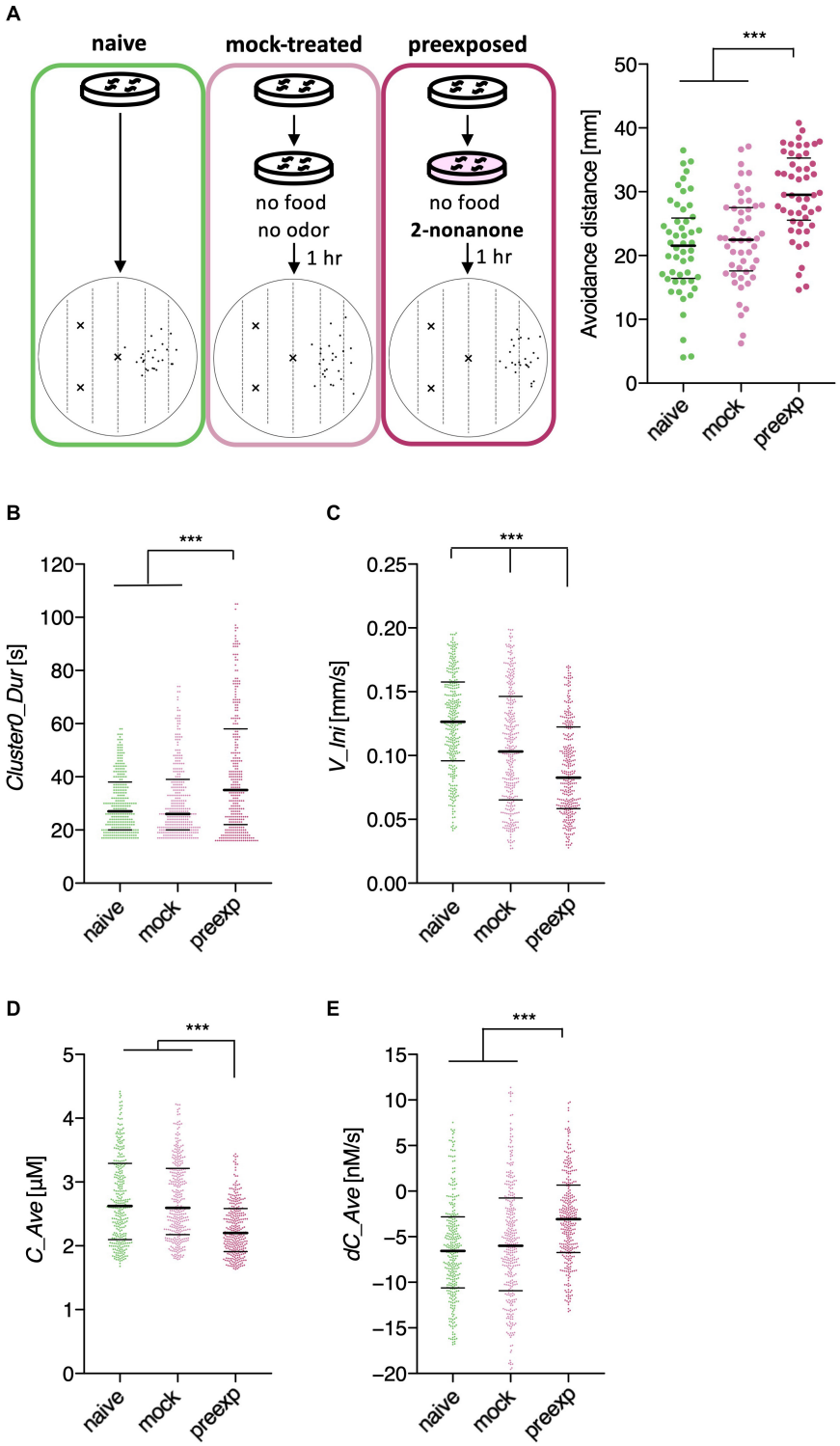


FIGURE 6 | Extracted features modulated by the odor learning. **(A)** Enhanced odor avoidance behavior in worms caused by odor pre-exposure. Left: End-points of 25 worms in each condition plotted on a schematic representation of the assay plate. Right: Avoidance distance (distance between the center line of the plate and the end-point of the behavior) of each worm. Each dot represents a worm. Significant differences were observed between the pre-exposed worms and the naive and mock-treated worms (***$p < 0.001$, Kruskal–Wallis test with *post hoc* Dunn's test). **(B–E)** Distributions of extracted features. Duration **(B)**, the initial value of velocity (**C**; average of time window 2), the average odor concentration **(D)**, and the average odor concentration change **(E)** of each run (***$p < 0.001$, Kruskal–Wallis test with *post hoc* Dunn's test). Each dot represents a cluster bout, and the bars represent the median and the first and third quartiles. The statistical details are described in **Supplementary Table 1**.

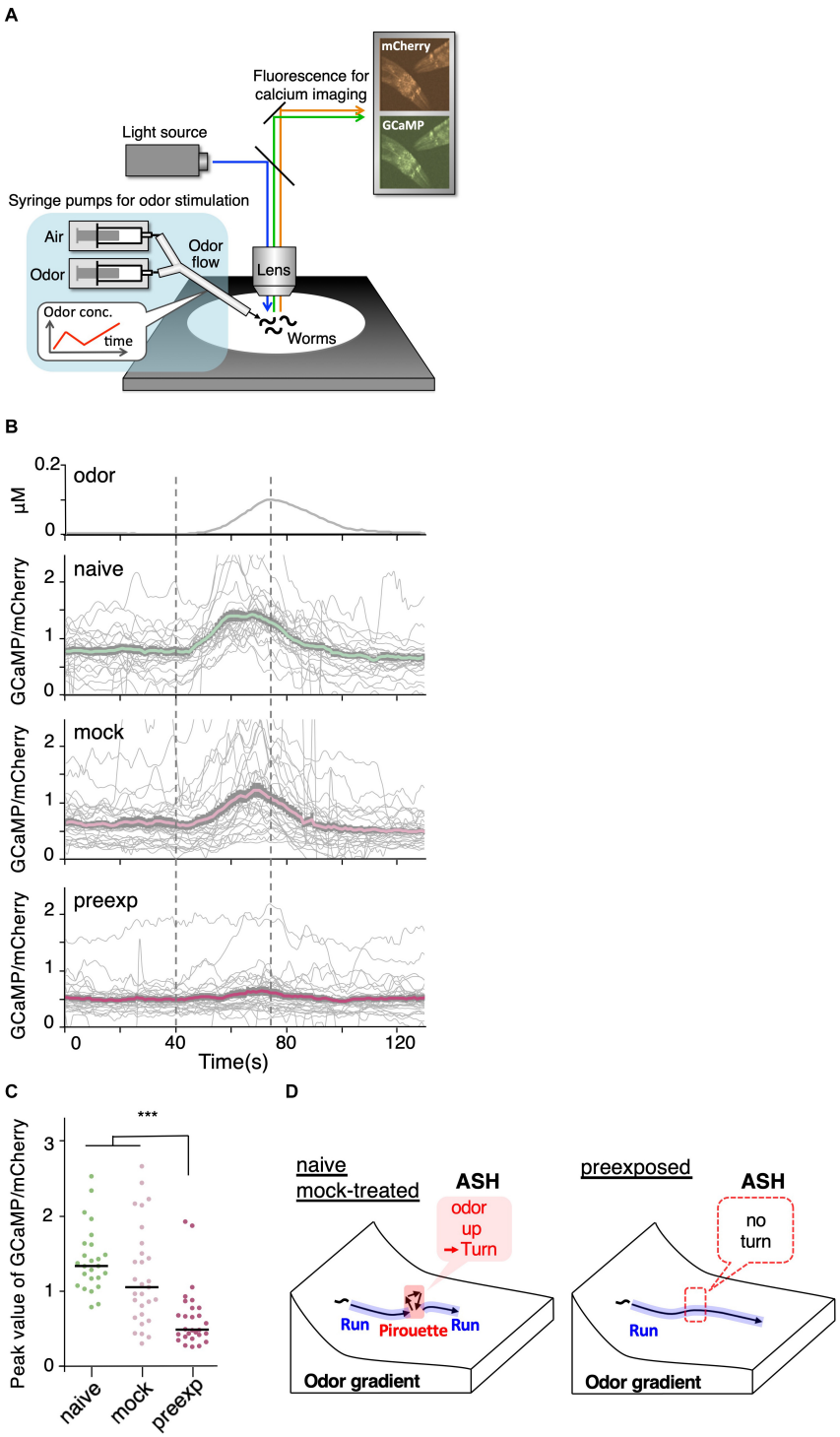


FIGURE 7 | Sensory responses to slight increases in odor concentration were reduced by pre-exposure to the odor. **(A)** A schematic drawing of calcium imaging of neural activity of worms under odor stimuli. Several immobilized worms were simultaneously exposed to an odor flow whose concentration was changed by controlling syringe pumps. **(B)** Responses (GCaMP/mCherry) of ASH neurons in naive (n = 25), mock-treated (n = 29), and pre-exposed (n = 26) worms. Thick lines with gray shadows indicate mean ± standard error of the mean, while thin lines indicate individual responses. **(C)** Distributions of peak values during the odor-increasing phase (t = 40–80 s) shown in **(B)**. The bars represent the median (***$p < 0.001$, Kruskal–Wallis test with *post hoc* Dunn's test). **(D)** A model relationship between odor concentration change and behavioral response during navigation along the odor gradient. When naive and mock-treated worms sensed a slight increase in the odor concentration, which is a sign of migrating in the wrong direction, they stopped a run and started a pirouette to search for a new direction. In contrast, the pre-exposed worms did not respond to a slight increase in odor concentration, leading to longer run durations (and shorter pirouette durations in total as a consequence), which likely contribute to the enhanced avoidance distance. The statistical details are described in **Supplementary Table 1**.

TABLE 3 | Patterns of learning-dependent behavioral features of cluster 0 in odor avoidance behavior of mutant worms.

Strain	Clst0 Dur	Clst1 Dur	V				dV				B				dB				Dir				C				dC			
	All	*All*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*	*All*	*Ini*	*Mid*	*Ter*
Wild-type	Up	D/U	Down	Down		Down									Down		Down						Down	Down	Down	Down	Up	Up	Up	Up
egl-3(n589)																							Up	Up	Up	Up				
egl-3(ok979)			Down	Down	Down	Down																	Up	Up	Up	Up				
egl-21																														
dop-3	Up		Down		Down	Down					Up	Up	Up					Up	Down			Down					Up		Up	Up
ocr-2			Up			Up																								
osm-9	Up		Up	Up	Up	Up																								
tax-4						Down									Down	U/D							D/U	D/U	D/U	D/U				
eat-4			Down	Down		Down																	Down							
pkc-1															Down		Down						Up				Down			

Darker and lighter colors mean larger and smaller information gain values, respectively. "Up" and "Down" indicates the value changes of pre-exposed worms compared with naive and mock-treated worms. "U/D" or "D/U" indicates the value of preexposed worms is higher than naive and lower than mock-treated worms, and vice versa. Blank cells, no value for information gain was obtained. The details of information gain are described in ***Supplementary Tables 2–11****.*

On an experimental tracking system (**Figure 8A**), male flies chased the target female flies' abdomens after tapping them with their forelegs to sense the cuticular pheromone, although males do not show such chasing behavior before tapping (Kohatsu et al., 2011). While this pheromone-driven behavior has been generally used for the observation of neural activity during courtship behavior in fruit flies, the behavioral features have not yet been elucidated comprehensively.

In this study, we used a tracking system as described in previous studies, where a male exhibited the moving female abdomen with eight times left-right round trip after a pheromone sensation (**Figure 8A**) (Kohatsu et al., 2011). We detected positive information gains in the velocity, changes in velocity, bearing, and changes in bearing in a pheromone sensation-dependent manner (**Figure 8B**). Unexpectedly, the information gains of velocity were higher in the earlier round trips and decreased over the trips (**Figure 8B**). This suggests that the pheromonal effect promoting chasing behavior decreases over time. To confirm the result of the STEFTR method, we re-analyzed the speed of the male locomotion along with time-series. Consistent with the STEFTR result, the velocity of males that had tapped the female significantly decreased over the trips (orange group in **Figure 8C**; significant differences between rounds 1, 4, and 8), whereas that of control flies (gray group in **Figure 8C**) remained mostly unchanged (significant difference only between rounds 1 and 5). Thus, the STEFTR method can even uncover behavioral features that fluctuate over time. The decreased tracking velocity may reflect a decrease in motivation in the fly brain, which can be assessed directly by observing the temporal changes in neuronal activity related to the courtship-motivation circuit in the fly brain (Yamamoto and Koganezawa, 2013; Zhang et al., 2018).

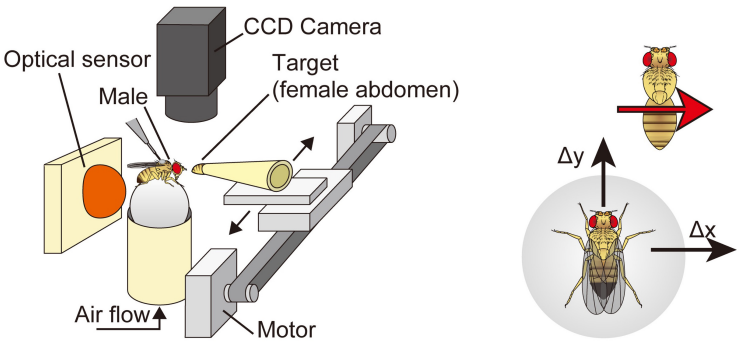
Feature Extraction of Learning-Dependent Modulation of Acoustic Navigation in Bats

To further demonstrate the general applicability of the method, we examined features of acoustic navigation in bats. We have previously reported that bats improve their flight trajectory in an indoor space with obstacles in a learning-dependent manner (Yamada, 2017). Here, we optimized features such as velocity (V), distance to the obstacle chain array (R_obs and R_x), and horizontal bearing of the flight (B_hori) for the experimental paradigm (**Figures 9A,B**). Interestingly, although the velocity (V) itself was modulated by flight experience, the change in velocity (dV) was not (**Figures 9C,D**), suggesting that bats determine flight speed before initiating navigation, but not during navigation, at least in this experimental condition. As the vocalizations of bats reflect their attention or decisions (Moss and Surlykke, 2010), our results suggest that the STEFTR method can be used to elucidate such higher brain functions during navigation.

DISCUSSION

Measuring and analyzing behavior is one of the most prominent steps in understanding brain function. In order

A



B

	V	*dV*	*B*	*dB*
Overall	0.279	0.260	0.146	0.176
Round 1	0.254	0.200	0.038	0.201
Round 2	0.268	0.218	0.177	0.277
Round 3	0.222	0.171	0.070	0.214
Round 4	0.175	0.235	0.184	0.204
Round 5	0.206	0.229	0.216	0.224
Round 6	0.168	0.221	0.118	0.240
Round 7	0.165	0.188	0.103	0.158
Round 8	0.027	0.087	0	0.062

C

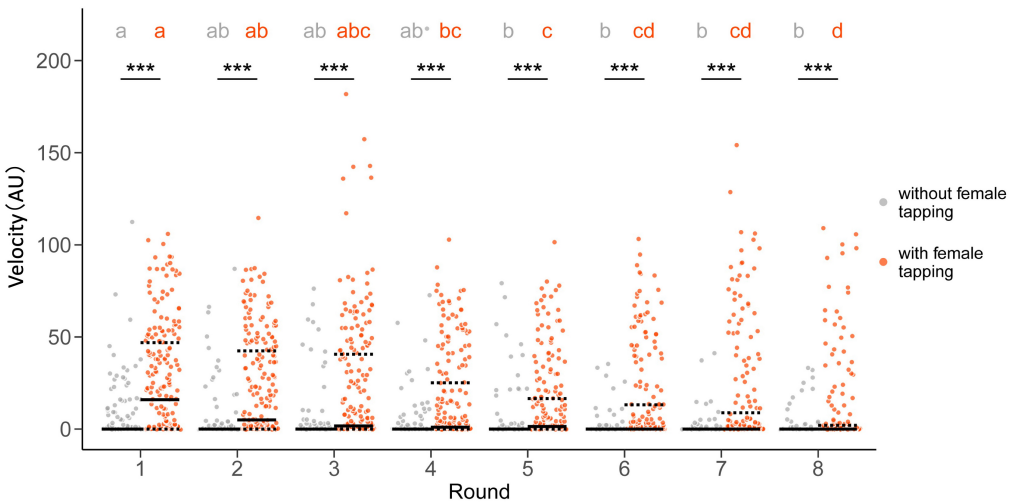
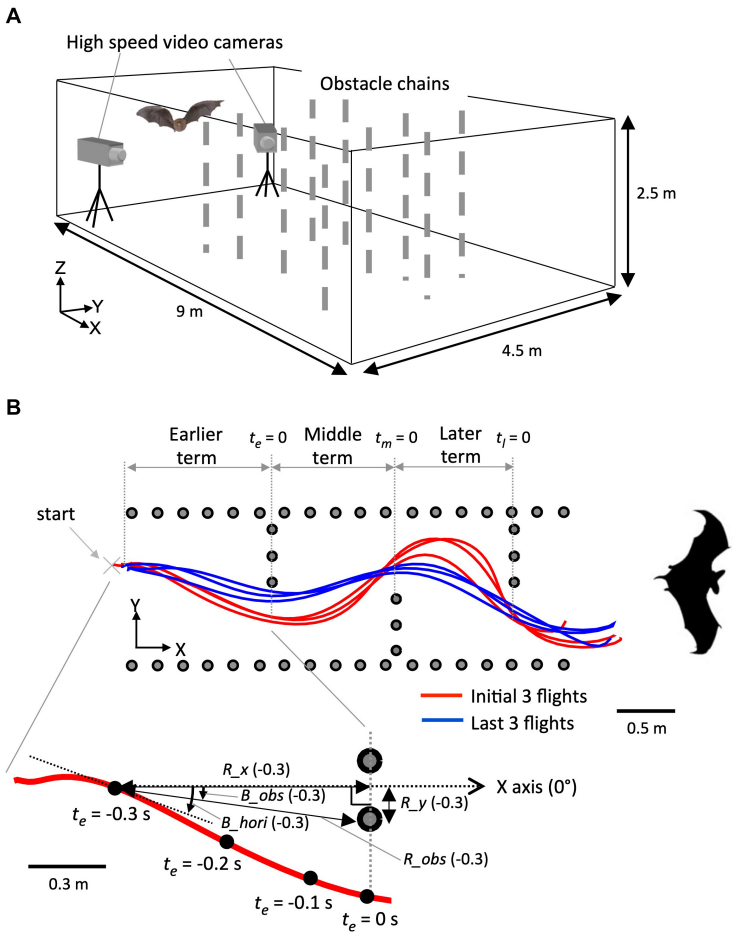


FIGURE 8 | Pheromone-driven responses of male fruit flies decreased over time. **(A)** A schematic drawing of the experimental setup. A female fly was actuated leftward and rightward in front of the male fly. The locomotion of the male fly was monitored by an optical sensor, which recorded lateral (Δx) and forward (Δy) movements at 4 Hz. **(B)** Information gain. Darker and lighter colors mean larger and smaller values, respectively. **(C)** Distribution of velocity in the chasing behavior of male flies. Control (without female tapping, gray dots) and experimental (with female tapping, orange dots) groups are shown. Solid and dotted lines represent the median and the first and third quartiles, respectively. Asterisks indicate the statistical significance between the control and test groups (Mann–Whitney *U* test followed by Bonferroni test for multiple comparison correction, $p < 0.05$). Different characters in each group indicate statistical significance among rounds (Steel–Dwass test, $p < 0.05$). The statistical details are described in **Supplementary Tables 12**, **13**.

to utilize "behavioral big data," we developed a hybrid supervised/unsupervised technique, the STEFTR method, to estimate behavioral states and to efficiently extract behavioral features solely from the trajectories of animal movement. The behavioral states of worms and penguins estimated with the STEFTR method were in reasonable agreement with the ones based on previous knowledge, highlighting the validity of our method. In addition, one of the learning-dependent behavioral features extracted from worms corresponded to a change in neural activity. Furthermore, we were able to identify temporally



C

Features	Info Gain
R_x (-0.3)	0.297
B_hori (-0.3)	0.285
R_obs (-0.3)	0.269
V (-0.3)	0.216
B_hori (-0.2)	0.213
ddB_hori (-0.3)	0.207
V (-0.2)	0.191
R_x (-0.2)	0.191
R_obs (-0.2)	0.191

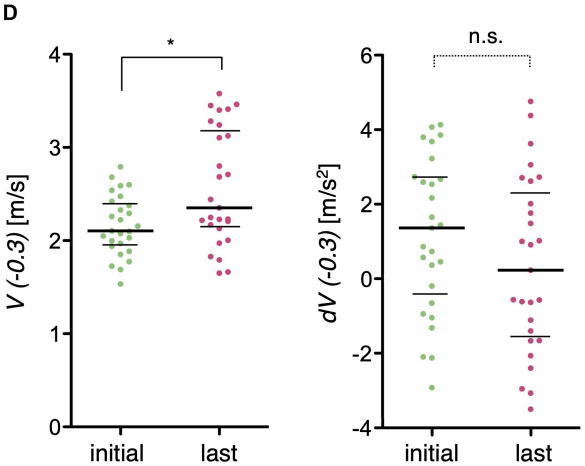


FIGURE 9 | Learning-dependent changes in bat acoustic navigation. **(A)** The experimental setup for monitoring the 3D flight trajectory of a bat during obstacle avoidance flight in a chamber. **(B)** Representative flight trajectories of a bat in the horizontal plane during repeated flights in the obstacle course. The figure on the top combines the first three (red) and last three (blue) flight trajectories. Each behavioral feature was collected in three segments: earlier, middle, and later terms. The figure on the bottom shows an expanded view of the earlier term in the first flight. Definition of the horizontal bearing of the flight (*B_hori*), distance (*R_obs*), and bearing (*B_obs*) of the bat to the nearest edge point of the obstacle chain array, longitudinal directional distance to the frontal chain array (*R_x*), and lateral directional distance to the inside pitch of the chains array (*R_y*) are indicated. Time windows for the analysis of each behavioral feature were 0.1, 0.2, or 0.3 s before or while (t = 0) passing through the chain array. **(C)** A list of extracted features of bat acoustic navigation modulated by flight experience. **(D)** Distributions of *V(–0.3)* and *dV(–0.3)* are plotted. The bars represent the median and the first and third quartiles (*$p < 0.05$, Kruskal–Wallis test with *post hoc* Dunn's test). The statistical details are described in **Supplementary Table 1**.

dynamic changes through feature extraction from fly courtship behavioral data.

One of the advantages of the STEFTR method is its versatility. Multiple methods have been reported for the behavioral analysis of specific animals under specific conditions (Baek et al., 2002; Stephens et al., 2008; Branson et al., 2009; Dankert et al., 2009; Brown et al., 2013; Kabra et al., 2013; Vogelstein et al., 2014; Wiltschko et al., 2015; Robie et al., 2017; Mathis et al., 2018). However, the animals and experimental conditions for which each of these methods can be applied are rather limited. For example, it is still not easy to effectively and robustly extract an animal's posture from a video image, especially in the wild. Even in laboratories, the parameters generally need to be adjusted again when imaging conditions changed (Dell et al., 2014; Egnor and Branson, 2016). In contrast to these methods, our method allows for behavioral analysis based on positional information that can be extracted from animal video images as well as from different methods such as a GPS device.

As a first step of the STEFTR method, we estimated behavioral states from animal trajectories. State estimation is one of the critical processes of movement analysis of animals in the wild as well as of cars and people in ecology and data science, respectively (Patterson et al., 2008; Jonsen et al., 2013; Zheng, 2015; Egnor and Branson, 2016). However, the analytical methods that can be applied to the analysis of various types of animals (and cars and people) are still in debate (Zheng, 2015; Gurarie et al., 2016). In the STEFTR method, we aimed to analyze behavior without previous knowledge of the animal and/or experimental condition and independent of the spatiotemporal scale of the behavior. For this purpose, we analyzed migratory velocity and direction, the most fundamental elements of moving objects, with appropriate moving-averaged data. Because different types of behavioral features are observed in different temporal scales from milliseconds to days or months (Buhusi and Meck, 2005), we assumed that different levels of behavioral states will be extracted with different temporal resolution. We further hypothesized that a proper combination of recording time, time unit, and temporal resolution may determine the type of behavioral state to be extracted. Our results suggest that the ratio among total recording time, time unit, and temporal window used in this study were proper to extract well-known behavioral states in worms and penguins (**Figures 2D–F**, **3F–H**). Furthermore, when we used ∼0.15% temporal window for analysis of worms, the clusters obtained from the analysis did not match to run and pirouette (**Supplementary Figure 1E**). This result also supports the idea that a proper temporal window is required to extract certain behavioral states. Still, those clusters with 0.15% temporal window may reflect other behavioral states of worms, which have not been studied yet.

State estimation based only on trajectory analysis using the STEFTR method is not perfect, as shown in the case of penguins (**Figures 3F–H**). However, the estimated behavioral states likely provide us with important information for further experiments, such as when and where in the spatiotemporal behavioral profile of the animal behavior should be analyzed in detail, especially in the case where the behavior has not been studied intensively in a quantitative manner. For example, relatively small movements at places distant from their nest in the wild may correspond to the feeding area. For neurobiological/physiological analysis, the transition from one state to the other could be triggered by a specific change in the sensory stimulus and associated with specific neural activities. It should also be noted that the STEFTR method allows semi-automatic STEFTR, which is suitable for large-scale behavioral analysis of mutant strains of laboratory model animals (**Table 3**).

Estimation of behavioral states (or behavioral modes) based on animal trajectories have been performed previously by various methods. The Expectation-Maximization binary Clustering (EMbC) method is the most similar to our method (Garriga et al., 2016). In the EMbC method, a few behavioral features, such as velocity and turn angle, were classified into two groups based on higher (H) or lower (L) values compared to a threshold, and the trajectories were segmented using combinations of the classifications (e.g., HH, HL, LH, or LL for velocity-turn angle). In that study, bat trajectories were analyzed by the EMbC method and compared with the labels by experts. True positive rates (TPR) for roost, forage, and commute were 9%, 94%, and 91%, respectively, which were at least comparable with the TPR of our method for worms and penguins ("sensitivity" in **Figures 2F**, **3H**). When state estimation from trajectory data with classic methods (first passage time, speed-tortuosity threshold, Hidden Markov Model, etc.) or with machine learning methods (linear discriminant analysis, classification and regression trees, support vector machine, etc.) were systematically compared with ground truth data, TPR of the classic methods were about 30–80%, and 80–90% for machine learning (Nathan et al., 2012; Bennison et al., 2017). Furthermore, even in a study that classified the behavioral state of *Drosophila* or mice based on high content video data with machine learning analysis, about 4–5% errors were observed (Kabra et al., 2013). This comparison of our method with earlier methods is not accurate as these studies used data different from ours. Still, we consider that our STEFTR method is one of the efficient methods to estimate behavioral states from animal trajectories.

For comprehensive feature extraction, we used information gain, an index used in decision tree analysis. Decision tree analysis is one of the machine learning techniques used for classification. Classification analysis involves classifying new, unlabeled data into appropriate classes using characteristic features and the parameters that have been extracted from the known class-labeled data. In the present study, however, the classification itself was not meaningful because the data were already classified (with/without learning or with/without sex pheromone). Instead, we focused on the procedure in the classification that identifies features useful for distinguishing between the two classes. In other words, behavioral features that are different between two classes (i.e., conditions) should be able to effectively classify the behavioral data of animals in two conditions. To our knowledge, our STEFTR method is the first pipeline to comprehensively extract behavioral features of a behavioral state in two different conditions.

Although the STEFTR method does not directly provide information about brain/neural activity underlying animal behavior, it provides us with clues required to formulate

hypotheses related to the experimental investigation of the neural activity, as shown in the case of learning-dependent changes in behavioral features and neural activity (**Figure 7**). For example, animals in the wild experience continuously changing visual, auditory, and olfactory stimuli, each of which contains multi-dimensional information (color, shape, tone, different chemical compounds, etc.). Therefore, it is difficult to identify which aspect(s) of the particular stimulus actually triggers a change in animal behavior. Estimation of behavioral states using the STEFTR method will allow us to identify the behavior-triggering stimulus by focusing on the timing and/or place of the behavioral transition. Similarly, large-scale recording of neural activities from moving animals in the laboratory itself is difficult to interpret. However, STEFTR of the behavior will greatly facilitate the identification of neural activities that are associated with behavioral transitions and/or specific behavioral features.

ETHICS STATEMENT

Penguins

This study was carried out in accordance with the recommendations of Bioscience Program Committee of Japanese Antarctic Research Expedition and the Law relating to Protection of the Environment in Antarctica. The protocol was approved by Bioscience Program Committee of Japanese Antarctic Research Expedition and the Ministry of the Environment, Government of Japan.

Flying Seabirds

This study was carried out in accordance with the recommendations of the guidelines of the Animal Experimental Committee of Nagoya University. The protocol was approved by the Animal Experimental Committee of Nagoya University. Fieldwork was conducted with permission from the Ministry of the Environment and Agency for Cultural Affairs, Japan.

Rats

This study was carried out in accordance with the recommendations of the Regulations on Animal Experiments in Osaka City University. The protocol was approved by the Animal Care and Ethics Committee of Osaka City University.

Bats

This study was carried out in accordance with the recommendations of Principles of Animal Care [publication no. 86-23 (revised 1985) of the National Institutes of Health] and all Japanese laws. The protocol was approved by the Animal Experiment Committee of Doshisha University.

AUTHOR CONTRIBUTIONS

SJY, TM, and KDK designed the experiments. SJY, KO, KI, NK, TK, DT, YY, YoI, FH, KF, YT, AY-U, KH, KS, KY, AT, YuI, AK, SH, and KDK performed the experiments. SJY and TM analyzed the data. SJY and KDK wrote the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by the Interdisciplinary graduate school program for systematic understanding of health and disease (for SJY), by the Tohoku Ecosystem-Associated Marine Science (for KS), by 58th Japanese Antarctic Research Expedition (for KI, NK, and AT) by KAKENHI JP16K16185 (for NK), JP25249020 and JP16H06536 (for KH), JP24681006, JP16H01769, JP16H06541 (for KY), 17H05983 (for AT), JP16H04655, JP18H05069, and JP17K19450 (for AK), JP16H06542 (for SH), JP16H06539 (for TM), JP16H06545 (for KDK) from the MEXT, by PRESTO 11008 (for SH) by JST, and by the Osaka University Co-Creation Program (for TM and KDK).

ACKNOWLEDGMENTS

We thank Drs. André Brown, Katsuyoshi Matsushita, Ken-ichi Hironaka, Takuma Degawa, Daisuke Yamamoto, Soh Kohatsu, and KDK laboratory members for suggestions and comments on this work. Nematode strains were provided by the Caenorhabditis Genetics Center (funded by the NIH Office of Research Infrastructure Programs P40 OD010440) and by the National BioResource Project funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.00626/full#supplementary-material

REFERENCES

Alivisatos, A. P., Chun, M., Church, G. M., Greenspan, R. J., Roukes, M. L., and Yuste, R. (2012). The Brain Activity Map Project and the Challenge of Functional Connectomics. *Neuron* 74, 970–974. doi: 10.1016/j.neuron.2012.06.006

Anderson, D. J., and Perona, P. (2014). Toward a Science of Computational Ethology. *Neuron* 84, 18–31. doi: 10.1016/j.neuron.2014.09.005

Baek, J.-H., Cosman, P., Feng, Z., Silver, J., and Schafer, W. R. (2002). Using machine vision to analyze and classify *Caenorhabditis elegans* behavioral phenotypes quantitatively. *J. Neurosci. Methods* 118, 9–21. doi: 10.1016/s0165-0270(02)00117-6

Bargmann, C. I. (2006). Chemosensation in *C. elegans*. *WormBook* 1–29. doi: 10.1895/wormbook.1.123.1

Bargmann, C. I., Hartwieg, E., and Horvitz, H. R. (1993). Odorant-selective genes and neurons mediate olfaction in *C. elegans*. *Cell* 74, 515–527. doi: 10.1016/0092-8674(93)80053-h

Bennison, A., Bearhop, S., Bodey, T. W., Votier, S. C., Grecian, W. J., Wakefield, E. D., et al. (2017). Search and foraging behaviors from movement data: A comparison of methods. *Ecol. Evol.* 8, 13–24. doi: 10.1002/ece3.3593

Berman, G. J. (2018). Measuring behavior across scales. *BMC Biol.* 16:23. doi: 10.1186/s12915-018-0494-497

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Delhi: Springer Verlag.

Branson, K., Robie, A. A., Bender, J., Perona, P., and Dickinson, M. H. (2009). High-throughput ethomics in large groups of *Drosophila*. *Nat Methods* 6, 451–457. doi: 10.1038/nmeth.1328

Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics* 77, 71–94.

Brown, A. E. X., and de Bivort, B. (2018). Ethology as a physical science. *Nat. Phys.* 14, 653–657. doi: 10.1038/s41567-018-0093-90

Brown, A. E. X., Yemini, E. I., Grundy, L. J., Jucikas, T., and Schafer, W. R. (2013). A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. *Proc. Natl. Acad Sci. U.S.A.* 110, 791–796. doi: 10.1073/pnas.1211447110

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765. doi: 10.1038/nrn1764

Colbert, H. A., Smith, T. L., and Bargmann, C. I. (1997). OSM-9, a novel protein with structural similarity to channels, is required for olfaction, mechanosensation, and olfactory adaptation in *Caenorhabditis elegans*. *J. Neurosci.* 17, 8259–8269. doi: 10.1523/jneurosci.17-21-08259.1997

Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J., and Perona, P. (2009). Automated monitoring and analysis of social behavior in *Drosophila*. *Nat. Meth.* 6, 297–303. doi: 10.1038/nmeth.1310

De Bono, M., and Maricq, A. V. (2005). Neuronal substrates of complex behaviors in *C. elegans*. *Annu. Rev. Neurosci.* 28, 451–501. doi: 10.1146/annurev.neuro.27.070203.144259

Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., de Polavieja, G. G., Noldus, L. P., et al. (2014). Automated image-based tracking and its application in ecology. *Trends Ecol. Evol.* 29, 417–428. doi: 10.1016/j.tree.2014.05.004

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* 39, 1–38. doi: 10.2307/3315718

Diba, K., and Buzsáki, G. (2008). Hippocampal network dynamics constrain the time lag between pyramidal cells across modified environments. *J. Neurosci.* 28, 13448–13456. doi: 10.1523/JNEUROSCI.3824-08.2008

Egnor, S. E. R., and Branson, K. (2016). Computational analysis of behavior. *Annu. Rev. Neurosci.* 39, 217–236. doi: 10.1146/annurev-neuro-070815-13845

Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA Workbench Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Burlington, MA: Morgan Kaufmann.

Garriga, J., Palmer, J. R. B., Oltra, A., and Bartumeus, F. (2016). Expectation-Maximization binary clustering for behavioural annotation. *PLoS One* 11:e151984. doi: 10.1371/journal.pone.0151984

Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M., and Mainen, Z. F. (2014). Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nat. Neurosci.* 17, 1455–1462. doi: 10.1038/nn.3812

Gorelik, R., and Gautreau, A. (2014). Quantitative and unbiased analysis of directional persistence in cell migration. *Nat. Protoc.* 9, 1931–1943. doi: 10.1038/nprot.2014.131

Gurarie, E., Bracis, C., Delgado, M., Meckley, T. D., Kojola, I., and Wagner, C. M. (2016). What is the animal doing? Tools for exploring behavioural structure in animal movements. *J. Anim. Ecol.* 85, 69–84. doi: 10.1111/1365-2656.12379

Iino, Y., and Yoshida, K. (2009). Parallel use of two behavioral mechanisms for chemotaxis in *Caenorhabditis elegans*. *J. Neurosci.* 29, 5370–5380. doi: 10.1523/JNEUROSCI.3633-08.2009

Jonsen, I. D., Basson, M., Bestley, S., Bravington, M. V., Patterson, T. A., Pedersen, M. W., et al. (2013). State-space models for bio-loggers: a methodological road map. *Deep Sea Res II Top. Stud. Oceanogr.* 8, 34–46. doi: 10.1016/j.dsr2.2012.07.008

Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat. Methods* 10, 64–67. doi: 10.1038/nmeth.2281

Kaplan, J. M. (1996). Sensory signaling in *Caenorhabditis elegans*. *Curr. Opin. Neurobiol.* 6, 494–499. doi: 10.1016/s0959-4388(96)80055-9

Kass, J., Jacob, T. C., Kim, P., and Kaplan, J. M. (2001). The EGL-3 proprotein convertase regulates mechanosensory responses of *Caenorhabditis elegans*. *J. Neurosci.* 21, 9265–9272. doi: 10.1523/jneurosci.21-23-09265.2001

Kimura, K. D., Fujita, K., and Katsura, I. (2010). Enhancement of odor avoidance regulated by dopamine signaling in *Caenorhabditis elegans*. *J. Neurosci.* 30, 16365–16375. doi: 10.1523/JNEUROSCI.6023-09.2010

Kohatsu, S., Koganezawa, M., and Yamamoto, D. (2011). Female contact activates male-specific interneurons that trigger stereotypic courtship behavior in *Drosophila*. *Neuron* 69, 498–508. doi: 10.1016/j.neuron.2010.12.017

Kohatsu, S., and Yamamoto, D. (2015). Visually induced initiation of *Drosophila* innate courtship-like following pursuit is mediated by central excitatory state. *Nat. Commun.* 6, 6457. doi: 10.1038/ncomms7457

Kokubun, N., Lee, W. Y., Kim, J.-H., and Takahashi, A. (2015). Chinstrap penguin foraging area associated with a seamount in Bransfield Strait. *Antarctica. Polar Sci.* 9, 393–400. doi: 10.1016/j.polar.2015.10.001

Komatsu, H., Mori, I., Rhee, J. S., Akaike, N., and Ohshima, Y. (1996). Mutations in a cyclic nucleotide-gated channel lead to abnormal thermosensation and chemosensation in *C. elegans*. *Neuron* 17, 707–718. doi: 10.1016/s0896-6273(00)80202-0

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Land, M., Islas-Trejo, A., and Rubin, C. S. (1994). Origin, properties, and regulated expression of multiple mRNAs encoded by the protein kinase C1 gene of *Caenorhabditis elegans*. *J. Biol. Chem.* 269, 14820–14827.

Landhuis, E. (2017). Neuroscience: big brain, big data. *Nature* 541, 559–561. doi: 10.1038/541559a

Lee, R. Y., Sawin, E. R., Chalfie, M., Horvitz, H. R., and Avery, L. (1999). EAT-4, a homolog of a mammalian sodium-dependent inorganic phosphate cotransporter, is necessary for glutamatergic neurotransmission in *Caenorhabditis elegans*. *J. Neurosci.* 19, 159–167. doi: 10.1523/jneurosci.19-01-00159.1999

Lewis, J. A., Wu, C. H., Berg, H., and Levine, J. H. (1980). The genetics of levamisole resistance in the nematode *Caenorhabditis elegans*. *Genetics* 95, 905–928.

Lockery, S. R. (2011). The computational worm: spatial orientation and its neuronal basis in *C. elegans*. *Curr. Opin. Neurobiol.* 21, 782–790. doi: 10.1016/j.conb.2011.06.009

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., et al. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21, 1–12. doi: 10.1038/s41593-018-0209-y

Matsumoto, S., Yamamoto, T., Yamamoto, M., Zavalaga, C. B., and Yoda, K. (2017). Sex-related differences in the foraging movement of streaked shearwaters *Calonectris leucomelas* breeding on Awashima Island in the sea of Japan. *Ornithol. Sci.* 16, 23–32. doi: 10.2326/osj.16.23

Mizuseki, K., Diba, K., Pastalkova, E., Teeters, J., Sirota, A., and Buzsáki, G. (2014). Neurosharing: large-scale data sets (spike, LFP) recorded from the hippocampal-entorhinal system in behaving rats. *F1000Res* 3:98. doi: 10.12688/f1000research.3895.1

Mizuseki, K., Sirota, A., Pastalkova, E., and Buzsáki, G. (2009). Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop. *Neuron* 64, 267–280. doi: 10.1016/j.neuron.2009.08.037

Moss, C. F., and Surlykke, A. (2010). Probing the natural scene by echolocation in bats. *Front. Behav. Neurosci.* 4:33. doi: 10.3389/fnbeh.2010.00033

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M., and Getz, W. M. (2012). Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *J. Exp. Biol.* 215, 986–996. doi: 10.1242/jeb.058602

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends Ecol. Evol.* 23, 87–94. doi: 10.1016/j.tree.2007.10.009

Pierce-Shimomura, J. T., Morse, T. M., and Lockery, S. R. (1999). The fundamental role of pirouettes in *Caenorhabditis elegans* chemotaxis. *J. Neurosci.* 19, 9557–9569.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn* 1, 81–106. doi: 10.1023/A:1022643204877

Robie, A. A., Hirokawa, J., Edwards, A. W., Umayam, L. A., Lee, A., Phillips, M. L., et al. (2017). Mapping the neural substrates of behavior. *Cell* 170, 393–406.e28. doi: 10.1016/j.cell.2017.06.032

Sasakura, H., and Mori, I. (2013). Behavioral plasticity, learning, and memory in *C. elegans*. *Curr. Opin. Neurobiol.* 23, 92–99. doi: 10.1016/j.conb.2012.09.005

Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and Tsien, R. Y. (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. *red fluorescent protein*. *Nat. Biotech.* 22, 1567–1572. doi: 10.1038/nbt1037

Stephens, G. J., Johnson-Kerner, B., Bialek, W., and Ryu, W. S. (2008). Dimensionality and dynamics in the behavior of *C. elegans*. *PLoS Comput. Biol.* 4:e1000028. doi: 10.1371/journal.pcbi.1000028

Suo, S., Ishiura, S., and Van Tol, H. H. M. (2004). Dopamine receptors in *C. elegans*. *Eur. J. Pharmacol.* 500, 159–166. doi: 10.1016/j.ejphar.2004.07.021

Tanimoto, Y., Yamazoe-Umemoto, A., Fujita, K., Kawazoe, Y., Miyanishi, Y., Yamazaki, S. J., et al. (2017). Calcium dynamics regulating the timing of decision-making in *C. elegans*.. *eLife* 6:e21629. doi: 10.7554/eLife.21629

Tian, L., Hires, S. A., Mao, T., Huber, D., Chiappe, M. E., Chalasani, S. H., et al. (2009). Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nat. Methods* 6, 875–881. doi: 10.1038/nmeth.1398

Tobin, D., Madsen, D., Kahn-Kirby, A., Peckol, E., Moulder, G., Barstead, R., et al. (2002). Combinatorial expression of TRPV channel proteins defines their sensory functions and subcellular localization in C. *elegans neurons*. *Neuron* 35, 307–318. doi: 10.1016/s0896-6273(02)00757-2

Vogelstein, J. T., Park, Y., Ohyama, T., Kerr, R. A., Truman, J. W., Priebe, C. E., et al. (2014). Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning. *Science* 344, 386–392. doi: 10.1126/science.1250298

Watanabe, S., Sato, K., and Ponganis, P. J. (2012). Activity time budget during foraging trips of emperor penguins. *PLoS One* 7:e50357. doi: 10.1371/journal.pone.0050357

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. Royal Soc. B* 314, 1–340. doi: 10.1098/rstb.1986.0056

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., et al. (2015). Mapping Sub-Second Structure in Mouse Behavior. *Neuron* 88, 1121–1135. doi: 10.1016/j.neuron.2015.11.031

Yamada, Y. (2017). *A Study on Cheap Robust Sensing for Obstacle Avoidance Guidance Based on Bio-Sonar Strategy of Bats*. Ph.D. thesis, Doshisha University, Kyoto.

Yamada, Y., Hiryu, S., and Watanabe, Y. (2016). Species-specific control of acoustic gaze by echolocating bats, *Rhinolophus ferrumequinum nippon* and *Pipistrellus abramus*, during flight. *J. Comp. Physiol. A* 202, 791–801. doi: 10.1007/s00359-016-1121-1120

Yamamoto, D., and Koganezawa, M. (2013). Genes and circuits of courtship behaviour in *Drosophila* males. *Nat. Rev. Neurosci.* 14, 681–692. doi: 10.1038/nrn3567

Yamazoe-Umemoto, A., Fujita, K., Iino, Y., Iwasaki, Y., and Kimura, K. D. (2015). Modulation of different behavioral components by neuropeptide and dopamine signalings in non-associative odor learning of *Caenorhabditis elegans*. *Neurosci. Res.* 99, 22–33. doi: 10.1016/j.neures.2015.05.009

Yamazoe-Umemoto, A., Iwasaki, Y., and Kimura, K. (2018). Measuring spatiotemporal dynamics of odor gradient for small animals by gas chromatography. *Bio-Protocol* 8:e2797. doi: 10.21769/bioprotoc.2797

Yoda, K. (2018). Advances in bio-logging techniques and their application to study navigation in wild seabirds. *Adv. Robot.* 58, 1–10. doi: 10.1080/01691864.2018.1553686

Yoda, K., Naito, Y., Sato, K., Takahashi, A., Nishikawa, J., Ropert-Coudert, Y., et al. (2001). A new technique for monitoring the behaviour of free-ranging Adélie penguins. *J. Exp. Biol.* 204, 685–690.

Yoda, K., Shiomi, K., and Sato, K. (2014). Foraging spots of streaked shearwaters in relation to ocean surface currents as identified using their drift movements. *Prog. Oceanogr.* 122, 54–64. doi: 10.1016/j.pocean.2013.12.002

Zhang, S. X., Miner, L. E., Boutros, C. L., Rogulja, D., and Crickmore, M. A. (2018). Motivation, perception, and chance converge to make a binary decision. *Neuron* 99, 376–388.e6. doi: 10.1016/j.neuron.2018.06.014

Zheng, Y. (2015). Trajectory Data Mining. *ACM Tras. Intell. Syst. Techno.* 6, 1–41. doi: 10.1145/2743025

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Yamazaki, Ohara, Ito, Kokubun, Kitanishi, Takaichi, Yamada, Ikejiri, Hiramatsu, Fujita, Tanimoto, Yamazoe-Umemoto, Hashimoto, Sato, Yoda, Takahashi, Ishikawa, Kamikouchi, Hiryu, Maekawa and Kimura. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neural Circuits

ORIGINAL RESEARCH
published: 03 July 2019
doi: 10.3389/fncir.2019.00044

Stimulus-Specific Adaptation Decreases the Coupling of Spikes to LFP Phase

Mohsen Parto Dezfouli[1†], Mohammad Zarei[2†], Mehran Jahed[2‡] and Mohammad Reza Daliri[1*‡]*

[1] Neuroscience and Neuroengineering Research Laboratory, Biomedical Engineering Department, School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, [2] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

Stimulus repetition suppresses the neural activity in different sensory areas of the brain. This mechanism of so-called stimulus-specific adaptation (SSA) has been observed in both spiking activity and local field potential (LFP) responses. However, much remains to be known about the effect of SSA on the spike–LFP relation. In this study, we approached this issue by investigating the spike-phase coupling (SPC) in control and adapting paradigms. For the control paradigm, pure tones were presented in a random unbiased sequence. In the adapting paradigm, the same stimuli were presented in a random pattern but it was biased to an adapter stimulus. In fact, the adapter occupied 80% of the adapting sequence. During the tasks, LFP and multi-unit activity were recorded simultaneously from the primary auditory cortex of 15 anesthetized rats. To clarify the effect of adaptation on the relation between spike and LFP responses, the SPC of the adapter stimulus in these two paradigms was evaluated. Here, we employed phase locking value method for calculating the SPC. Our data show a strong coupling of spikes to LFP phase most prominently in beta band. This coupling was observed to decrease in the adapting condition compared to the control one. Importantly, we found that adaptation reduces spikes dominantly from the preferred phase of LFP in which spikes are more likely to be present there. As a result, the preferred phase of LFP may play a key role in coordinating neuronal spiking activity in neural adaptation mechanism. This finding is important for interpretation of the underlying neural mechanism of adaptation and also can be used in the context of the network and related connectivity models.

Keywords: stimulus-specific adaptation, spike-phase coupling (SPC), primary auditory cortex, local field potential (LFP), multi-unit activity (MUA)

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Nikolay Bibikov,
N. N. Andreyev Acoustics Institute, Russia
Mehdi Adibi,
University of New South Wales Sydney, Australia
Ehsan Arabzadeh,
Australian National University, Australia

***Correspondence:**
Mehran Jahed
jahed@sharif.edu
Mohammad Reza Daliri
daliri@iust.ac.ir

†Co-first authors

‡Co-senior authors

Received: *17 March 2019*
Accepted: *18 June 2019*
Published: *03 July 2019*

Citation:
Parto Dezfouli M, Zarei M, Jahed M and Daliri MR (2019) Stimulus-Specific Adaptation Decreases the Coupling of Spikes to LFP Phase. Front. Neural Circuits 13:44. doi: 10.3389/fncir.2019.00044

INTRODUCTION

Neural synchrony/desynchrony has been targeted of many recent brain studies and has deeply influenced modern knowledge in various functions, such as sensory coding, decision making, working memory, and selective attention (Eckhorn and Obermueller, 1993; Baker et al., 1999; Cutsuridis and Hasselmo, 2011; Muthukumaraswamy, 2011; Li et al., 2014;

Mendoza-Halliday et al., 2014; Ruff and Cohen, 2014; Fazlali et al., 2016; Ding et al., 2017; Bahmani et al., 2018; Johnson et al., 2018a,b). Neural activity is either measured by spiking activity or local field potentials (LFPs), in order to encode sensory information (Csicsvari et al., 2003; Buzsáki and Draguhn, 2004; Liu and Newsome, 2006; Katzner et al., 2009; Whittingstall and Logothetis, 2009; Buzsáki et al., 2012). LFPs reflect the activity of a population of neurons based on spatially averaged synaptic activity (Buzsáki and Draguhn, 2004; Buzsáki et al., 2012; Jansen et al., 2014). They are continuous cyclic signals with various frequency bands. The low frequencies of the LFP as in delta, theta, alpha, and beta bands are a compound signal of slower events from a large population of cells. Therefore, the activity in the low-frequency bands of the LFP indicates the combination of neural activity across larger networks of neurons (Buzsáki and Draguhn, 2004). On the other hand, the high frequencies of the LFP such as low-gamma and high-gamma bands reflect higher local activity (Csicsvari et al., 2003; Liu and Newsome, 2006; Whittingstall and Logothetis, 2009).

The relation between spike times and the phase of LFPs has pointed out to several cognitive functions in various brain regions (Pesaran et al., 2002; Ray and Maunsell, 2010; Womelsdorf et al., 2010; Li et al., 2017), including the prefrontal cortex (Siegel et al., 2009), the visual cortex (Whittingstall and Logothetis, 2009), and hippocampus (Sirota et al., 2008; Cutsuridis and Hasselmo, 2011). For instance, the coupling between spikes to phases of LFP in the theta band encodes spatial memory in the hippocampus (Cutsuridis and Hasselmo, 2011). This spike–LFP phase relation, so-called spike-phase coupling (SPC), shows how activities of single neurons are harmonized for the averaged synaptic activity or LFPs to derive various cognitive functions. Measuring the coupling of single neuron's spiking activities to the LFP is a method to investigate neuronal synchronization. Locking of spiking activities to LFP is a feature of such inter-neuronal synchrony. Several spike–LFP synchronization measures have been introduced in previous studies namely phase locking value (PLV) (Lachaux et al., 1999), spike field coherence (Grasse and Moxon, 2010), and pairwise phase consistency (Vinck et al., 2010, 2012). PLV, as one of the most important synchronization measures, represents the resultant length of the circular averaging of the instantaneous phases simultaneous to spikes. PLV vector is close to 1 for a trial when most spikes are coupled to a certain phase, and it is close to 0 when most spikes are spread across various phases (Lachaux et al., 1999). The shortcoming of PLV method is its bias on the number of spikes (Vinck et al., 2010, 2012; Zarei et al., 2018). To minimize this limitation, in general, an extra step is utilized in order to equalize the spikes of objective conditions for a certain count. Spike field coherence quantifies the synchronous activity between LFPs and spikes as a function of frequency. This method adds the power of LFP in different frequencies surrounding the corresponding spikes to the average LFP and normalizes this sum to the total number of spikes (Grasse and Moxon, 2010). Spike field coherence represents the coupling of spikes in regions around the LFPs. Hence, this method cannot exclusively represent the coupling of spikes to phase fluctuations. Lastly, the pairwise phase consistency method calculates the cosine of the absolute angular distance of the LFP phase across all possible pairs of spikes (Vinck et al., 2010). This method has a high variance for low spike counts and may yield negative values which are not physiologically justified. Therefore, this study employs SPC based on PLV method in order to uncover the effect of adaptation on the spike–LFP relation. Estimation of SPC through the PLV method reflects the phases of LFP for which spikes have occurred.

Neural adaptation is a common phenomenon which has been extensively observed in the mammalian sensory areas such as visual (Müller, 1999; Kayser et al., 2009), auditory (Ulanovsky et al., 2003; Dean et al., 2005; Anderson et al., 2009; Malmierca et al., 2009; Parto Dezfouli and Daliri, 2015), and somatosensory (Katz et al., 2006; Adibi et al., 2013) systems. Generally, adaptation tends to suppress neuronal activities in various sensory systems. Notably, suppression is only the case for a limited range of stimuli with repeated or prolonged stimulation (Dean et al., 2005; Adibi et al., 2013). Adapting to the environment as a result of the frequent representation of one stimulus, such as light, smell, or sound, is a vital brain function that the lack of it could be disturbing. Adaptation causes certain variations in neural properties in order to reduce attention to frequent stimuli. Particularly, it leads to an increase in the neural sensitivity against unexpected changes for deviance detection (Ulanovsky et al., 2003). Research on adaptation and change detection points to evoked potential signals in mismatch negativity studies, which includes human (Näätänen et al., 2007), primate (Javit et al., 1994), and cat (Csépe et al., 1987) experiments. The mismatch negativity is a component of event-related potential that occurs in response to a rare stimulus in a sequence. In the auditory system, two frequencies with almost similar responses are used in an oddball paradigm, represented by different probabilities. In some sequences, one tone is presented repeatedly as a standard stimulus and the other one is presented rarely and this role is swapped in other sequences. The difference between the responses in standard and rare conditions results in mismatch negativity which arises independent to the subject's attention (Näätänen et al., 2007; Zhou et al., 2013). In recent years, the oddball paradigm has been extensively used in electrophysiological studies. As such, researchers have pinpointed interesting phenomena in the auditory cortical neurons, known as stimulus-specific adaptation (SSA) (Ulanovsky et al., 2003, 2004; Nelken and Ulanovsky, 2007; Parto Dezfouli and Daliri, 2015). Here, we used the term "adaptation" for "SSA" concept.

It is known that brain systematically suppresses neural responses to a repeated stimulus. This mechanism exists in various sensory areas of the brain which seems to suppress both field potentials and spiking activities. But how does this mechanism affect the relation between spike and LFP? We addressed this issue by investigating the SPC in the primary auditory cortex of the rat during an experiment consisting of two control and adapting conditions. In what follows, we will introduce our analytical and experimental methods in detail followed by the results of experiments and analyses. Finally, in the section "Discussion," a thorough review of this work with specific references to key points will be presented.

MATERIALS AND METHODS

The procedure of surgery, experimental paradigm, data acquisition, and preprocessing of data are described in Parto Dezfouli and Daliri (2015). Here, we provide further details necessary to evaluate and present the current data analysis.

Electrophysiological Recording

The data were recorded from the left auditory cortex (A1 area) of 15 adult male and female Wistar rats weighing 250–350 g. A linear multi-electrode array, consisting of four tungsten electrodes (FHC, 5M, United States; ∼5–10 μm tip diameter) were used for extracellular recording. The electrodes were directed into the cortex using a Microdrive (SM-21, Narishige, Japan). LFP and multi-unit activity (MUA) were recorded simultaneously using the USB-ME64-PGA recording system (Multichannel System, Germany). Furthermore, an online data visualization was utilized through a multichannel software known as "MCRack." The recorded raw signals of each channel (with 10 kHz sampling rate) were initially pre-amplified by an eight-channel Miniature Preamplifier. Next, the amplified signals were band-pass filtered from 1 to 5 kHz and amplified with a gain of 1000 with a Programmable Gain Amplifier device. Finally, the recorded data were transmitted to the computer for subsequent off-line analyses.

Experimental Paradigm

Before the main task, in order to detect the selective neurons, broad-band noise bursts were presented with 300 ms duration and 500 ms inter-stimulus interval with an amplitude of 50 dB. For all four electrodes, only the channels for which average LFP amplitudes were large enough were recorded. For every recording, various arrangements of 11 frequencies (200 Hz to 20 kHz), with the frequency difference of $f = \frac{(f2-f1)}{f1} = 44\%$ and seven intensities of 10–70 dB SPL at 10 dB steps, were considered in order to measure the frequency response area of every recording site. Each frequency–intensity combination was presented 10 times in a quasi-random sequence. Finally, based on the frequency response area of each single site, five selected frequencies around the characteristic frequency were considered for the main task. As a result, for each recording site, these five selected frequencies, namely f1–f5, in four higher intensities of 40–70 dB SPL in 10 dB steps, provided 20 frequency–intensity combinations. **Figure 1A** shows the timeline of the task in which these pure tone combinations were presented with 50 ms duration and 300 ms inter-stimulus interval. Because of several combinations of intensities and frequencies, the sequences had to be as short as possible in order to obtain comparable data from two paradigms in the same recording site. As such this study utilized a repetition of 2–3 Hz as an effective investigation of SSA paradigm (Taaseh et al., 2011; Shen et al., 2015; Takaura and Fujii, 2016). **Figure 1B** illustrates the main assessing task consisting of two control and adapting paradigms. For the control paradigm, 20 combinations consisting of five selected frequencies in four intensities were presented in a uniformly distributed manner. Each frequency–intensity combination was presented for 30 times. In the adapting paradigm, the same combinations were used somehow the middle frequency (f3) corresponding to the level of 60 dB SPL was considered as the adapter and the other combinations were assumed as its neighbors. Unlike the unbiased control paradigm, the adapting paradigm biased to the adapter. The adapter tone was presented four times per each presentation and therefore occupied 80% of all tone probability in this sequence. For instance, in the control paradigm with 600 trials (20 combinations in 30 times), the stimuli were presented with the same probability of 5%. While in the adapting paradigm with 2850 stimuli, the adapter was presented 2280 times (80%) and all other 19 combinations were presented for 570 times (20%). As a result, the adapter stimulus (as the target condition of this study) was presented with the probability of 5 and 80% in control and adapting sequences, respectively. In order to ensure the stability of recording conditions, the first paradigm was presented once before and once after the adapting paradigm. The sites which responded with $>30\%$ variation were excluded.

Data Analysis

The LFP and MUA were recorded from the primary auditory cortex of anesthetized rats using a simultaneous four-channel recording. For this study, we considered only the sites with responsive activity for both MUA and LFP. Overall 98 sites were selected for the main analyses. Each 50 ms stimulus presentation with its following 300 ms delay interval was considered a trial. Accordingly, every recording site consisted of ∼30 and 2280 trials of the characteristic frequency (f3) in the control and adapting paradigms, respectively, where the number of 2280 trials is based on four iterations for every 30 trials of 19 combinations. All analyses were carried out in MATLAB 2016b (Mathworks, Natick, MA, United States). For LFP, the raw data were passed from a low-pass filter with a cutoff frequency of 300 Hz. Likewise for MUA analyses, first the raw signals were filtered using a band pass filter with cut-off frequencies of 300 and 3000 Hz. Subsequently, spike times were detected by a threshold crossing method. Ultimately, all responses were aligned to the stimulus onset and their baselines were corrected (Dezfouli et al., 2014). In order to correct for the baseline, as close as possible to the onset of the response, the average response of the first 5 ms duration from the stimulus onset was subtracted from it. The shortest response latencies were consistently observed to be longer than 8 ms. The standard error of the mean (SEM) was considered as a criterion for measuring the variability of the response.

The time window of 0–100 ms with respect to stimulus onset was selected as the duration of the analysis. To remove the 50 Hz noise and for the purpose of filtering the LFPs in different frequency bands, a band-pass and non-causal finite impulse response (FIR) filter was used. The traditional bands were considered for the analysis of LFP signals, namely delta (δ; 1–4 Hz), theta (θ; 5–8 Hz), alpha (α; 9–12 Hz), beta (β; 13–30 Hz), low-gamma (γ_L; 31–70 Hz), and high-gamma (γ_H; 70–120 Hz). Since we found the coupling of the spike to LFP phase in beta range, we focused our analyses within the beta band (13–30 Hz), but the frequencies up to 120 Hz were also analyzed. Likewise, we filtered the beta range of LFPs into separate 4 Hz windows. For each frequency-based window, SPC was computed for the

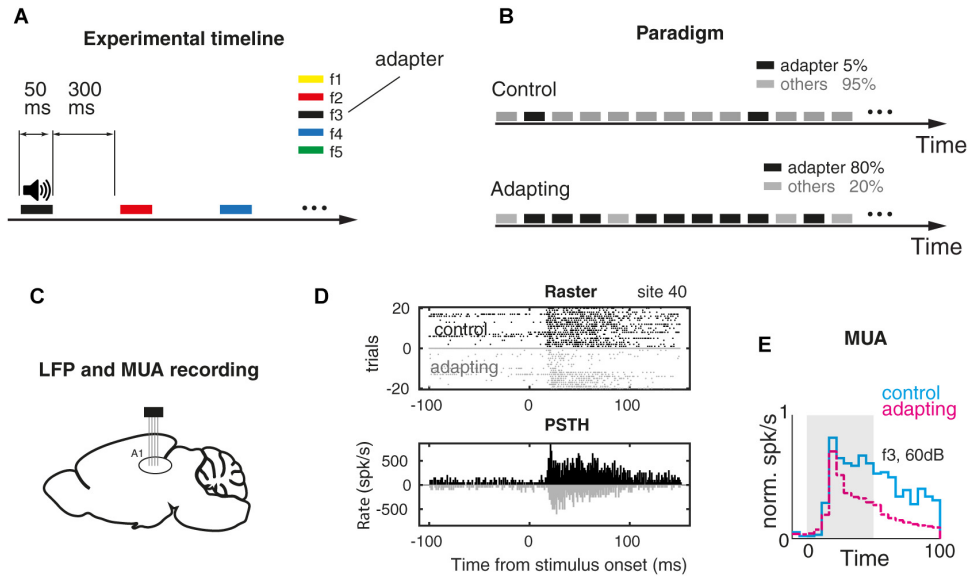


FIGURE 1 | Adaptation experiment and MUA response. **(A)** Timeline of the auditory task. Pure tone stimuli were presented randomly for 50 ms and with 300 ms inter-stimulus interval. Stimuli were pure tones with a particular frequency and level out of five selected frequencies (f1–f5) and four intensities (40–70 dB SPL). **(B)** Two sequences of stimuli that were utilized for investigating the adaptation effect. In the first sequence (control) pure tones were presented with an equal probability. Each frequency–intensity combination was presented for 30 times. In the adapting sequence, the same combinations were presented with this a difference that the middle frequency at the level of 60 dB SPL (as the adapter) was presented with the probability of 80% of the whole sequence. **(C)** LFP and MUA were recorded from the primary auditory cortex of the anesthetized rat. **(D)** Raster plot and peristimulus time histogram (PSTH) of frequency f3 in 60 dB SPL intensity in a sample recording site. **(E)** Comparing PSTH of control vs. adapting conditions in the population of recording sites (n = 98).

control and the adaptation conditions. Notably, the sites with a responsive spiking activity to stimuli and LFPs with a negative wave in deep layers were selected for SPC analysis. Moreover, in order to avoid edge effects, created by the cut-off at the starting and end of an LFP time section, we added 900 ms data from the same trial to both ends of the LFP (±450) and later eliminated the corresponding parts from the filtered signal.

Spike-Phase Coupling (SPC) Based on the PLV Method

As noted before, SPC shows how activities of neurons are coordinated for the averaged synaptic activity or LFPs. Here, we employed PLV to measure SPC. PLV method measures the strength of consistency or locking of phases in spike times, by calculating the angular summation between phases of LFP in spike times. The amplitude of PLV shows the SPC strength and its angle reflects the phases of LFPs for which spikes have occurred. To evaluate PLV in selected frequency bands, the Hilbert transform was utilized for the analytic signal of LFP. The Hilbert transform is defined as,

$$\mathrm{HT}(x(t)) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau}\, \mathrm{d}\tau$$

It converts a real-valued signal ($x(t)$) into a complex analytical signal ($\mathrm{HT}(x(t))$) where P is the Cauchy principal value of the singular integral. Next, instantaneous phases $\phi(t)$ were obtained by calculating the angles corresponding to the aforementioned analytic signal,

$$\phi(t) = \arctan\left(\frac{\mathrm{HT}(x(t))}{x(t)}\right)$$

Furthermore, PLV method is described by the following formula:

$$\mathrm{PLV} = \frac{1}{N}\left|\sum_{\mathrm{n}=1}^{\mathrm{N}} \exp\left(j\phi\right)\right|$$

where N is the number of spikes and φ is the LFP phase at the times of spike occurrence.

It is worth to note that one limitation of SPC estimation is that the value of SPC is dependent on spike numbers. For instance, if one compares two conditions with different spike numbers, it is more likely that the condition with a larger spike number offers less SPC. This issue was addressed in two recent studies on SPC (Vinck et al., 2012; Zarei et al., 2018). To overcome this issue we equalized the number of spikes in the control and adapting conditions. Hence, after identifying a threshold for the average spike number, trials with spike numbers < threshold were removed and spikes in trials with spike numbers > threshold were reduced (removed randomly) to the threshold value. Notably, the LFP amplitudes for each recording site were normalized by subtracting the mean and dividing the result by the standard deviation in order to create normalized LFP signals. Normalization was performed to set aside the possibility of LFP power effect on the SPC strength.

Spike-Triggered Average (STA) LFP

The spike-triggered average (STA) LFP is a quantity which links a spike train with the LFP, recorded simultaneously. It indicates the average LFP value chosen at the times of the occurrence of spikes. STA is calculated by averaging the LFP amplitude in a trace surrounding spiking times. In this study, the LFP trace from −20 to 20 ms relative to the spike times was averaged to estimate the STA where this sum was subsequently divided by the total number of spikes. Notably, the STA would produce a flat result for independent neuronal activities. Otherwise, the result could indicate the coupling of spikes to a particular phase of LFP.

RESULTS

To elucidate the effect of adaptation on SPC, we performed an auditory experiment consisting of two control and adaptation conditions (Parto Dezfouli and Daliri, 2015). The timeline of the experiment and the two control and adapting paradigms are shown in **Figures 1A,B**. Twenty combinations of pure tones were used in the experiment. In the control paradigm, they were presented randomly with a similar probability that each combination of pure tones occupies 5% of the sequence. In the adapting paradigm, the same stimuli were presented randomly with this difference that in the adapting paradigm the middle frequency (f3) at the level of 60 dB SPL (as the adapter) occupied 80% of the total sequence. The rest of the combinations occupied 20% of this sequence. We collected LFP and MUA from 98 recording sites over the primary auditory cortex of 15 anesthetized rats while they were presented by auditory pure tones in two conditions (**Figure 1C** – see the section "Materials and Methods"). The original signal of an example recording site (site 14; control paradigm) before its conversion to LFP and spiking activity is shown in **Supplementary Figure S1**. The raster plot and peristimulus time histogram (for 20 trials) of the characteristic frequency related to a single recording site (site 40) were shown for the control and adapting conditions, separately (**Figure 1D**). Comparison of the population spiking activity between control and adapting conditions, in line with previous studies (Ulanovsky et al., 2003; Nelken and Ulanovsky, 2007), shows a suppression in the spiking activity due to adaptation (**Figure 1E**).

Spike-Phase Coupling Suppressed by Adaptation

Comparison of the SPC of a particular stimulus between control and adapting conditions indicated that SPC was reduced due to adaptation (SPC of adaptation < control). We employed PLV to calculate SPC and since it is believed that the PLV method is biased to the spike count, first one needs to equalize the spike counts of trials within each recording site. To this end, we determined the average number of spikes for each recording site (**Figure 2A**). Using this average, it is possible to estimate the value of spike rates for each recording site to drop extra spikes. Indeed, this is the easiest way to calculate the spike count threshold (T) and is frequently preferred by researchers while it is not optimal.

Here, we used an optimal thresholding method which results in an optimal compromise between the number of spikes and trials. We assumed different spike count thresholds for finding the optimal threshold. Due to PLV's bias on the spike count, we first balanced the spike count across different sites. To obtain the optimal threshold for the spikes number, we computed the total spike count from the neural data that would be remained after utilizing different thresholds (**Figure 2B**). Accordingly, the number of T = 14 spk/bin was selected as the optimal threshold value. Here, the bin is the 100 ms window from stimulus onset. Consequently, the trials with the number of spikes below the threshold were removed. Also, the trials with spikes number higher than the threshold were equalized relative to spike number to the threshold value by removing the additional spikes randomly. Thus, all resulting trials had the same spike rates and we could use an existing method such as PLV in order to calculate the SPC.

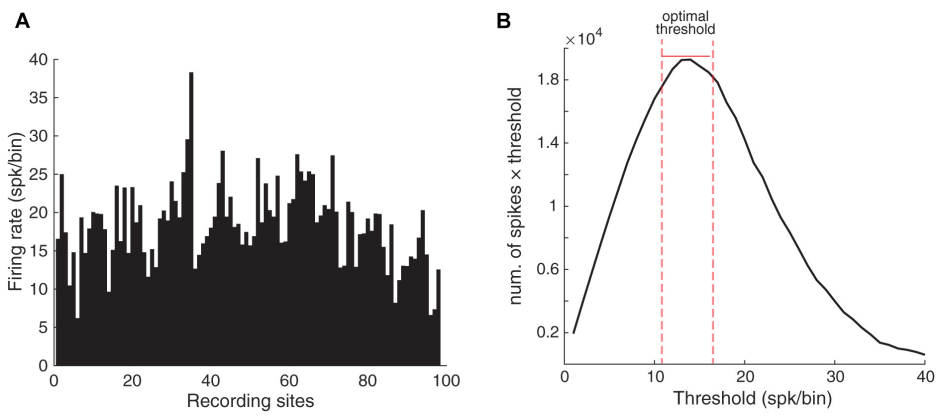


FIGURE 2 | Selecting the optimal threshold. **(A)** The average number of spikes in each electrode during the analysis period (100 ms). The *X*-axis shows the site indices and the *Y*-axis shows the spike rate. **(B)** The optimal spike count threshold for computing SPC. The *X*-axis indicates different values for the spike count threshold and *Y*-axis indicates the total spike count considered for each threshold (T). The optimal threshold (T = 14 spk/bin) produced the maximum spike count from the neural data. The bin is the 100 ms window from stimulus onset.

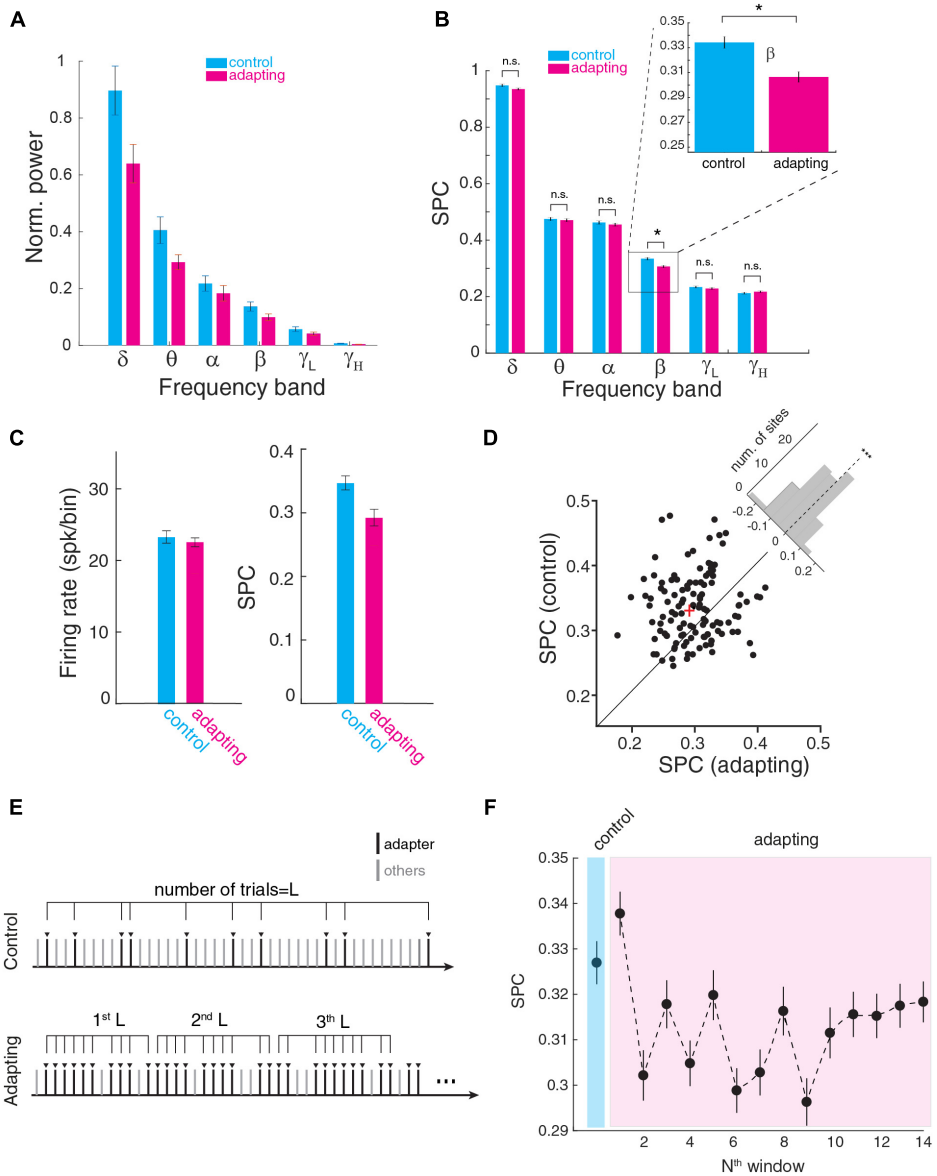


FIGURE 3 | Spike-phase coupling at different frequency bands. **(A)** Power induced in six frequency bands for control and adapting conditions. The average band power was normalized to the maximum power. Significant power reduction observed in all frequency bands except alpha band. Maximum power difference between control and adapting condition arose in delta band and the minimum was in alpha band. **(B)** SPC strength for control and adapting conditions averaged across neurons for different frequency ranges of delta (δ; 1–4 Hz), theta (θ; 5–8 Hz), alpha (α; 9–12 Hz), beta (β; 13–30 Hz), low-gamma (γ_L; 31–70 Hz), and high-gamma (γ_H; 70–120 Hz). SPC strength is shown to be significantly different for the two conditions within the beta band ($p < 0.0001$; Wilcoxon rank sum test). **(C)** Comparison of SPC between the f3 condition of adapting sequence and a condition with similar spike rate in the control condition. The SPC in the adapting condition reduces (right panel) while the spiking activity was similar (left panel). **(D)** Scatter plot of the SPC strength (control vs. adapting) within the beta band (13–30 Hz) for all recording sites. The "cross" sign marks the average SPC (control vs. adapting) of the population of the recording site. The histogram in the upper right shows the distribution of recording sites toward two objective conditions, namely control and adapting ($p < 0.001$; t-test, $n = 98$, mean = –0.029, std = 0.071). It illustrates the distribution of SPC difference ($SPC_{adapting} - SPC_{control}$) across all recording sites. Each data point corresponds to a recording site. **(E)** Schematic of division of the adapting trials into partitions with the same length (L) based on the number of test trials in the control sequence (L). **(F)** Time course of adaptation effect on SPC. For all adapter trials in the adapting sequence, the data were partitioned based on its trials' number in the control sequence (L trials in each bin). SPC strength in the adapting sequence is shown to be lower than the control sequence as the desired window increases (exponentially descends).

We next measured the SPC strength within six frequency bands; delta to gamma. First, we calculated the power-induced LFP signal in these frequency bands and compared the band power between control and adapting conditions (**Figure 3A**). Results show a significant reduction of LFP power almost in all frequency bands except alpha band ($p < 0.05$, t-test). On the other hand, as shown in **Figure 1E** and also previous reports (Taaseh et al., 2011; Parto Dezfouli and Daliri, 2015) adaptation

reduces the firing rate. Higher rate of activity leads to less SPC strength based on PLV (**Supplementary Figure S2**). Accordingly, to minimize any effect of LFP power or firing rate difference on SPC values, along with equalizing spike rate, we normalized LFPs and considered just their phase feature. Next, we estimated the SPC strength in each frequency band (**Figure 3B**). **Figure 3B** shows the SPC strength for the control and adapting conditions which were averaged across recording sites. Here, SPC strength was significantly different for the control and adapting conditions within the beta (13–30 Hz) band ($p << 0.01$; Wilcoxon rank sum test; $p < 0.05$; Bonferroni correction). Other bands showed no significant SPC difference between the control and adapting conditions (δ; $p = 0.14$), theta (θ; $p = 0.61$), alpha (α; $p = 0.52$), low-gamma (γ_L; $p = 0.57$), and high-gamma (γ_H; $p = 0.35$).

To investigate whether the phase decoupling is a consequence of reduced responsiveness of neurons (both in terms of LFP and spiking activity), or is governed by additional mechanisms, we compared the SPC of adapter stimulus under adaption condition with the SPC for a stimulus under a control condition with the matched firing rates. By comparing the lower intensities (intensity at the level of 50 dB) in the control sequence which have similar firing rate as that of the adapting condition, the coupling was less for adapting condition while their firing rates were matched (**Figure 3C**). This result shows that the adaptation-induced decoupling is mediated by a separate mechanism than suppression of excitability and responsiveness of individual cells. Subsequently, the SPC strength of control and adapting condition was compared across all recording sites. **Figure 3D** shows the SPC of all recording sites ($n = 98$) within the beta band (13–30 Hz). Each dot represents the SPC strength of the control vs. adapting condition for a recording site. The histogram in the upper right shows the distribution of two conditions. The results show that across the sites, SPC strength for the control condition is significantly higher than adapting condition ($p < 0.001$; t-test, $n = 98$, mean = −0.029, std = 0.071).

We further performed a systematic examination of the time-course of the adaptation effect on the SPC strength. For this purpose, as depicted in **Figure 3E**, we partitioned the probing trials in the adapting paradigm based on its number in the control sequence ($L = 30$), namely [1: L], [L: 2L], etc. Next, SPC was computed for L control probe trials and compared with 14 L portions of adapting trials (**Figure 3F**). The result shows that a reduction of SPC strength emerges from the second part of L trials in the adapting paradigm. This value converges to a stable value from ∼10th partition after some fluctuations.

Spikes Are Coupled to the Phase of LFPs Within the Beta Band

Previous studies have revealed that the spike rate is attenuated in the adapted condition compared to the control condition (Taaseh et al., 2011; Parto Dezfouli and Daliri, 2015). This study suggests that adaption causes SPC to reduce. This reduction occurs in the coupling of spikes to LFP phase within the beta band.

To examine the relation of this reduction of spikes to LFP phase, we first tested if the preferred phase for two conditions is matched. We defined the preferred phase as the phase in which spikes prominently occur (α) and consequently give the most SPC value. Accordingly, anti-preferred phase was defined as the phase with 180° distance from the preferred phase ($180 - \alpha$). For this purpose, the histogram of LFP phases in spike times was computed for both conditions. **Figure 4A** shows the histogram of LFP phases in the beta band across sites for both control and adapting conditions. As depicted, the distribution of phases across recording sites differs significantly from the uniform distribution in both conditions. The mean locking phase is 2.89 and 2.76 rad (165°, 158°) for control and adapting conditions, respectively, and there is no statistically significant difference between them ($p = 0.3$, t-test). This suggests that spiking activity in the auditory cortex tends to occur more frequently at a similar phase within the beta band (13–30 Hz) of LFPs, independent of adapting or control conditions. The results of the SPC for

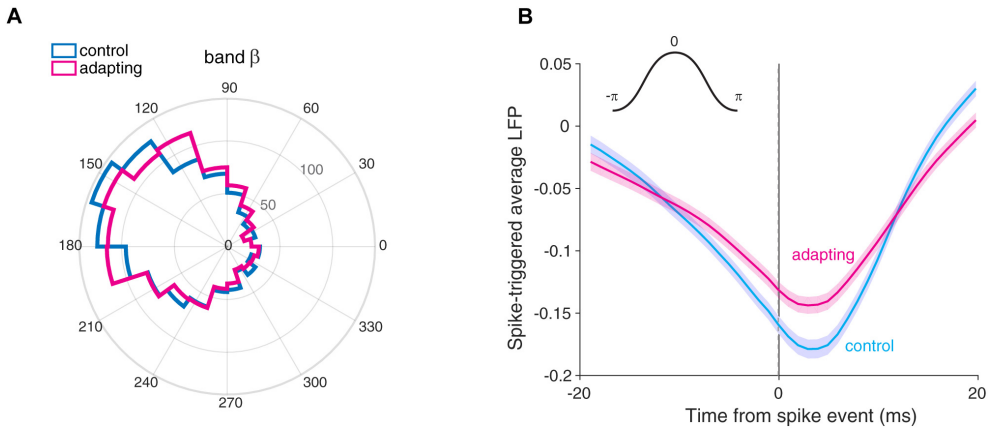


FIGURE 4 | The preferred phase of LFP. **(A)** The histogram of phases for the preferred LFP phase in control (cyan) and adapting (magenta) conditions within the beta band (13–30 Hz). The mean locking phases are 0.91 and 0.88 rad (165° and 158°) for control and adapting conditions, respectively. **(B)** Spike-triggered average (STA) of normalized LFP across sites for control (cyan) and adapting (magenta) conditions. STA curves show the spikes are coupled to the ∼160° of the phase of LFP for two conditions. The inset in the upper section illustrates the phase values in a cycle. Shade areas depict standard error of mean across all recording sites.

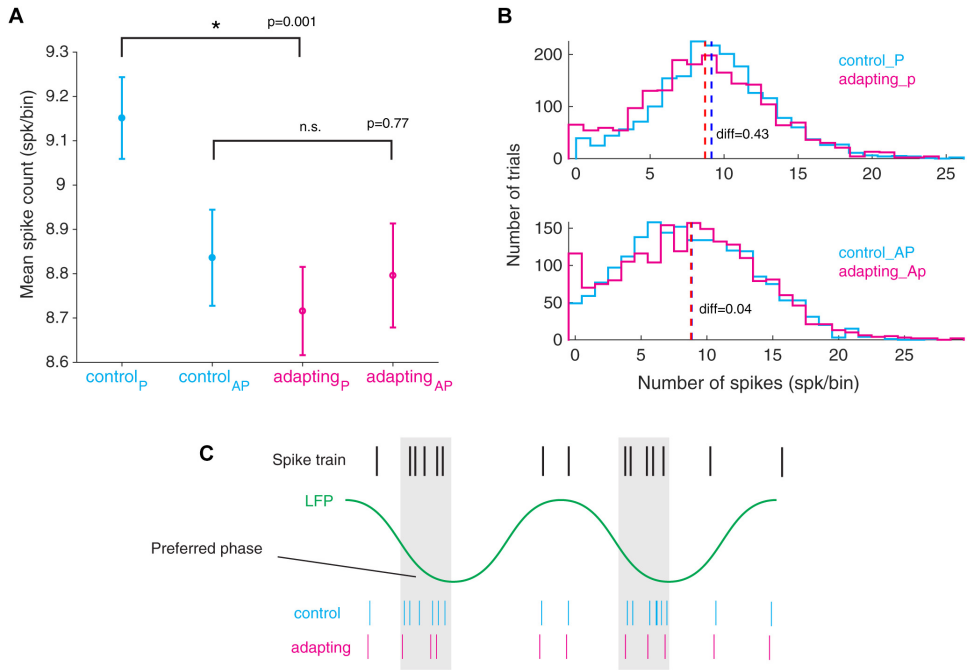


FIGURE 5 | Declining in the number of spikes from the preferred phase of LFP during adaptation. **(A)** Mean standard error of spike counts at both the preferred and the anti-preferred LFP phase for the control vs. adapting conditions. All LFP phases are divided into two groups of preferred phase $\pm\pi$ and anti – preferred phase $\pm\pi$ where the spike counts in these two groups are compared. **(B)** Histogram representation of spike counts in the preferred phase and anti-preferred phase for control (cyan) and adapting (magenta) conditions, respectively. **(C)** Schematic illustration of spike suppression from the preferred phase of LFP in the adapting condition compared to the control condition.

each 4 Hz frequency window from 1 to 60 Hz are shown in **Supplementary Figure S3A**. Considering 4 Hz frequency windows shows this phase consistency occurs prominently in 16–20 Hz of the beta band (**Supplementary Figure S3B**).

We also measured the preferred phase of LFP based on STA method. To this end, we averaged the LFPs within a window (±20) around each spike occurrence. As depicted in **Figure 4B**, we computed the STA across all recording sites (n = 98) for control (cyan) and adapting (magenta) conditions. The difference between the peak and trough for the control (cyan) and adapting (magenta) conditions shows that the SPC for the control condition is larger. Moreover, **Figure 4B** shows spikes coupled to the falling phase (∼160°) (consistent with **Figure 4A**) of LFP for both conditions.

Spikes Are Suppressed From the Preferred Phase of LFP

Subsequently, we investigated the relation between the spike suppression and the LFP phase. To this end, we compared the spike counts of the preferred phase for control and adapting conditions. Likewise, we conducted this procedure for the anti-preferred phase (**Figure 5A**). As such, one could realize whether the suppression of spikes occurs in the preferred or anti-preferred phase. The difference between the spike counts before and after adaptation (control and adapting conditions) for each of the preferred phase and anti-preferred phase depicts the number of suppressed spikes for the corresponding phase. Results show that a significant reduction of spikes occurs in the preferred phase (Wilcoxon rank sum test, $p < 0.001$) while no visible difference is observed for the anti-preferred phase. **Figure 5B** shows the distributions of preferred and anti-preferred phases for control (cyan) and adapting (magenta) conditions, respectively. Comparing the values of mean spike count in the population of sites (difference$= \sum \text{spk}_{\text{control}} - \sum \text{spk}_{\text{adapting}}$) shows that the suppression of spikes occurs dominantly in the preferred phase of LFP. **Figure 5C** depicts how the preferred phase in LFP is defined according to the spike train and depicts the relation of spiking activity in control and adapting conditions.

DISCUSSION

This study reported that SSA suppresses the spike to LFP phase coupling most prominently in beta range. It also revealed that the adaptation-induced spike reduction occurred for the preferred phase of LFP.

As noted before, SPC shows how activities of spikes are coordinated in the LFPs for different cognitive functions such as sensory coding, attention, working memory, and adaptation (Baker et al., 1999; Kayser et al., 2009; Cutsuridis and Hasselmo, 2011; Muthukumaraswamy, 2011; Mendoza-Halliday et al., 2014; Ruff and Cohen, 2014; Fazlali et al., 2016; Ding et al., 2017; Bahmani et al., 2018). This coupling has also been found to be reduced with attention in visual area V4 (Fries et al., 2001). Also,

it is indicated that the correlation in spiking activity between neurons is reduced by attention (Cohen and Maunsell, 2009). Likewise, the coupling has been documented to encode working memory contents in spike–LFP relation within an area (Bahmani et al., 2018) as well as phase–phase synchrony between two areas (Johnson et al., 2018a,b). It is worth to note that the high frequencies of the LFP such as low- and high-gamma bands reflect more local activity (Csicsvari et al., 2003; Liu and Newsome, 2006; Whittingstall and Logothetis, 2009). On the other hand, the low frequencies of the LFP such as delta, theta, alpha, and beta ranges are a compound signal of slower events from a very large population of cells. Therefore, the activity in the low-frequency bands of the LFP indicates the combination of neural activity across larger networks of neurons (Buzsáki and Draguhn, 2004). As a result, synchronization of spiking activity with LFP in low frequency ranges appears better and with more strength, in different cortical areas, and through various cognitive functions.

The length of the PLV vector represents an estimation of the strength of SPC. It computes a value between 0 and 1 for a given number of spikes for a trial. The limitation of the PLV method is its bias on the number of spikes (Zarei et al., 2018) and studies that use this method, equalize the spikes at a certain count. Therefore, to compute the SPC by this method an equalizing plan was used in order to obtain the spike counts based on a threshold. This study utilized an optimal thresholding scheme which provided an optimal compromise between the number of spikes and the number of trials (Zarei et al., 2018). As such, the trials whose number of spikes were below the threshold were removed, and the trials with their spikes higher than the threshold were equalized to it. It may be argued that how does this method affect the number of trials per condition (control vs. adapting) individually? To address this issue, as depicted in **Figure 2** there is a tradeoff between the threshold value and the number of remaining trials after spike-count equalization. The optimal thresholding selected a threshold somehow the minimum number of trials needs to be removed. **Supplementary Figure S4A** shows the relation between trial numbers in each condition for 10 threshold values (between 1 and 46). It shows that with the threshold value of 14 (spk/bin) no significant difference is visible between two control and adapting conditions. Moreover, we equalized the number of trials between control and adapting conditions before SPC calculation. The SPC strength in beta band shows a significant difference from chance level SPC in threshold values between 7 and 30 spk/bin (**Figure 3** and **Supplementary Figure S4B**). Hence, calculation of SPC by the PLV method revealed the phases of the LFP that spikes have occurred.

The main purpose of this study was to investigate the role of synchrony in neural adaptation mechanism. Here, we examined the effect of adaptation on SPC over the primary auditory cortex area. Commonly, adaptation suppresses the neuronal activities, namely spikes and LFPs in sensory systems. To assess the effect of adaptation on the relation of spikes to LFP phases (SPC), this study calculated the strength of locking between spikes and field potentials by quantifying PLV (Lachaux et al., 1999) in control and adapting conditions.

Recently, the effect of cognitive functions such as adaptation and attention on population synchrony has reported in various brain areas. Adibi et al. (2013) showed that stimulus presentation reduced individual neuron trial-to-trial response variability (captured by Fano factor) and correlations in the population response variability (noise correlation). It has been shown that adaptation conveyed the neuronal operating scheme to lower rates with higher Fano factor and noise correlations (Adibi et al., 2013). Furthermore, Gutnisky and Dragoi (2008) showed that noise correlations are independent of stimulus orientation and caused a strong reduction in correlations after adaptation. Mitchell et al. (2007) found that correlations in spiking averaged are significantly reduced with attention directed into the receptive field. Consistent with previous research in the response function of cortical neurons, our results suggest that the SPC within the beta band is significantly reduced with adaptation (Adibi et al., 2013, 2014; Ponce-Alvarez et al., 2013).

Our results indicate that spikes are coupled to the falling phases of LFP in the primary auditory cortex of the rat (**Figure 4**). This finding is consistent with the previous studies in a different cortical area of the macaque (Lakatos et al., 2008; Whittingstall and Logothetis, 2009). This SPC within the beta band frequency is suppressed by neural adaptation. Furthermore, we found that the reduction of spikes in adaptation occurred at the preferred phase of LFP rather than the anti-preferred phase, or other phases (**Figure 5**). This result provides valuable information toward a better understanding of the underlying neural mechanism of adaptation and could be utilized in the context of biological neural modeling. In addition, this finding may suggest the information coded by single neurons fluctuates relative to the preferred phase of LFP. As a result, the preferred phase of LFP may play a key role in coordinating neuronal spiking activity in different brain functions.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Research with animal subjects represents a small but indispensable component of neuroscience research. The scientists in this study are aware and are committed to the great responsibility they have in ensuring the best possible science with the least possible harm to the animals (Calapai et al., 2017). The main goal of the current study was to elucidate the effect of adaptation on SPC as a crucial component of measuring brain function. All experiments were carried out in Iran Neural Technology Center (INTC) at Iran University of Science and Technology (IUST) with the approval of, and using methods conforming to the standards of the ministry of health and medical education. The animals are maintained in the cages in facilities of the Neural Technology Center. The animal care and use committee of Neuroscience Research Laboratory, IUST

approved all surgical procedures and experimental protocols in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All surgery was performed under urethane pentobarbital anesthesia, and all efforts were made to minimize suffering. At the end of the experimental period, the animal was sacrificed under ether anesthesia (Parto Dezfouli and Daliri, 2015).

AUTHOR CONTRIBUTIONS

MPD, MZ, MJ, and MRD contributed to the concept of the work. MPD performed the experimental recording. MPD and MZ analyzed the data. All authors contributed to drafting, revising, and finalizing the manuscript. MRD and MJ supervised the work.

FUNDING

This work was supported by the Iran University of Science and Technology (IUST) and the Sharif University of Technology.

ACKNOWLEDGMENTS

We would like to thank Prof. Abbas Erfanian [Iran Neural Technology Center (INTC)] for his support of the experimental recording of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncir.2019.00044/full#supplementary-material

FIGURE S1 | An example of original recording and spikes to phase relation. **(A)** Data were recorded with 10 kHz sampling rate to cover both LFP and spike activities. The lower panel shows the raw data of a sample recording site (site 14-control paradigm) vs. time. The upper panel indicates ∼700 ms recording time with the presentation of two pure tones. **(B)** Raw LFP, filtered LFP in beta band, its extracted phase in this band for a sample trial, and spike event times for a three sample trials with the same LFP oscillations in beta band.

FIGURE S2 | The SPC reduction due to adaptation is independent of firing rate. By dividing the trials of adapting sequence into two groups; lower than median and higher than median firing rate, we evaluated the effect of firing rate on SPC value. Results show that the low firing rate (LFR) group leads to more SPC value than high firing rate (HFR) group. Two conditions were selected from the adapting sequence.

FIGURE S3 | SPC for 4 Hz segments and per 4 Hz stepping size. **(A)** Based on Bonferroni criterion (0.05/15–0.003), the SPC within 16–20 and 20–24 Hz shows significant difference between control and adapting conditions. **(B)** Polarity maps of three sample bands (4–8, 16–20, and 48–52 Hz) that indicate dispersed phase for other bands except beta band.

FIGURE S4 | Effects of the threshold value. **(A)** The relation between remained trial numbers in each control and adapting conditions for 10 threshold values (from 1 to 46) in the optimal thresholding plan. It shows that almost the same number of trials were decreased from both control and adapting conditions. Notably, we also equalized the number of trials between control and adapting conditions for SPC calculation. **(B)** SPC strength for control and adapting conditions compared to the chance level of SPC relative to changing of the threshold value.

REFERENCES

Adibi, M., McDonald, J. S., Clifford, C. W., and Arabzadeh, E. (2014). Population decoding in rat barrel cortex: optimizing the linear readout of correlated population responses. *PLoS Comput. Biol.* 10:e1003415. doi: 10.1371/journal.pcbi.1003415

Adibi, M., McDonald, J. S., Clifford, C. W. G., and Arabzadeh, E. (2013). Adaptation improves neural coding efficiency despite increasing correlations in variability. *J. Neurosci.* 33, 2108–2120. doi: 10.1523/JNEUROSCI.3449-12.2013

Anderson, L. A., Christianson, G. B., and Linden, J. F. (2009). Stimulus-specific adaptation occurs in the auditory thalamus. *J. Neurosci.* 29, 7359–7363. doi: 10.1523/JNEUROSCI.0793-09.2009

Bahmani, Z., Daliri, M. R., Merrikhi, Y., Clark, K., and Noudoost, B. (2018). Working memory enhances cortical representations via spatially specific coordination of spike times. *Neuron* 97, 967–979.e6. doi: 10.1016/j.neuron.2018.01.012

Baker, S. N., Kilner, J. M., Pinches, E. M., and Lemon, R. N. (1999). The role of synchrony and oscillations in the motor output. *Exp. Brain Res.* 128, 109–117. doi: 10.1007/s002210050825

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. doi: 10.1038/nrn3241

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Calapai, A., Berger, M., Niessing, M., Heisig, K., Brockhausen, R., Treue, S., et al. (2017). A cage-based training, cognitive testing and enrichment system optimized for rhesus macaques in neuroscience research. *Behav. Res. Methods* 49, 35–45. doi: 10.3758/s13428-016-0707-703

Cohen, M. R., and Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* 12, 1594–1600. doi: 10.1038/nn.2439

Csépe, V., Karmos, G., and Molnár, M. (1987). Evoked potential correlates of stimulus deviance during wakefulness and sleep in cat — animal model of mismatch negativity. *Electroencephalogr. Clin. Neurophysiol.* 66, 571–578. doi: 10.1016/0013-4694(87)90103-90109

Csicsvari, J., Jamieson, B., Wise, K. D., and Buzsáki, G. (2003). Mechanisms of gamma oscillations in the hippocampus of the behaving rat. *Neuron* 37, 311–322. doi: 10.1016/S0896-6273(02)01169-8

Cutsuridis, V., and Hasselmo, M. (2011). Spatial memory sequence encoding and replay during modeled theta and ripple oscillations. *Cognit. Comput.* 3, 554–574. doi: 10.1007/s12559-011-9114-9113

Dean, I., Harper, N. S., and McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nat. Neurosci.* 8, 1684–1689. doi: 10.1038/nn1541

Dezfouli, M. A. P., Dezfouli, M. P., and Rad, H. S. (2014). "A novel approach for baseline correction in 1 H-MRS signals based on ensemble empirical mode decomposition," in *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (Chicago: IEEE).

Ding, S., Meng, L., Han, Y., and Xue, Y. (2017). A review on feature binding theory and its functions observed in perceptual process. *Cognit. Comput.* 9, 194–206. doi: 10.1007/s12559-016-9446-9440

Eckhorn, R., and Obermueller, A. (1993). Single neurons are differently involved in stimulus-specific oscillations in cat visual cortex. *Exp. Brain Res.* 95, 177–182. doi: 10.1007/BF00229667

Fazlali, Z., Ranjbar-Slamloo, Y., Adibi, M., and Arabzadeh, E. (2016). Correlation between cortical state and locus coeruleus activity: implications for sensory coding in rat barrel cortex. *Front. Neural Circuits* 10:14. doi: 10.3389/fncir.2016.00014

Fries, P., Reynolds, J. H., Rorie, A. E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291, 1560–1563. doi: 10.1126/science.291.5508.1560

Grasse, D. W., and Moxon, K. A. (2010). Correcting the bias of spike field coherence estimators due to a finite number of spikes. *J. Neurophysiol.* 104, 548–558. doi: 10.1152/jn.00610.2009

Gutnisky, D. A., and Dragoi, V. (2008). Adaptive coding of visual information in neural populations. *Nature* 452, 220–224. doi: 10.1038/nature06563

Jansen, M., Li, X., Lashgari, R., Kremkow, J., Bereshpolova, Y., Swadlow, H. A., et al. (2014). Chromatic and achromatic spatial resolution of local field potentials in awake cortex. *Cereb. Cortex* 25, 3877–3893. doi: 10.1093/cercor/bhy221

Javit, D. C., Steinschneider, M., Schroeder, C. E., Vaughan, H. G., and Arezzo, J. C. (1994). Detection of stimulus deviance within primate primary auditory cortex: intracortical mechanisms of mismatch negativity (MMN) generation. *Brain Res.* 667, 192–200. doi: 10.1016/0006-8993(94)91496-91496

Johnson, E. L., Adams, J. N., Solbakk, A.-K., Endestad, T., Larsson, P. G., Ivanovic, J., et al. (2018a). Dynamic frontotemporal systems process space and time in working memory. *PLoS Biol.* 16:e2004274. doi: 10.1371/journal.pbio.2004274

Johnson, E. L., King-Stephens, D., Weber, P. B., Laxer, K. D., Lin, J. J., and Knight, R. T. (2018b). Spectral imprints of working memory for everyday associations in the frontoparietal network. *Front. Syst. Neurosci.* 12:65. doi: 10.3389/fnsys.2018.00065

Katz, Y., Heiss, J. E., and Lampl, I. (2006). Cross-whisker adaptation of neurons in the rat barrel cortex. *J. Neurosci.* 26, 13363–13372. doi: 10.1523/JNEUROSCI.4056-06.2006

Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D. L., and Carandini, M. (2009). Local origin of field potentials in visual cortex. *Neuron* 61, 35–41. doi: 10.1016/j.neuron.2008.11.016

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* 61, 597–608. doi: 10.1016/j.neuron.2009.01.008

Lachaux, J.-P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* 8, 194–208. doi: 10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113. doi: 10.1126/science.1154735

Li, X., Chen, Y., Lashgari, R., Bereshpolova, Y., Swadlow, H. A., Lee, B. B., et al. (2014). Mixing of chromatic and luminance retinal signals in primate area V1. *Cereb. Cortex* 25, 1920–1937. doi: 10.1093/cercor/bhu002

Li, Y., Yu, C., Zhou, Z. C., Stitt, I., Sellers, K. K., Gilmore, J. H., et al. (2017). Early development of network oscillations in the ferret visual cortex. *Sci. Rep.* 7:17766. doi: 10.1038/s41598-017-17502y

Liu, J., and Newsome, W. T. (2006). Local field potential in cortical area MT: stimulus tuning and behavioral correlations. *J. Neurosci.* 26, 7779–7790. doi: 10.1523/JNEUROSCI.5052-05.2006

Malmierca, M. S., Cristaudo, S., Perez-Gonzalez, D., and Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *J. Neurosci.* 29, 5483–5493. doi: 10.1523/JNEUROSCI.4153-08.2009

Mendoza-Halliday, D., Torres, S., and Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat. Neurosci.* 17, 1255–1262. doi: 10.1038/nn.3785

Mitchell, J. F., Sundberg, K. A., and Reynolds, J. H. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* 55, 131–141. doi: 10.1016/j.neuron.2007.06.018

Müller, J. R. (1999). Rapid adaptation in visual cortex to the structure of images. *Science* 285, 1405–1408. doi: 10.1126/science.285.5432.1405

Muthukumaraswamy, S. D. (2011). Temporal dynamics of primary motor cortex gamma oscillation amplitude and piper corticomuscular coherence changes during motor control. *Exp. Brain Res.* 212, 623–633. doi: 10.1007/s00221-011-2775-z

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Nelken, I., and Ulanovsky, N. (2007). Mismatch negativity and stimulus-specific adaptation in animal models. *J. Psychophysiol.* 21, 214–223. doi: 10.1027/0269-8803.21.34.214

Parto Dezfouli, M., and Daliri, M. R. (2015). The effect of adaptation on the tuning curves of rat auditory cortex. *PLoS One* 10:e0115621. doi: 10.1371/journal.pone.0115621

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., and Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat. Neurosci.* 5, 805–811. doi: 10.1038/nn890

Ponce-Alvarez, A., Thiele, A., Albright, T. D., Stoner, G. R., and Deco, G. (2013). Stimulus-dependent variability and noise correlations in cortical MT neurons. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13162–13167. doi: 10.1073/pnas.1300098110

Ray, S., and Maunsell, J. H. R. (2010). Differences in gamma frequencies across visual cortex restrict their possible use in computation. *Neuron* 67, 885–896. doi: 10.1016/j.neuron.2010.08.004

Ruff, D. A., and Cohen, M. R. (2014). Attention can either increase or decrease spike count correlations in visual cortex. *Nat. Neurosci.* 17, 1591–1597. doi: 10.1038/nn.3835

Shen, L., Zhao, L., and Hong, B. (2015). Frequency-specific adaptation and its underlying circuit model in the auditory midbrain. *Front. Neural Circuits* 9:55. doi: 10.3389/fncir.2015.00055

Siegel, M., Warden, M. R., and Miller, E. K. (2009). Phase-dependent neuronal coding of objects in short-term memory. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21341–21346. doi: 10.1073/pnas.0908193106

Sirota, A., Montgomery, S., Fujisawa, S., Isomura, Y., Zugaro, M., and Buzsáki, G. (2008). Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron* 60, 683–697. doi: 10.1016/j.neuron.2008.09.014

Taaseh, N., Yaron, A., and Nelken, I. (2011). Stimulus-specific adaptation and deviance detection in the rat auditory cortex. *PLoS One* 6:e23369. doi: 10.1371/journal.pone.0023369

Takaura, K., and Fujii, N. (2016). Facilitative effect of repetitive presentation of one stimulus on cortical responses to other stimuli in macaque monkeys–a possible neural mechanism for mismatch negativity. *Eur. J. Neurosci.* 43, 516–528. doi: 10.1111/ejn.13136

Ulanovsky, N., Las, L., Farkas, D., and Nelken, I. (2004). Multiple time scales of adaptation in auditory cortex neurons. *J. Neurosci.* 24, 10440–10453. doi: 10.1523/JNEUROSCI.1905-04.2004

Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* 6, 391–398. doi: 10.1038/nn1032

Vinck, M., Battaglia, F. P., Womelsdorf, T., and Pennartz, C. (2012). Improved measures of phase-coupling between spikes and the local field potential. *J. Comput. Neurosci.* 33, 53–75. doi: 10.1007/s10827-011-0374-374

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., and Pennartz, C. M. A. (2010). The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. *Neuroimage* 51, 112–122. doi: 10.1016/j.neuroimage.2010.01.073

Whittingstall, K., and Logothetis, N. K. (2009). Frequency-band coupling in surface eeg reflects spiking activity in monkey visual cortex. *Neuron* 64, 281–289. doi: 10.1016/j.neuron.2009.08.016

Womelsdorf, T., Johnston, K., Vinck, M., and Everling, S. (2010). Theta-activity in anterior cingulate cortex predicts task rules and their adjustments following errors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5248–5253. doi: 10.1073/pnas.0906194107

Zarei, M., Jahed, M., and Daliri, M. R. (2018). Introducing a comprehensive framework to measure spike-lfp coupling. *Front. Comput. Neurosci.* 12:78. doi: 10.3389/fncom.2018.00078

Zhou, Z., Zhu, H., and Chen, L. (2013). Effect of aripiprazole on mismatch negativity (mmn) in schizophrenia. *PLoS One* 8:e52186. doi: 10.1371/journal.pone.0052186

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Parto Dezfouli, Zarei, Jahed and Daliri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

REVIEW
published: 20 August 2019
doi: 10.3389/fnins.2019.00878

Discovering Implied Serial Order Through Model-Free and Model-Based Learning

Greg Jensen[1,2]*, *Herbert S. Terrace*[1,3] and *Vincent P. Ferrera*[2,3]

[1] Department of Psychology, Columbia University, New York, NY, United States, [2] Department of Neuroscience, Columbia University, New York, NY, United States, [3] Department of Psychiatry, Columbia University, New York, NY, United States

Humans and animals can learn to order a list of items without relying on explicit spatial or temporal cues. To do so, they appear to make use of transitivity, a property of all ordered sets. Here, we summarize relevant research on the transitive inference (TI) paradigm and its relationship to learning the underlying order of an arbitrary set of items. We compare six computational models of TI performance, three of which are model-free (*Q*-learning, Value Transfer, and REMERGE) and three of which are model-based (RL-Elo, Sequential Monte Carlo, and Betasort). Our goal is to assess the ability of these models to produce empirically observed features of TI behavior. Model-based approaches perform better under a wider range of scenarios, but no single model explains the full scope of behaviors reported in the TI literature.

Keywords: reinforcement learning, model-free learning, model-based learning, cognitive maps, transitive inference

OPEN ACCESS

Edited by:
Ali Ghazizadeh,
Sharif University of Technology, Iran

Reviewed by:
Abdol-Hossein Vahabie,
Institute for Research in Fundamental Sciences (IPM), Iran
G. Elliott Wimmer,
Stanford University, United States

***Correspondence:**
Greg Jensen
greg.guichard.jensen@gmail.com

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: *14 April 2019*
Accepted: *05 August 2019*
Published: *20 August 2019*

Citation:
Jensen G, Terrace HS and Ferrera VP (2019) Discovering Implied Serial Order Through Model-Free and Model-Based Learning. Front. Neurosci. 13:878. doi: 10.3389/fnins.2019.00878

INTRODUCTION

Transitivity is a property of all ordered sets, including number systems, social hierarchies, rational economic preferences, and spatial position. Exploiting transitivity can reduce task complexity when learning serial order, since knowing that X > Y and Y > Z is sufficient to infer that X > Z. Using this implied information is called transitive inference (TI) and is thought to underlie some forms of serial learning. In this review, we describe and implement six reinforcement learning algorithms for serial learning. Three are model-free, and depend on expected reward value to make their judgments. The other three are model-based, and represent the rank associated with each stimulus using a spatial continuum. These algorithms are compared side-by-side using simulation under identical experimental conditions. The performance of the algorithms illustrates the utility of one-dimensional cognitive maps for representing abstract relationships.

Serial learning is ubiquitous, but the mechanisms underlying TI remain poorly understood. In humans, reasoning about serial order or rank has been studied for over 100 years, with a central role in the history of intelligence testing (Burt, 1911) and child development (Piaget, 1921). Following the demonstration that squirrel monkeys can perform TI (McGonigle and Chalmers, 1977), a wealth of results with animal subjects has revealed the breadth of serial learning. To date, every vertebrate species tested has shown evidence for some form of TI (Vasconcelos, 2008; Jensen, 2017). Given this broad comparative literature showing serial learning under various experimental procedures (Terrace, 2012), it seems reasonable that these abilities reflect cognitive mechanisms that have deep evolutionary roots.

Despite the ubiquity of this phenomenon, no consensus has emerged about a mechanism for inferring serial order. Some explain serial learning in animals using model-free learning

(e.g., Vasconcelos, 2008). In the context of TI, we use the term "model-free learning" to refer to an algorithm that estimates a probability or rate of some outcome, conditional only on past observed events and on cues discernable in the environment. For example, a subject might learn that p(food) is low because food is rarely available in its past experience, but also learn that p(food | buzzer) is high because food is often made available while the buzzer is audible. Given sufficient training, even naïve conditional probability estimation of this kind can solve a wide variety of tasks without any deeper understanding of causal relationships.

Another approach for explaining behavior consistent with TI is "model-based learning," in which observed events are presumed to reflect some underlying set of rules or relations. Subjects make additional assumptions about what the stimuli mean and how they relate to one another, and these assumptions support more complex inferences. Although cognitive models of serial learning often follow the themes of model-based learning, few are described in sufficient detail to simulate behavior, making it difficult to compare cognitive models to their model-free counterparts (Gazes et al., 2012). For serial learning, we believe that the most promising approach is to assume that each item in the list has some "spatial position" along an abstract continuum (Jensen et al., 2013, 2015).

Solving TI using a spatial representation is akin to an abstract cognitive map (Redish, 1999), an organizational scheme that could be used for much more than spatial navigation (Behrens et al., 2018). According to this proposal, positions of stimuli are "mapped" along a single dimension, a capacity that has been attributed to hippocampal computation (Eichenbaum et al., 2016). Recent studies on the computational capabilities of circuits in hippocampus (Oliva et al., 2016) and entorhinal cortex (Bellmund et al., 2016) further suggest that spatial modeling is a common substrate across many different cognitive domains (Constantinescu et al., 2016).

Viewed as spatial computations, TI problems become an important limiting case. In open fields (Chalmers et al., 2016) or radial arm mazes (Ferguson et al., 2019), the success of reinforcement learning often depends on physical and temporal cues. These allow a model-free algorithm to "navigate a space" without encoding a representation of that space, because the cues provide many ways to discover rewards. For example, early difficulties experienced by model-free learning of the 1983 Atari game *Ms. Pac-Man* (e.g., Mnih et al., 2015) have largely been overcome by making better use of the information on screen (e.g., van Seijen et al., 2017). In effect, the environment bears the burden of representing itself, freeing the agent from having to do so using an encoding scheme.

Algorithms of this kind struggle when task information is sparse. The 1984 Atari game *Montezuma's Revenge*, for example, remains largely unsolved by reinforcement learning because it provides the user with very few informative cues (Justensen et al., 2019). Similarly, TI tasks present the bare minimum of information that only implies the list order, while other cues (such as time, stimulus position, and trial order) are carefully controlled to provide no information usable by the subject. As a result, TI procedures eliminate the cues that make open field problems with spatial landmarks solvable by model-free methods. TI tasks thus have the potential to reveal clues about the machinery that organisms use for spatial cognition in general, beyond the narrow scope of ordered lists.

In this review, we will first characterize the problem posed by TI tasks (see Section Problem: How to Infer Serial Ordering from Pairwise Comparisons), then describe three model-free learning algorithms (see Section Model-Free Solutions: Expected Value Estimation) and three model-based learning algorithms (see Section Model-Based Solutions: Inference About Ordering) that have been suggested as models of this kind of inference. We then recover the relevant algorithm parameters from empirical data and use them to perform simulations to evaluate how well these algorithms do under a variety of experimental conditions (see Section Simulation of Serial Order Tasks). We conclude with a brief model comparison analysis (see Section Model Comparison and Ensemble Modeling), a note on computational complexity (see Section A Note on Computational Complexity), and our concluding thoughts (see Section Conclusion).

PROBLEM: HOW TO INFER SERIAL ORDERING FROM PAIRWISE COMPARISONS

When organisms solve the TI problem, they in effect perform a statistical procedure: Inferring ranks of items in an ordered list, based on partial information provided by a series of pairwise comparisons. Such statistical inferences can be accomplished with a variety of strategies (Bürkner and Vuorre, 2018). Because many candidate algorithms exist, above-chance performance in a TI task is not sufficient to determine how organisms solve the task. Instead, common features of behavior that accompany this inference provide vital clues about its underlying computations.

Experimental Evidence

In the standard TI task, subjects are presented with pairs of stimuli drawn at random from an ordered list, and are rewarded for selecting the "correct" item. A stimulus is correct if its rank is lower than the distractor with which it is paired. So, for example, given the 7-item ordered list ABCDEFG (e.g., **Figure 1A**), the stimulus C is a correct answer in the context of the pair CD, but is incorrect in the context of the pair BC (**Figure 1B**).

Although there are 21 possible pairs in a 7-item list, training all the pairs in parallel is not sufficient to conclude that a "transitive inference" has occurred. A model-free learning mechanism that tracks the expected value of each stimulus can succeed in an all-pairs design, because A is correct in more pairs than B, which in turn is correct in more pairs than C, and so forth. To provide a strong test that an organism's inference of list order depends on transitivity (rather than comparing each item's association with obtained rewards), the classic test of TI is to train *only* the adjacent pairs (in this case, AB, BC, CD, DE, EF, and FG). This contrast between training sets and testing sets is shown in **Figure 1C**. Under adjacent-pair training, terminal items A and G are associated with perfect information (100 and 0% rewards, respectively), while all non-terminal items are correct in 50%

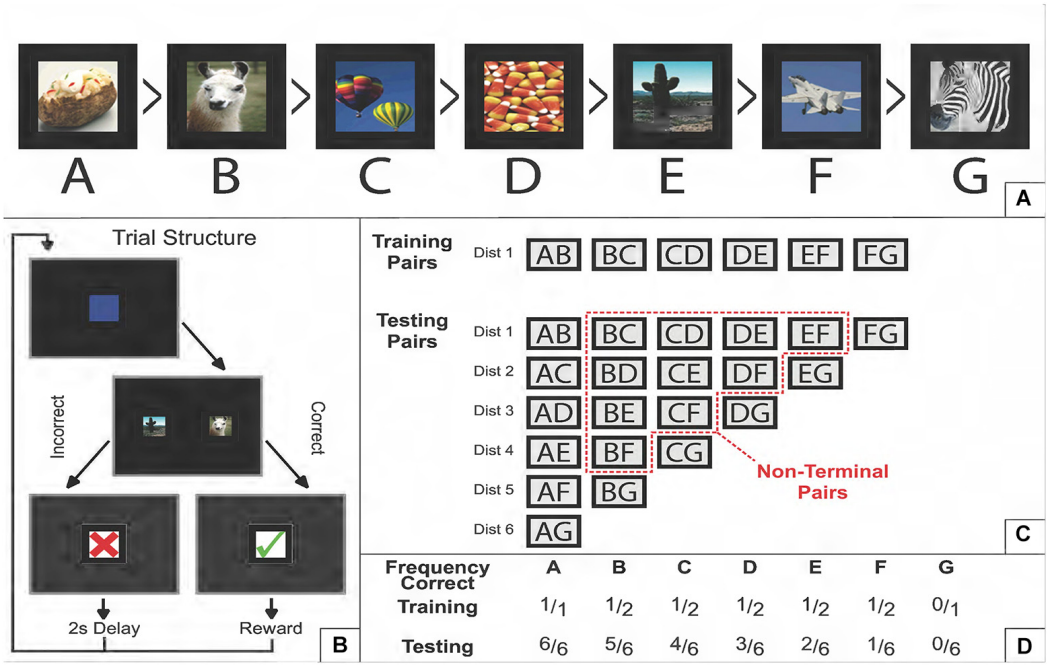


FIGURE 1 | Schematic description of a transitive inference task. **(A)** Example of a 7-item ordered list of photographic stimuli. "Greater Than" symbols denote which items dominate which other items (e.g., baked potato is correct relative to every other stimulus, but fighter plane is only correct when paired with zebra). **(B)** Example of a single trial. Following a start stimulus, two stimuli appear on screen. If the correct item is selected, positive feedback is provided; an incorrect choice leads to negative feedback. In every trial, there is always one correct and one incorrect answer. No more than two stimuli are ever presented at a time for the duration of the experiment. **(C)** The training and testing sets of pairs of stimuli. During training, only adjacent pairs (AB, BC, etc.) are presented typically in a randomized order that counterbalances how often each stimulus appears on each side of the screen. During testing, all 21 pairs are presented. The "symbolic distance" (denoted as "Dist") refers to how many steps in the list one has to make to get from one item to another. Adjacent pairs have a symbolic distance of 1, whereas pairing both terminal items has a distance of 6. The "non-terminal pairs" are those that exclude both the first and last list item. If a non-terminal pair was also excluded from the training set, then it considered is a "critical pair" at test. **(D)** The frequency with which each stimulus is correct, relative to the number of pairs it appears in during that phase. During training, non-terminal items are correct 50% of the time, and consequently comparisons of the overall reward rate for non-terminal are equivocal, given only the information available during training. Reproduced with permission (DOI: 10.6084/m9.figshare.7992005.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

of the training cases (**Figure 1D**). Upon completion of training, B and D have identical expected values, providing no clue about which response is correct. We henceforth refer to pairs that are non-adjacent (i.e., outside the training set) and non-terminal (i.e., ambiguous with respect to expected value) to be the *critical pairs*. If a subject exceeds chance on those pairs, some inference beyond mere reward associations must have taken place. In a 7-item list, the six critical pairs are BD, CE, DF, BE, CF, and BF.

Aside from above-chance performance on the critical pairs, two other major clues are routinely evident in studies of TI. The first of these is the *terminal item effect* (Wynne, 1997; Acuna et al., 2002), which occurs when pairs that include terminal items yield more correct responses. Since terminal items have an unambiguous reward history (e.g., A is always correct), these discriminations should be easier because of their differential reward associations, even if a non-associative mechanism is also at work. Terminal item effects are usually visible throughout training and testing. A model purporting to explain how subjects perform TI should display these effects.

The second consistent behavioral phenomenon is the *symbolic distance effect* (D'Amato and Colombo, 1990), which is present when the difference in the ranks of two items (or the pair's "symbolic distance") is associated with response accuracy. Non-human studies of TI almost always display lower accuracy for pairs of adjacent items than for pairs that are spaced apart more widely in the list (Jensen, 2017), even when adjacent pairs are extensively trained. Unlike terminal item effects (which can arise from reward associations), symbolic distance effects provide compelling evidence of a cognitive strategy (Terrace, 2010; Jensen et al., 2013), particularly when those distance effects can be seen among critical pairs that did not appear in the training set (e.g., Jensen et al., 2017; Kao et al., 2018).

One of the surprising predictions of a positive symbolic distance effect is that response accuracy should be lower for adjacent pairs than for any other distance, despite extensive training on those pairs. In non-human animals, even when training consists of tens of thousands of presentations of a small set of adjacent pairs (e.g., Merritt et al., 2007; Lazareva and Wasserman, 2012), those pairs nevertheless have the highest error rates. In some cases, response accuracies appear barely above chance throughout training, only to leap to higher accuracies as soon as pairs with larger symbolic distances are tested (e.g., Jensen et al., 2015). These results, particularly the high starting accuracy for non-adjacent pairs, suggests a

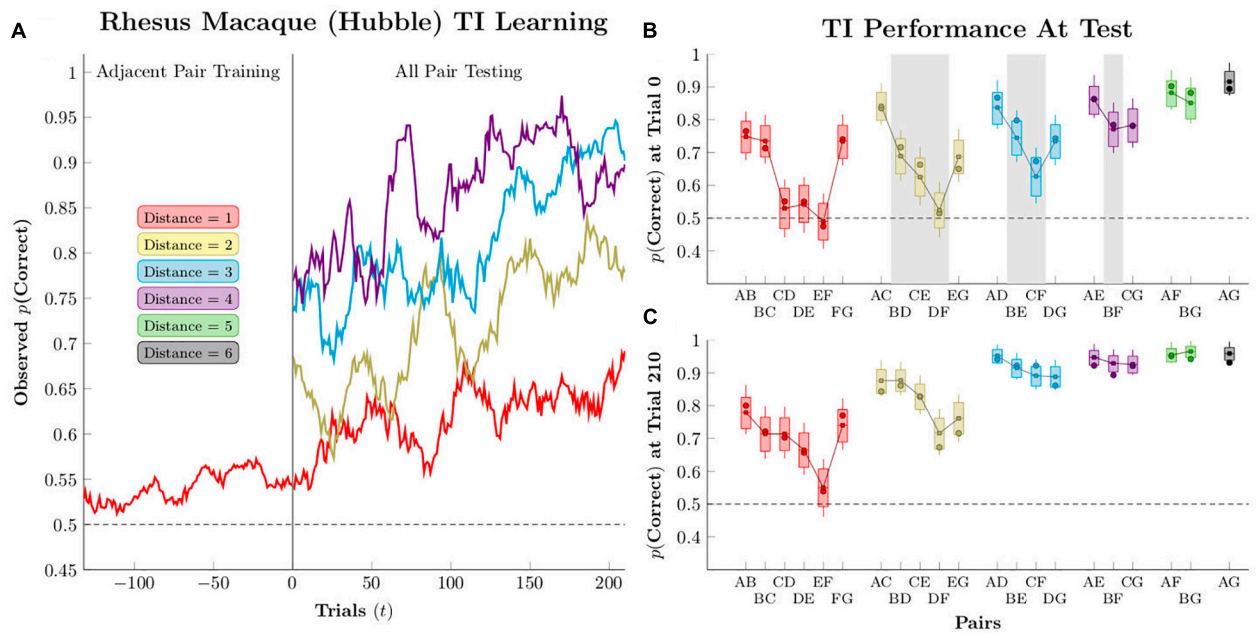


FIGURE 2 | Estimated performance for a single subject (Hubble), as reported by Jensen et al. (2015). Circles represent empirical averages of observed performance, while box-and-whisker plots represent the estimated performance according to pairwise logistic regression. Boxes represent the 80% credible interval, whereas whiskers represent the 95% credible interval. **(A)** Mean response accuracy for non-terminal pairs of distances 1–4, smoothed over a 21-response moving window **(B)** Estimated performance at the start of testing ("trial zero"). Pairs shaded with gray are the "critical pairs" that are expected to remain at chance if subjects are engaged only in a reward association learning strategy; if their performance is above chance, it suggests that an actual inference is being performed. **(C)** Estimated performance after 210 trials of testing. Reproduced with permission (DOI: 10.6084/m9.figshare.7992008.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

representation that approximates the overall order successfully but retains uncertainty about the positions of individual items.

Figure 2 gives an example of typical TI performance (from Jensen et al., 2015). This monkey ("Hubble") was given 150 to 300 trials of training on adjacent pairs (red points) in 7-item lists. Response accuracies were fit for each pair independently using logistic regression in the Stan programing language (Carpenter et al., 2017), as described in the **Supplementary Material**. Using the models for each of the pairs, **Figure 2** projects the accuracy on the first trial of testing ("trial 0"), when each non-adjacent pair had not been seen before, as well as the last trial of testing ("trial 210"), at which point all pairs were familiar. The critical pairs (highlighted in gray) are above chance in five out of six cases at trial 0, and pairs with a larger symbolic distance (e.g., BF) tend to display higher performance than those with smaller symbolic distances (e.g., CE).

MODEL-FREE SOLUTIONS: EXPECTED VALUE ESTIMATION

Model-free approaches seek to explain TI (and by extension, serial learning) purely in terms of observable associative factors, making no reference to internal states or to representations of ordering (Wynne, 1995; Vasconcelos, 2008). These theories assume that model-free learning can account for above-chance performance on critical pairs. In studies of this kind, the phrase "transitive inference" is used to identify the behavioral phenomenon of above-chance performance on critical pairs, not to imply a logical inference *per se*.

Most models of this kind choose stimuli on the basis of the "associative strength" each has with rewards. That is, model-free learning relies exclusively on using feedback from the environment to evaluate the relative frequency of reward as a function of past experience. Experienced value is taken to predict future expected value, and choices should thus be made to maximize future rewards.

Acting to maximize rewards can be deceptively powerful, even in the absence of a representational (i.e., model-based) framework for understanding the task at hand. As such, we must be cautious about interpreting results that are consistent with a cognitive account. In some cases, TI results can also be explained by model-free accounts. We consider three model-free algorithms below.

Q-Learning: Expected Value Weighted by Recency

One of the simplest incarnations of model-free learning that is relevant to TI problems is *Q*-learning (Watkins and Dayan, 1992). As it is typically implemented, *Q*-learning uses both a retrospective and a prospective learning mechanism. The retrospective element maintains a running average of the outcomes for each of its responses, weighted to favor recent events according to an exponential function (Glimcher, 2011). This retrospective element is identical to the "associative

strength" estimated by the Rescorla–Wagner model (Widrow and Hoff, 1960; Rescorla and Wagner, 1972), with roots in the 'mathematical model of simple learning' proposed by Bush and Mosteller (1951). Its prospective element uses a discounted projection of the future state to which the current choice leads. Although *Q*-learning cannot explain TI in a well-controlled experiment, it provides a useful baseline against which to compare other algorithms.

Under *Q*-learning, each behavioral alternative has a numerical "quality," which corresponds to the value or utility expected from performing that behavior, contingent on the "state" the actor is in. Typically, the algorithm's state is determined by environmental context, such as position in space or the presence of an informative cue. At a time t, the value associated with an action a_t, given some contingent contextual state s_t, is denoted by $Q(a_t|s_t)$ (to be read as "the value of action a_t, given the state s_t"). When a choice is made, its consequence c_t is used to update the value of Q according to the following function:

$$\begin{aligned} Q(a_t|s_t) \leftarrow\ & Q(a_t|s_t) \\ & +\delta[c_t + \gamma(\max(Q(\text{all } a_{t+1}|s_{t+1}))) - Q(a_t|s_t)] \quad (1) \\ & \alpha \in (0,1),\ \gamma \in (0,1) \end{aligned}$$

The term c_t is sometimes called the "immediate reward," but c_t may also take on a negative value in the event that the consequence is aversive.

Q-learning's value updating function has two parameters. The 'learning rate' δ controls retrospective learning, and governs how quickly the actor forgets the old information encoded in $Q(a_t|s_t)$ in favor of new information. When δ is 0.0, the actor cannot learn from feedback; when it is 1.0, it discards all past information every time it observes a new outcome. Giving δ a low value (e.g., below 0.2) results in the gradual progress typical of trial-and-error learning.

The 'discount factor' γ controls *Q*-learning's prospective learning, and governs the influence that future states have on estimates of the value of the current response. The quantity $\max\left(Q\left(\text{all } a_{t+1}|s_{t+1}\right)\right)$ refers to the value of the best projected response option, contingent on the actor's state in the next time step, s_{t+1}. This also contributes to the value of Q because a current choice might bring the actor *closer* to a desirable contextual state, even if that choice doesn't yield an immediate reward (e.g., moving the agent closer to the exit of a maze in order to hasten the receipt of a reward upon exiting). If $\gamma = 0.0$, *Q*-learning is identical to Rescorla–Wagner-style associative learning:

$$Q(a_t|s_t) \ \leftarrow\ Q(a_t|s_t) + \delta\left[c_t - Q(a_t|s_t)\right], \quad \alpha \in (0,1) \quad (2)$$

Despite its extreme simplicity, *Q*-learning can solve many difficult problems if given sufficient training and a sufficiently complex matrix of states and actions. This is because, when an actor's choices (e.g., "move forward") influence its contextual state ("distance and direction of food"), the algorithm can discover not only which actions yield rewards, but can extrapolate backwards from those rewards to discover which *sequences* of behavior lead to rewards. However, *Q*-learning is at a particular disadvantage in the case of TI tasks because most experimental designs predetermine and randomize the presentation order of stimuli, such that the actor's choice a_t has *no effect whatsoever* on the subsequent state s_{t+1}. If, for example, a subject sees the stimulus pair AB on the first trial, its choice has no influence on which pair is presented on the second trial, or on which side of the screen each stimulus appears. Put another way, because each future state s_{t+1} does not depend in any way on a_t, the value of $\max\left(Q\left(\text{all } a_{t+1}|s_{t+1}\right)\right)$ in Equation 1 is entirely determined by the task structure, and not by the actor's choices. TI is thus akin to a maze in which, after each choice, the experimenter teleports the subject to another part of the maze.

Using *Q*-learning to try to solve TI tasks also requires a decision by its designer about what constitutes a "state." If, for example, we treat each unique stimulus pairing (e.g., "BC" or "CD") as a unique state, then the actor will treat $Q(\text{C}|\text{BC})$ as being entirely unrelated to $Q(\text{C}|\text{CD})$. When presented with a stimulus pairing it has never previously observed (e.g., BD), such an actor would always begin at chance, with no possibility of generalizing. We are interested in an algorithm that *can* generalize (since this is the only way this algorithm has any hope of solving novel scenarios), we have coded our *Q*-learning implementation to ignore context. The result is that Q is a vector, with one value for each response alternative.

Finally, the designer must decide whether to update response alternatives that are *not* chosen. For example, if the actor is presented with BC and chooses B (yielding a reward), it should update $Q(\text{B})$ using $c_t = 1.0$, but should it *also* update $Q(\text{C})$ using $c_t = 0.0$? A *Q*-learning algorithm that only updates the chosen response is hereafter identified as "asymmetric." Asymmetric learning is commonly assumed in the behavior analysis literature, because the tradition of behaviorism insists that behavior should only be explained by observables, not by counterfactuals. We call algorithms that update the values of both choices "symmetric," since they exploit the either/or symmetry of the TI task, effectively extracting two trials worth of information from each trial. We feel this choice can be justified on the grounds that subjects familiar with the task have ample reason to believe that feedback is not probabilistic, and thus that every trial has one correct and one incorrect answer. Unless otherwise indicated, we assume symmetrical updating.

Still missing from this algorithm is the *action policy*, the rule by which values Q are translated into actions. Our implementation of *Q*-learning makes use of the softmax decision rule (Sutton and Barto, 1998), originally due to Luce (1959). This gives the probability of taking an action a_t, from the set of all actions available at that time A_t, as follows:

$$p(a_t|s_t) = \ \text{softmax}(Q,\theta) = \frac{\exp(\theta \cdot Q(a_t|s_t))}{\sum_{i \in A_t} \exp(Q(i|s_t))}, \quad \theta \geq 0.0 \quad (3)$$

This action policy is governed by one parameter θ, which determines how strongly the response alternatives contrast with one another. When θ is large, the policy responds almost exclusively to the choice with the highest expected value. As θ gets

smaller, the odds of non-maximizing behaviors increase, until $\theta = 0.0$ yields equiprobable chance responding.

The Value Transfer Model: Value by Association

Following the demonstration of TI on monkeys by McGonigle and Chalmers (1977), there was heated debate over the correct definition and implications of "animal intelligence" (Macphail, 1987). TI in non-human animals was particularly provocative because it could not be easily explained by the reward associations that had dominated behavior analysis (and thus animal psychology generally) in the preceding decades. The "Value Transfer Model" (VTM) was proposed as a counterargument against a cognitive account (von Fersen et al., 1991), ostensibly providing an associative account of how TI might arise.

Under VTM, stimuli become associated with rewards according to a Rescorla–Wagner mechanism, and the strength of this association governs the relative rate at which stimuli are selected. However, they *also* accrue value from the associative strength of the stimuli they are paired with. In other words, value "transfers" between stimuli that appear together. Like most behavior-analysis studies of its time, this was demonstrated by first collecting data and then fitting a function to the resulting summary statistics. Under this analysis, VTM appears able to describe behavior in the classical experimental demonstrations of TI in animals (Wynne, 1995). However, VTM was *not* originally conceptualized as a process model that could be used in simulation.

We implemented VTM as a process model using the design described by Kumaran et al. (2016). This builds directly on the retrospective formalism described in Equation 2, and introduces an additional term τ to the value updating function:

$$Q(a_t|s_t) \leftarrow Q(a_t|s_t) + \delta\left[c_t + \tau\left(\sum Q(\neg a_t|s_t)\right) - Q(a_t|s_t)\right]$$
$$\alpha \in (0,1), \tau \in (0,1) \quad (4)$$

Here, the parameter τ corresponds to rate of "transfer" between stimuli and $\sum Q(\neg a_t|s_t)$ refers to the sum of all values of Q for the actions that were available to the actor but were *not* chosen. If $\tau = 0.0$, the function is identical to Equation 2. As τ grows, the value of any given Q grows merely by association with the other presented stimuli, *independent* of the consequences of choosing a_t. This implementation uses the softmax rule (Equation 3) as its action policy. VTM may be implemented symmetrically (updating every response option) or asymmetrically (updating only the chosen item). Unless otherwise noted, we used symmetrical updating.

Although some variant of VTM is typically presented as the go-to "behaviorist alternative" to cognitive theories of TI (e.g., André et al., 2012), VTM makes incorrect predictions under a wide range of experimental preparations (e.g., Weaver et al., 1997; Daniels et al., 2014; Vasconcelos and Monteiro, 2014). A particularly important failure was reported by Lazareva and Wasserman (2006, 2012), who introduced a block of trials between training and testing that consisted exclusively of massed presentation of a single pair of stimuli. This design did not undermine performance by subjects, despite VTM's predictions that there should be a dramatic increase in error rates. We demonstrate this failure, among others, in the simulations below.

REMERGE: Value by Configural Similarity

Although VTM is the most commonly cited associative model of TI, another approach was proposed at about the same time. Couvillon and Bitterman (1992) proposed that, rather than TI arising from cross-talk between reward associations, it could instead be explained by the similarity of the contexts in which behavior was occurring (i.e., the stimulus pairings). Under this "configural" theory of TI, subjects simultaneously learn about response alternatives *and* about stimulus pairs. The configural similarity of stimuli provides clues that could be used to exceed chance performance for novel pairs.

This theme of context configuration has been a long-standing proposal in computational neuroscience for how spatial or temporal processing might be implemented in the hippocampus (Wu and Levy, 1998; Rodriguez and Levy, 2004). In the context of the TI task, its most fully realized form is the REMERGE model, due to Kumaran and McClelland (2012).

REMERGE is a neural net model with three layers, as depicted in **Figure 3**. The "Feature layer" receives inputs from the environment, with each item in the list receiving its own node. The "Conjunctive layer" consists of one node for each stimulus pair. Each Feature node excites every Conjunctive node associated with a pair that its stimulus belongs to, *and* each Conjunctive node stimulates both Feature nodes associated with its pair of items. Meanwhile, activity in the Conjunctive layer stimulates nodes in the "Response layer," with one node associated with each response alternative. Each Conjunctive node sends excitatory signals to the Response node associated with a correct response while also sending an inhibitory signal to the Response node associated with an incorrect response. The mutually excitatory feedback between the Feature and Conjunctive layers is renormalized until a steady state is reached, and at that point the relative activity of nodes in the Response layer provide the basis for the action policy to select one of the stimuli. The details for this implementation are given in the **Supplementary Material**.

As described by Kumaran and McClelland (2012), REMERGE has two free parameters governing the activity of the network; λ and ω govern the behavior of the feedback between network nodes. λ denotes the "temperature" of the network, which governs how much activity in the network is permitted to vary as it converges toward a steady state. ω corresponds to a constant "saturation" variable that determines how much hedging should occur in the activity of the conjunctive layer, effectively putting an upper limit on the resulting renormalized values. Kumaran and McClelland (2012) also give one free parameter that governs choice behavior, the θ parameter of a softmax action policy. This policy is identical to that given by Equation 3, except that instead of using Q, REMERGE applies softmax to the activity associated with all nodes in the Response layers that are available to the actor during the current trial. Our symbols λ, ω, and θ correspond to the parameters given by Kumaran and McClelland as $\frac{1}{\tau}$, C, and

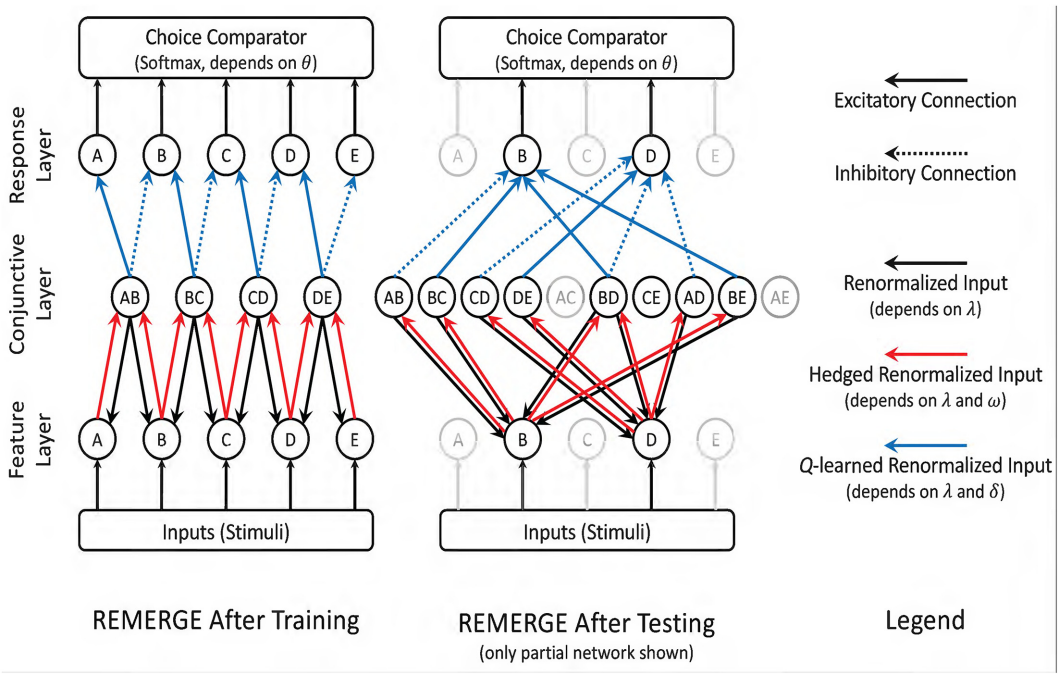


FIGURE 3 | Schematic depiction of RL-REMERGE. Nodes in the Feature layer are initially stimulated by inputs (i.e., if a stimulus is present, its Feature node is excited). All solid-line connections are excitatory, whereas dotted-line connections are inhibitory. When stimuli in the Feature layer are excited, this flows through excitatory connections to the Conjunctive layer, with one node for each pairing the algorithm has previously observed. Excitation then flows back to the Feature layer, and so forth. Excitation entering the Conjunctive layer (red arrows) is "hedged" during renormalization, leading to dissipating excitation for nodes that are many steps removed from environmental input. Once the feedback between the Feature and Conjunctive layers has stabilized, excitatory and inhibitory inputs to the Response layer are summed, and a choice is made using the softmax rule. Because REMERGE is initially ignorant about the values between the Conjunctive and Response layers, those arrows (in blue) begin with values of 0.0 and their excitatory or inhibitory strength is subsequently discovered using *Q*-learning. In this scenario REMERGE learned a 5-item list. **(Left)** Following training, the Conjunctive layer only has four nodes, corresponding to the four adjacent pairs. This is sufficient to infer that B > D because information flow between the Feature and Conjunctive layers when B and D are externally stimulated yield a higher overall level of excitation for B than it does for D. **(Right)**. Partial network depicting *only* those nodes that are directly connected to options B and D. Since all ten pairs have now been observed, every stimulus in the Feature layer has links to four Conjunctive layer nodes. In practice, information flow still passes through all Feature and Conjunctive nodes, but exposure to all stimulus pairs makes that diagram too complex to depict in a figure. Reproduced with permission (DOI: 10.6084/m9.figshare.7992011.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

β respectively, but we have elected not to use their notation to minimize confusion with other symbols used in this manuscript.

Despite being a model of TI, REMERGE was *not* a model of learning in its original presentation, because the arrangement of the network presumed that training was complete and sought only to predict the actor's behavior on the next trial. In order to make REMERGE a learning model, we made two changes. The first change was that each node in the Conjunctive layer was excluded from calculation until the first trial in which that stimulus pair was presented. Thus, given training only on adjacent pairs, REMERGE would only be able to make inferences using those pairs. This explains the difference between the two versions of the network depicted in **Figure 3**. Following training, the Conjunctive layer has only four nodes because it has only been exposed to four combinations of stimuli. Contrastingly, after testing, the Conjunctive layer includes a node for each of the ten possible pairs. This is because by that time the network has been exposed to all of those explicit cases.

Our second change was that rather than hard-coding the links between the Conjunctive and Response layers with the correct and incorrect responses, we require that REMERGE learn the values of those links by *Q*-learning (where incorrect responses were given a value of −1, rather than 0, to ensure incorrect links would become inhibitory), with a corresponding learning rate δ. Thus, our "reinforcement learning" version of REMERGE had four free parameters in total.

REMERGE straddles the line between model-free and model-based learning. Although it does not implement a model of the list *order*, it does implement a kind of logical syllogism. As **Figure 3** depicts, the presentation of the pair BD allows these two stimuli to be indirectly associated by information flowing back and forth through the Feature and Conjunctive layers (following the path B ↔ BC ↔ C ↔ CD ↔ D). Importantly, however, this syllogism is subject to error propagating through the network. In particular, due to the hedging term ω, activity flowing along longer sequences of nodes (i.e., inferences involving response pairs with a larger symbolic distance) tends to wash out. The obvious prediction that follows from this is that REMERGE should display a *negative* symbolic distance effect at transfer, the opposite of the pattern observed in the empirical data (e.g., in **Figure 2**).

MODEL-BASED SOLUTIONS: INFERENCE ABOUT ORDERING

At the core of the debate over whether TI is better explained by model-free or by model-based learning is a disagreement about parsimony. Advocates of model-free TI (such as VTM) see the invocation of "mental representations" as a needless complication, whereas advocates of cognitive models argue that even though they are generally more computationally complex, they are nevertheless more plausible in light of their generality, as well as support from a host of experimental studies (Terrace, 2012).

Despite this, most cognitive models of serial learning have been conceptual in nature, consisting of qualitative descriptions of processes that give rise to group averages. Much of the empirical support cited for these theories relies on the failure of associative models to account for the available experimental results (Gazes et al., 2012). Assessing the complexity of these cognitive theories is difficult because most cognitive models of serial learning have not been characterized in computational terms. This paucity of model-based learning algorithms that are based on cognitive theories may be in part due to the difficulty of designing a plausible model-based algorithm for *learning* TI that does not presuppose the list order. As such, in order for a model-based algorithm to provide a satisfactory description of serial learning, it must be able to begin with novel stimuli, deduce the correct order given some pattern of stimulus pairings, and display the characteristic pattern of errors seen in real subjects while doing so.

First, it is important to characterize the range of possibilities for a model-based algorithm. Since TI is presumed to be an inference about the order of the items in a list, an algorithm that bases its response on *any* model of ordering, rather than on expected value, may be considered a cognitive model. The simplest such model is simply an ordered list that is able to adjust its ordering on the basis of feedback.

RL-Elo: Semiparametric Rank Estimation

The Elo rating system was first developed for use in chess (Elo, 1978) and has since been used in a variety of other competitive settings. Each competitor has a rating (e.g., Garry Kasparov retired from chess with a rating of 2812), which can be compared to the rating of others in order to predict who is more likely to win in a match-up. When a sanctioned match between players occurs, the amount by which each player's score changes is a function of how surprising the result is, relative to the rating's prediction. Although designed for human competition, the dynamic character of the Elo rating system (able to change over time as new information comes in) has made it a popular model for other scientific problems of ordinal rank, such as the analysis of dominance hierarchies in animal groups (Neumann et al., 2011).

To adapt this approach to the problem of TI, Kumaran et al. (2016) implemented a reinforcement learning version of this system, which they called "RL-Elo." Under this system, each stimulus was assigned a score of 0.0 in a vector V at the start of training, and choices were made according to the softmax rule (Equation 3) applied to the elements of V that were associated with the stimuli visible during each trial (with a corresponding free parameter θ). When ratings are evaluated on a pairwise basis (e.g., a_t and $\neg a_t$), this reduces to the logistic function, defined in terms of the difference between the two values of V:

$$p(a_t) = \frac{1}{1 + \exp(-\theta(V(a_t) - V(\neg a_t)))} \tag{5}$$

This action policy is effectively the softmax rule (Equation 3), albeit limited to two alternatives. However, rather than updating the stimuli as a function of their expected value, RL-Elo updates them with respect to their expected probability of a correct response. Thus, if a_t is the correct response, then given the probability that a correct response $p(a_t)$,

$$\begin{aligned} V(a_t) &\leftarrow V(a_t) + \delta\left(1 - p(a_t)\right) \\ V(\neg a_t) &\leftarrow V(\neg a_t) + \delta\left(p(a_t) - 1\right) \\ \delta &\in (0, 1) \end{aligned} \tag{6}$$

Here, the learning rate δ makes another appearance, but this time it is being used to update the *score* associated with a stimulus, without consideration for its expected value. Consequently, the engine that powers RL-Elo is not expected value, but rather expected difference in rating. Under this implementation, RL-Elo updates stimuli symmetrically. Since this difference is effectively tracked as a difference on the logit scale, the vector V can be seen as a semiparametric model for estimating the ranks of the stimuli. From this perspective, $V(a_t)$ makes no distinction between a representation of the rank of a stimulus and the preference for selecting that stimulus during trial t. The result is a gradient-following algorithm (Williams, 1992) that adjusts the scores in V until such time as difference in the scores between any two items yields close to perfect discrimination of their rank. The logic of this updating scheme is depicted in **Figure 4** (left).

Sequential Monte Carlo: Iterative Crowdsourcing

Most statistical analyses of ordinal rank adopt a semiparametric strategy, as RL-Elo does. However, estimating some scalar quantity as a proxy for the ordinal rank has limitations. One is that, like most model-free strategies, adjusting the values associated with stimuli only after they were presented can be unreliable when some item pairs are presented more often than others (i.e., unequal base rates of presentation). That limitation is overcome by "Sequential Monte Carlo" (SMC) algorithms (Kumaran et al., 2016). Sometimes called "particle filter" algorithms, these provide an approximation to Bayesian reasoning about ordinal rank similar to the approach used by a Kalman filter (Doucet et al., 2000). The details of its implementation are given in the **Supplementary Material**.

Rather than consider a single vector of scalar values, as RL-Elo does, SMC simultaneously considers a very large number of them. Our implementation uses 10,000 such "hypothetical orderings," generated randomly from a normal distribution (with variance $\sigma^2_{\emptyset}$) at the outset of training. Rather than adjust those

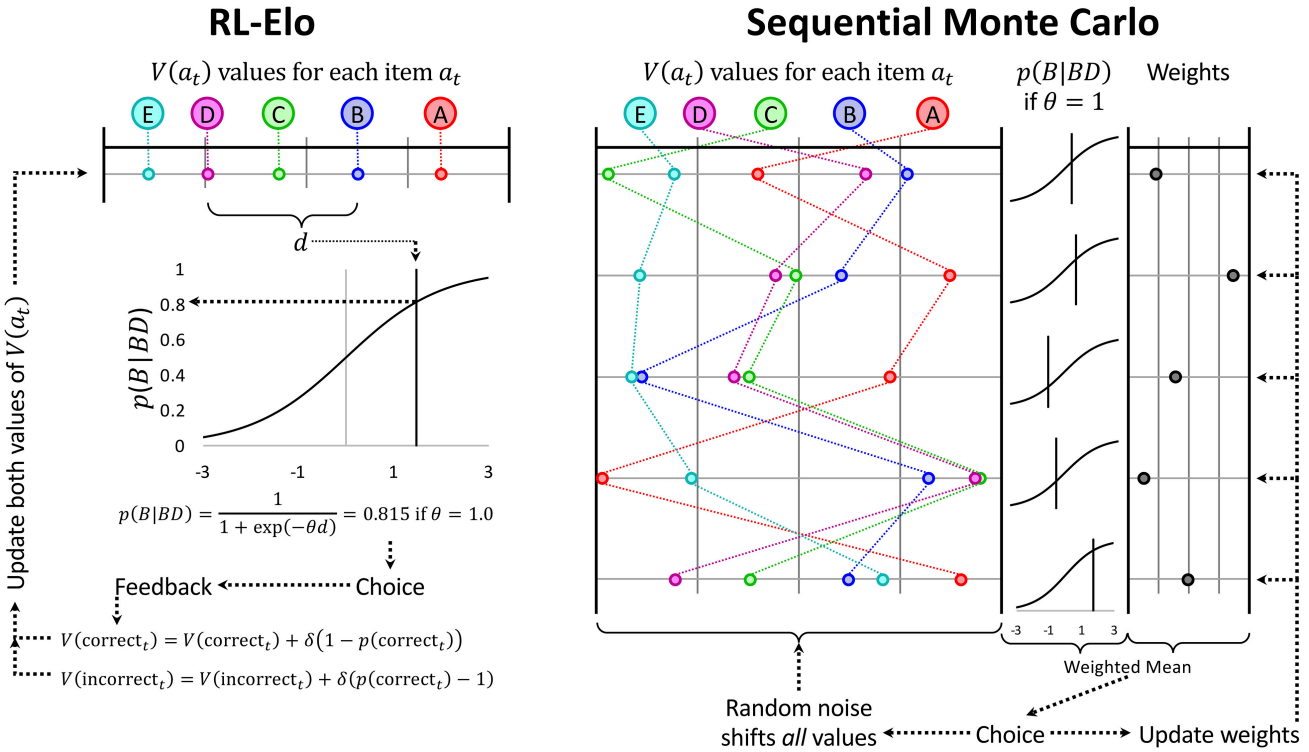


FIGURE 4 | Schematic depiction of the updating rules used for the RL-Elo and Sequential Monte Carlo (SMC) algorithms. **(Left)** RL-Elo represents each list item a_t as a stimulus as having a value $V(a_t)$ along some unitless continuum. The probability of choosing the item is obtained by computing the difference d between that item and its alternative (d can be negative if the item has the lower of the two values) and passing d through the logistic function. The parameter θ adjusts the contrast. If the choice made was correct, its value is adjusted as a function of how far that value was from its expectation, adjusted by a learning rate δ. **(Right)** The updating process during a single trial for SMC is demonstrated using a sample of five hypothetical orderings (out of the 10,000 that our implementation used). As with RL-Elo, each difference score yields a probability of selection using the logistic function. Additionally, each hypothetical ordering has an associated weight based on how well it has predicted previous trials. The choice is made on the basis of a weighted average of the probabilities, and the feedback is used to update the weights (strengthening or weakening each ordering's influence as a function of how well it predicted the outcome of the trial). The orderings themselves are then modified using Gaussian noise. Not shown is the algorithm periodically throwing out the existing 10,000 orderings in favor of a new set of alternatives. Reproduced with permission (DOI: 10.6084/m9.figshare.7992014.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

orderings adaptively based on feedback, the values of all items in all hypothetical orderings are adjusted at random by adding normally distributed random noise (with variance σ^2) to each value. In general, $\sigma_\emptyset$ is large relative to σ, in order to ensure that the various hypothetical orderings differ considerably from one another. SMC does not use the feedback to update the orderings directly, but instead uses it to update the *weight* associated with each hypothetical ordering. All orderings begin with equal weights, but over time, those whose predictions are consistent with feedback are given more weight, while those that are not consistent are given less weight. Orderings that happen by chance to be mostly correct come to dominate the overall representation of order, while those that make consistently poor predictions are reduced to having almost no influence. The choice that the actor then makes is based on the weighted average of the estimated stimulus positions, based on a softmax decision rule (Equation 3) with θ as a free parameter. This implementation thus has three free parameters: $\sigma_\emptyset$, σ, and θ .

SMC can be seen as 10,000 RL-Elo estimates, except that all stimuli are being updated at random all the time. Those orderings that are consistently right are given greater weight than those that are consistently wrong, and it is the *weight* that is updated on the basis of feedback, using the delta rule given in Equation 6. Rather than trying to update a single ordering optimally (as RL-Elo does), SMC updates a crowd of possible orderings randomly, and then updates the "fitness" or "reputation" of those random orderings on the basis of feedback to separate out the good vs. the bad models. **Figure 4** (right) gives a schematic example of this updating process (albeit using only five hypothetical orderings, rather than 10,000).

Even with 10,000 simultaneous models, however, there is a risk that *none* of the hypothetical orderings are ordered correctly. We know that monkeys can learn 9-item lists (Jensen et al., 2013) and 15-item lists (Treichler et al., 2003), but for lists of those lengths there are 9! and 15! (i.e., 362,880 and 1.3 trillion) possible orderings, respectively. As such, SMC includes a criterion to resample its list orderings periodically, generating a new set of 10,000 hypotheses using its existing 10,000 hypotheses and their respective weights as a seed. Put another way, the new set of 10,000 hypotheses are drawn, with replacement, from

the previous set of 10,000, using the learned weights. Thus, orderings that have close to zero weight have little chance of being resampled, whereas those with high weight are likely to appear multiple times in the new generation of hypotheses (after which they diverge due to the added noise).

SMC algorithms give reasonable approximations to behavior in a variety of contexts (Yi et al., 2009). Insofar as they constitute a Monte Carlo estimation process, filters with a sufficiently large number of particles should converge on optimal estimates given sufficient evidence. SMC's periodic resampling allows the algorithm to weed out bad hypotheses and gradually converge on a high-precision estimate of the list ordering, giving it an advantage over a conventional Kalman filter. However, this advantage relies on the assumption that the target ordering is static. In the event that the ordering changes, this resampling can result in the algorithm painting itself into a corner, from which it can only gradually escape. Consequently, a reasonable prediction is that SMC should be robust to many experimental manipulations, but should respond poorly to experimental designs that change the ordering of the list.

Betasort: Modeling Position, Uncertainty, and Transitivity

Although score-differential models of rank like RL-Elo and SMC have the potential to be effective in sorting well-ordered lists, they have a number of limitations. The most substantial of these is that they assume that the uncertainty of each score is uniform. If, for example, a subject is trained extensively on a five-item list ACDEF, and is only later introduced to the stimulus B, those models have no way of representing that the position of B is much more uncertain than the positions of the other stimuli. To fix this, the position of each stimulus would need to be represented by a minimum of two parameters. Attempts to extend the Elo rating system directly, such as Glickman's "Glicko" and "Glicko-2" algorithms (described by Samothrakis et al., 2016) provide a framework for introducing additional parameters, but their additional model complexity can create scenarios in which ratings become trapped in local minima, as well as relying on hyperparameters that must be set *ad hoc* to govern how dispersion evolves over time. It would be preferable to keep action selection and value updating be as simple and as computationally efficient as possible.

Jensen et al. (2015) proposed the "Betasort" algorithm as a computational model of serial learning. Under this framework, the position of each item in the list is represented by a beta distribution whose shape is determined by two values, U and L (corresponding to the "upper" and "lower" ends of a unit scale). An item's probability density $D(x)$ between 0.0 and 1.0 is given by the following formula:

$$\begin{aligned} D(x) &= \text{BetaPDF}(x|U, L) \\ &= \frac{x^{U-0.5} \cdot (1-x)^{L-0.5}\, \Gamma(U+L+1)}{\Gamma(U+0.5) \cdot \Gamma(L+0.5)} \end{aligned} \tag{7}$$

Note that $\text{BetaPDF}(x|U, L)$ assumes the inclusion of a Jeffrey's prior, which introduces a slightly conservative factor to the estimate that facilitates numerical estimation. This is accomplished by adding a value of 1/2 to each term, on the principle that the beta distribution is its own conjugate prior, such that $\text{BetaPDF}(x|U+0.5, L+0.5) \propto \text{BetaPDF}(x|U, L)\,\text{BetaPDF}(x|0.5, 0.5)$. Although it would be most appropriate to say that the entire distribution $D(x)$ represents the estimate of an item's position, a summary is given by the mean, $\frac{U+0.5}{U+L+1}$.

The action policy under Betasort is to draw a random value from each distribution that corresponds to an available action, and to take the action with the largest value. The net effect of a random draw procedure is a baseline probability of choosing an action a_t given by the following formula (Raineri et al., 2014):

$$\begin{aligned} p_{\text{beta}}(a_t) = &\int_0^1 \text{BetaPDF}\left(x, U_{a_t}, L_{a_t}\right) \\ &\cdot \left[\int_0^x \textit{BetaPDF}\left(y, U_{\neg a_t}, L_{\neg a_t}\right) dy\right] dx \end{aligned} \tag{8}$$

Although determining the value of this double integral is computationally prohibitive, doing so is not necessary to simulate behavior using this procedure. This is because making random draws from two beta distributions is much less computationally intensive than solving the integral above.

Unlike RL-Elo and SMC, Betasort applies its value updating function to all stimuli (including those that are not presently visible) and treats positive feedback differently from negative feedback. **Figure 5** gives a visual sense of these updating steps. Following every trial, the existing values of U and L are relaxed by a "recall" parameter φ, whose value is less than 1.0. This effectively makes U and L leaky accumulators (Grice, 1972), although in Betasort's case, the strength of the forgetting imposed by φ is further adjusted as a function of how frequently the algorithm makes correct responses. See Jensen et al. (2015) or the **Supplementary Materials** for details.

If a choice yielded positive feedback, every value of U is increased by $\frac{U}{U+L}$ and every value of L is increased by $\frac{L}{U+L}$. This has the effect of increasing the model's confidence of all position estimates, because it reduces the variability of all position distributions without shifting any of the distribution means. If, on the other hand, the feedback was negative, several updates happen in succession. First, the chosen item has its L increased by 1 while the item that was not chosen has its U increased by 1. Next, all of the stimuli that were *not* chosen have their values updated as well. Items above the visible stimulus pair have their values of U increased by 1, items below the pair have their values of L increased by 1, and those falling in between have their positions consolidated as though the feedback had been positive. This "implicit updating" stage ensures that, in general, the full list will preserve its ordering. As demonstrated below, implicit updating is needed to explain why experimental designs like the massed trials studies by Lazareva and Wasserman (2006, 2012) do not yield preference reversals.

One of the commonly reported features of TI studies, particularly those involving non-human animals, is a background error rate that persists even after extensive training. Whether

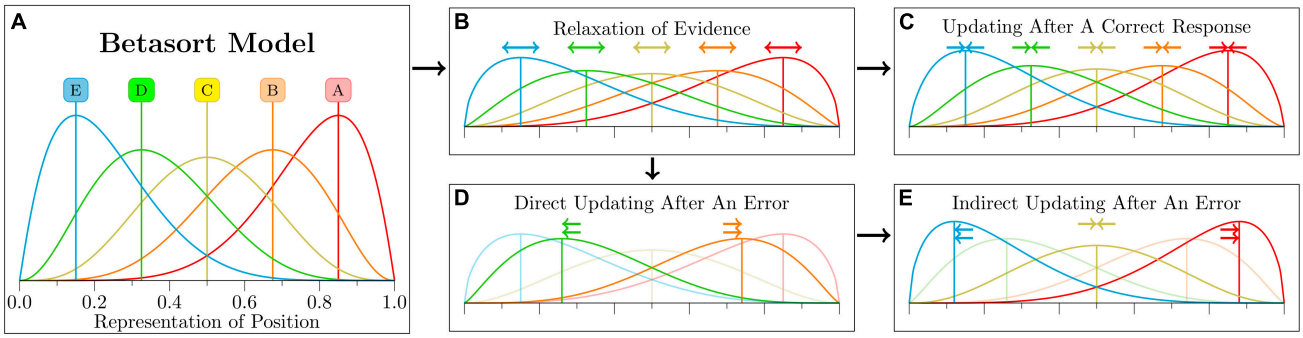


FIGURE 5 | Schematic depiction of the Betasort algorithm's representation of position. Quantitative details about each step are described by Jensen et al. (2015). **(A)** Each stimulus has an associated beta distribution, whose parameters encode both the stimulus position and its uncertainty. In order to choose between available options, the algorithm draws a value for each possibility, then chooses the item whose random draw yielded the highest value. Additionally, there is a probability ρ of the subject ignoring these values and making a response at random. **(B)** Following a choice, the representation is updated. The first stage of updating is "relaxation," which increases the uncertainty of the estimated position without changing the means of any of the distributions. This applies to all items, regardless of whether the stimuli were present during that trial or not. **(C)** If the subject makes a correct choice, then the distributions for all items are consolidated (again, whether or not any given stimulus was presented). What subject make consistently correct responses, the relaxation and consolidation steps effectively balance one another out. **(D)** If, however, the subject makes an incorrect choice, then updating proceeds in two stages. During direct updating, the means of the stimuli presented during that trial are moved apart to reduce the odds of another erroneous response. **(E)** The preceding step is them immediately followed by incorrect updating of all stimuli that were *not* visible during the trial. Stimuli that were between the two items are consolidated, items whose current mean lies below that of the pair are shifted further downward, and pairs whose mean lies above the pair are shifted upward. Reproduced with permission (DOI: 10.6084/m9.figshare.7992017.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

mistakes are made due to memory errors or to shifting attention, they are not captured by models that converge on perfect discrimination. Consequently, Betasort introduces a "random response" parameter ρ, which is the probability that the subject, on a given trial, ignores its current knowledge and instead makes a response at random. This could also be considered a "lapse rate," governing the frequency with which the subject lapses into making an uninformed response. When taking this additional factor into consideration, the probability of choosing a_t given by Equation 7 is updated in the following way:

$$p_{\text{beta}}(a_t, \rho) = \frac{\rho}{2} + (1 - \rho) \cdot p_{\text{beta}}(a_t) \tag{9}$$

The ρ parameter effectively puts a ceiling on performance and scales the remaining probabilities toward chance.

SIMULATION OF SERIAL ORDER TASKS

The algorithms described in the previous section are process models, and their behavior evolves over the course of training. One of the consistent themes of using process models to describe natural processes is that their results are often contrary to our intuitions. As such, the evaluation of these algorithms is best performed by simulating the consequences of various experimental designs and observing how they perform, rather than merely thinking about them intuitively. We subjected the algorithms to various experimental manipulations to demonstrate their relative efficacy and performance idiosyncrasies. In all cases, the figures below represent mean accuracies over 1000 simulations, each of which had its own randomized trial order of stimulus presentations.

Parameter Selection

The algorithms each had 2–4 free parameters, which afforded varying degrees of flexibility in describing performance. Our objective was not to discover each algorithm's optimal performance, but instead to use parameters that best described a real-data target. Put another way, rather than describe each algorithm's normative performance, we tested how well each approximates an empirical estimate of TI performance in an actual organism. In particular, we were interested in those parameters that recapitulate "performance at transfer" in a classic TI design. With this in mind, we selected one monkey (named "Hubble") whose TI data was published by Jensen et al. (2015). Likelihoods were calculated for the six algorithms with respect to the sequence of trials and the subject's choices during the first 21 trials of all-pairs testing, immediately after the end of training. This ensured that, however, performance appeared early in training or late in testing, it resembled the critical transfer test as closely as possible.

In order to perform statistical inference using process models, the response likelihoods must also be calculated on a trial-by-trial basis, unfolding as each models' representation is updated. Due to random processes that vary from one simulation to the next (particularly for SMC), calculations of the likelihoods were noisy. Additionally, the 3- and 4-parameter algorithms were not well suited to grid approximation. Consequently, maximum-likelihood parameters were obtained by quadratic approximation. As a shorthand, we will subsequently describe these as each algorithm's "Hubble parameters."

According to this procedure, Q-learning's Hubble parameters were $\delta = 0.138$, $\theta = 2.280$, and $\gamma = 0.373$. As a shorthand, error prediction learning rates like δ can be thought of in terms of their inverse. That is, if $\delta = 0.138$, then the actor will base its estimates

on about the previous $\frac{1}{\delta} = 7.2$ updates of that value. This is a relatively rapid learning rate, focused only on recent events. The decision rule $\theta = 2.280$ corresponds to an exaggerated preference for whichever alternative has a large value of Q (i.e., "winner-takes-all"). For example, if $Q_A = 1$ and $Q_B = 0.5$, then chance of choosing A over B would only be 0.622 if $\theta = 1$, whereas the probability for A would be 0.758 if $\theta = 2.280$. A prospective discounting rate of $\gamma = 0.373$ would normally suggest an actor that values future states at about $^1/_3$ of the value of current choices. However, since current choices in the TI task exert no control over future states, this term effectively introduces a small amount of noise into performance, shifting all behavior slightly closer to chance.

The Hubble parameters for VTM were $\delta = 0.137$, $\theta = 2.260$, and $\tau = 0.406$. the learning rate and decision rule terms were similar to those obtained for Q-learning, namely relatively fast learning and an exaggerated preference for the stimulus with the maximum Q estimate. The transfer rate τ was relatively large. For each unit of feedback that was learned about the value of a stimulus, 0.406 units were also learned from the alternative with which it was paired.

The Hubble parameters for RL-REMERGE were $\lambda = 2.690$, $\omega = 15$, $\theta = 4.971$, and $\delta = 0.500$. Because λ constituted the "inverse temperature," higher values correspond to less volatility while the network approached equilibrium. The very large value of ω ensured that information flowing through the network washed out after only a few steps through the conjunctive layer. This is certainly bad for the actor's performance, and was likely the most efficient way for REMERGE to approximate the persistent error rates observed in subjects after extensive training. The decision rule θ was quite large, indicating a strong preference for the Response node with maximal activity. The learning rate δ was very high relative to the other model-free algorithms. Although this would ordinarily make behavior highly volatile, the fact that the Conjunctive layer had a node for every observed stimulus pair meant that the algorithm was at an advantage because it could memorize pairs on a case by case basis. A high value of δ enabled the network to pick out and keep those contextually correct responses, even given only a few exposures.

The Hubble parameters for RL-Elo were $\delta = 0.079$ and $\theta = 1.529$. Although these values may seem smaller than those found for Q-learning and VTM, it is important to remember that the model-free algorithms are comparing rewards (1 in most cases) to non-rewards (0), whereas Elo is comparing wins (1) with losses (-1). Since the units for RL-Elo effectively cover twice the numerical range of the model-free algorithms, its units need only be half as big to have a comparable effect. Viewed from this perspective, RL-Elo's parameters are quite similar to those of the model-free algorithms.

The Hubble parameters for SMC were $\sigma_\emptyset = 0.5$, $\sigma = 0.032$, and $\theta = 1.410$. The $\sigma_\emptyset$ parameter was used to initialize the 10,000 candidate orderings, and its relatively small value ensured that all of the candidates were relatively confusable with one another. Meanwhile, the even lower value of σ ensured that the amount by which these representations changed from one trial to the next was small. The net effect was that SMC's prodigious ability to solve ordinal sorting problems was dampened, allowing it to better approximate Hubble's gradual learning and tendency to make errors. The decision rule θ was very similar to that of RL-Elo, which is sensible because the two made similar use of a logistic discrimination function.

The Hubble parameters for Betasort were $\rho = 0.188$ and $\varphi = 0.495$. The randomness parameter ρ means that, for about every two trials out of eleven, Hubble was expected to give a random response regardless of the stimuli, effectively putting a ceiling on his performance. Meanwhile, the recall parameter φ had a very low value, suggesting that the representation lost roughly half of its evidence for each trial. This had the effect of keeping all of the stimulus positions uncertain, which increased the error rate but also increased how rapidly positions could be adjusted on the basis of new information.

Classical Transitive Inference

The classic procedure for demonstrating TI is to set up an ordered list (e.g., ABCDEFG) and train subjects by presenting only the adjacent pairs of items (e.g., AB, BC, CD, DE, EF, and FG), giving positive or negative feedback for single responses. After training, the actor is tested with all possible pairs of items (21 in this case). Both training and testing consist of "blocks," during which each stimulus appears a fixed number of times. Since it is standard to counterbalance the spatial arrangements on a screen (AB vs. BA), we presented the actor with each pair twice during a block. Thus, training blocks consisted of 12 trials (each pair appearing twice), with the order of trials in any given block randomly permuted. This guaranteed that the actor's choice a_t had no influence on their state s_{t+1}. Training consisted of 11 blocks, and thus of 132 trials.

During testing, actors were exposed to all 21 possible item pairs (AB, AC, AD, etc.). As in training, each pair appeared twice per block, so testing blocks consisted of 42 trials, randomly permuted. Testing consisted of five blocks, and thus of 210 trials. The testing set includes six "critical" pairs (BD, BE, BF, CE, CF, and DF) which (1) were not presented during training and (2) do not include the terminal items A or G. In this and all subsequent simulations, rewards were delivered for correct responses during the testing phase, consistent with the data collected from Hubble.

Figure 6 (left column) depicts the mean predicted response accuracy by model-free algorithms for all pairs composed of non-terminal items (e.g., BC and BD, but not AB or EG), grouped by symbolic distance. Trial "zero" in this case consists of the state of knowledge after the last trial of training but before the first trial of testing, and thus represents estimated accuracy at transfer. Dashed lines correspond to asymmetric updating (in the tradition of behavior analysis), which generally display slower learning rates because asymmetric updating results in less evidence extracted from each trial. **Figure 6** (middle column) displays the response accuracy for each of the 21 pairs at trial zero, with the critical pairs shaded in gray. Hubble's estimated response accuracy for the 21 pairs is also plotted allowing comparison between the algorithms and the subject they are trying to approximate. Dark points correspond to symmetric updating, and light points correspond to asymmetric updating.

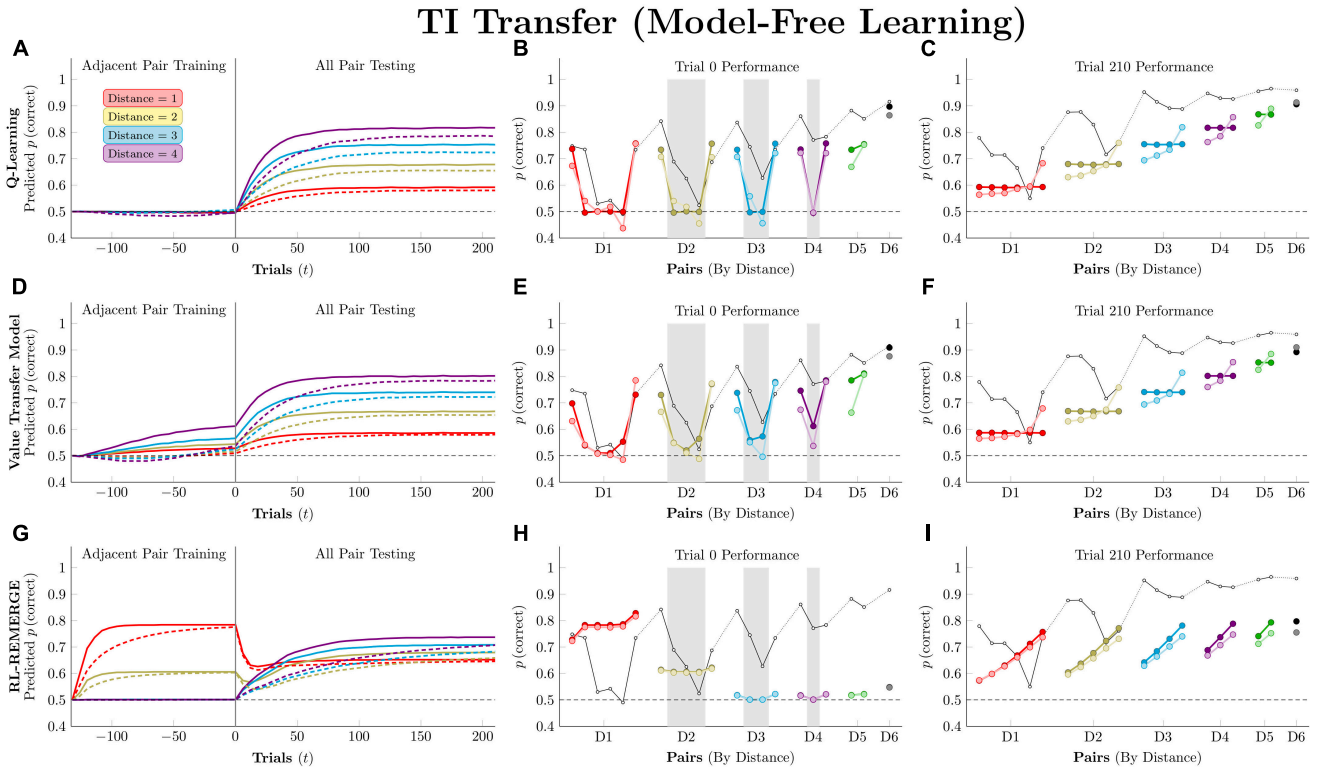


FIGURE 6 | Performance of model-free algorithms on a 7-item TI transfer task (adjacent pair training, all pair testing). **(A)** Mean performance of *Q*-learning on non-terminal pairs throughout training and testing, sorted by symbolic distance. Solid lines denote learning with symmetric updating of both alternatives, whereas dashed lines denote asymmetric updating of values associated with the chosen stimulus. **(B)** Mean performance of *Q*-learning for each of the 21 pairs at transfer, sorted by symbolic distance. Open points are empirical estimates of performance by Hubble, a monkey reported by Jensen et al. (2015). Dark points correspond to symmetric updating, while light points denote asymmetric updating. Critical pairs are shaded in gray. **(C)** Mean performance of *Q*-learning for each of the 21 pairs at the end of training. Open points are empirical estimates of performance by Hubble. Dark points correspond to symmetric updating, while light points denote asymmetric updating. **(D–F)** As above, but reporting simulated performance of the Value Transfer Model (VTM). **(G–I)** As above, but reporting the simulated performance of RL-REMERGE. Reproduced with permission (DOI: 10.6084/m9.figshare.7992020.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

Overall, this simulation confirms previous reports (Jensen et al., 2015, 2018) that *Q*-learning shows no TI at the start of testing and no learning on non-terminal pairs throughout training. VTM performance yielded a small transfer effect (with the critical non-adjacent pairs above chance), a positive symbolic distance effect, and a substantial terminal item effect when run using its Hubble parameters. Although the distance effect is real, it is much smaller than that displayed by Hubble. On this basis, it is fair to say that VTM can "solve" this TI task (insofar as accuracy on critical pairs exceeds chance), particularly when using symmetric updating, but still underperforms on pairs with large symbolic distances.

Superficially, RL-REMERGE also "solved" the TI task, insofar as some of its critical pairs were above chance at test. However, it displayed a negative symbolic distance effect and no clear terminal item effect. For pairs with a symbolic distance greater than 2, performance was close to chance. This is due to the high value of ω used in the simulations. Herein lies the essential tension in using REMERGE as an explanatory model: If ω is small, performance of proximate pairs (which are more numerous) will be much too high, whereas if ω is large, distant pairs of items will be close to chance. Since REMERGE is expected to display a negative distance effect at test in all cases, it cannot come to resemble empirical results in their particulars.

Figure 6 (right column) displays response accuracy at the end of training (on trial 210). Here, *Q*-learning and VTM appear nearly indistinguishable, both showing positive distance effects. RL-REMERGE provides an interesting contrast. Despite receiving symmetric feedback for item pairs, the network displays an effect of rank.

Figure 7 plots the simulation results for the three model-based algorithms, all of which show evidence of TI at the start of testing. Of these, RL-Elo appears to provide the closest approximation to the qualitative features of the data (showing both a positive distance effect and a terminal item effect), as well as a close approximation to the performance at trial 210. However, its performance at transfer is too low for many of the pairs. Meanwhile, SMC also displays both distance and terminal item effects, but its performance is too high in all cases. Finally, Betasort provides a good approximation to the *overall* response accuracy at test (including distance effects), but does not give a clear terminal item effect. Furthermore, its terminal item effect is negative at trial 210, a result not seen in empirical TI results.

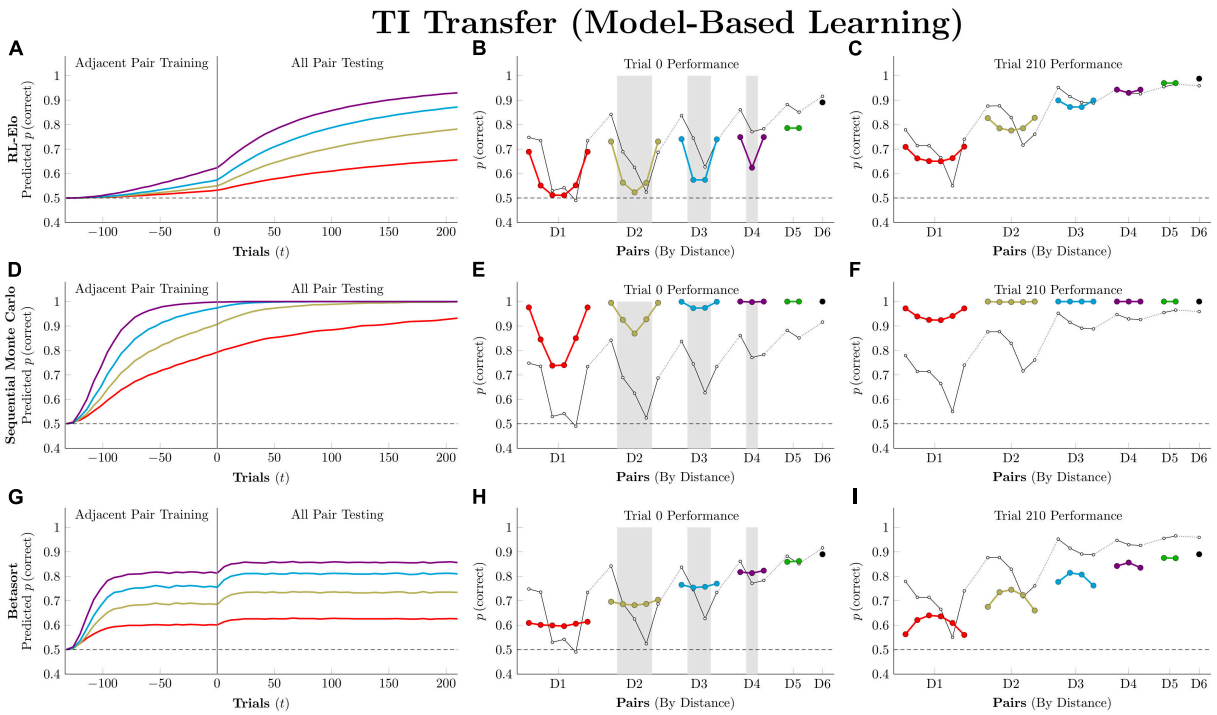


FIGURE 7 | Performance of model-based algorithms on a 7-item TI transfer task (adjacent pair training, all pair testing). **(A)** Mean performance of RL-Elo on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of RL-Elo for each of the 21 pairs at transfer, sorted by symbolic distance. Open points are empirical estimates of performance by Hubble, a monkey reported by Jensen et al. (2015). Critical pairs are shaded in gray. **(C)** Mean performance of RL-Elo for each of the 21 pairs at the end of training. Open points are empirical estimates of performance by Hubble. **(D–F)** As above, but reporting simulated performance of the Sequential Monte Carlo (SMC) algorithm. **(G–I)** As above, but reporting the simulated performance of Betasort. Reproduced with permission (DOI: 10.6084/m9.figshare.7992023.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

Distorting the Expected Reward Gradient

Model-free learning relies on comparisons of expected value as a proxy for preference. If different stimuli are associated with the delivery of different amounts of reward (i.e., to have different "reward magnitudes"), model-free judgments of preference should be badly disrupted if the reward information is not concordant with the optimal preference ordering. The worst-case scenario is a "reversed reward gradient," in which an item's rank also corresponds to its quantity of reward. So, stimulus A is always correct because its rank is 1, but is only worth 1 unit of reward. By contrast, F has a rank of 6, and as such is worth 6 units of reward in the pair FG, but is incorrect (and thus worth nothing) in the pairs AF, BF, CF, DF, and EF. Jensen et al. (2018) report that monkeys are able to solve TI tasks in spite of a reversed reward gradient, while *Q*-learning was unable to solve the problem in all cases regardless of the parameters that were used.

Figure 8 depicts performance of the model-free algorithms in an experimental procedure that is identical to that described in the preceding sections (7-item list, 11 blocks of adjacent-pair training, 5 blocks of all-pair testing, each pair presented twice per block). Because it would not be reasonable for actors to know the amount of reward associated with each stimulus in advance, these simulations were performed using asymmetric updating of only the chosen item.

For both *Q*-learning and VTM, response accuracy on the critical pairs is badly disrupted, but the resulting disruption takes on a surprising form: A sawtooth of alternating high-value and low-value estimates for pairs at each symbolic distance. This pattern is a consequence of the asymmetric updating during training. The value of F is driven very high by its hefty 6-unit rewards, and this causes it to be strongly preferred over E. As a consequence, D in the pair DE is also preferred. If all rewards were equal, then feedback during EF trials would cancel this effect out, but F's larger rewards tend to overshadow the smaller values earned for selecting E (which happens less and less often as F grows in value). The overall effect is near-exclusive selection of F and an alternating pattern of preference propagating down the list, gradually becoming washed out by the time AB is being considered. Once testing begins and all pairs start being presented, this effect rapidly washes out and subjects instead come to favor later list items in general. Because G never accrues any worth, pairs that include it remain above chance, but the remaining stimuli are consistently distorted by their reward values, causing many pairs to be selected at below-chance rates throughout testing.

RL-REMERGE displays a strikingly different pattern that is much more robust against this manipulation. Although some pairs are pushed slightly below chance during training, the algorithm is broadly resistant to the intervention, maintaining

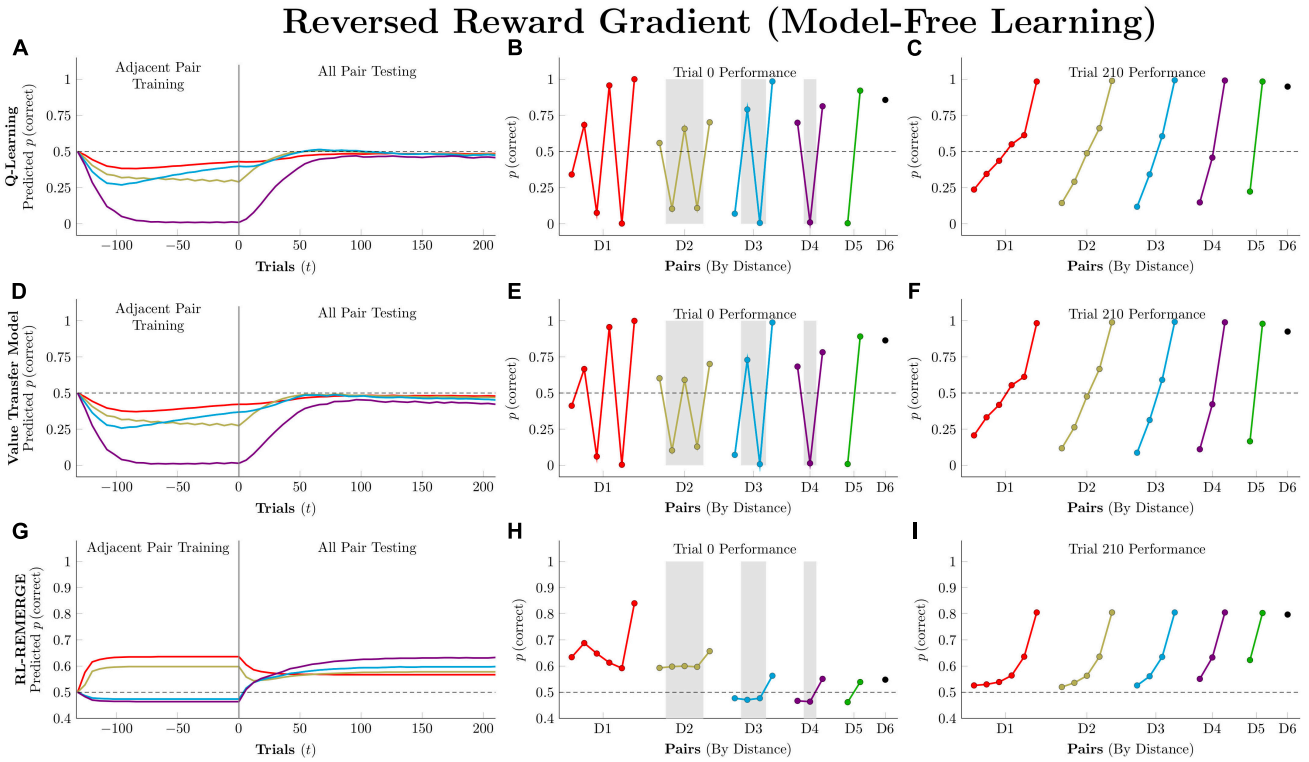


FIGURE 8 | Performance of model-free algorithms on a 7-item TI transfer task (adjacent pair training, all pair testing) in which the amount of reward delivered corresponded to the rank of the correct response (i.e., 1 unit for A, 2 units for B, and so forth). **(A)** Mean performance of *Q*-learning on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of *Q*-learning for each of the 21 pairs at transfer, sorted by symbolic distance. Critical pairs are shaded in gray. **(C)** Mean performance of *Q*-learning for each of the 21 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Value Transfer Model (VTM). **(G–I)** As above, but reporting the simulated performance of RL-REMERGE. Reproduced with permission (DOI: 10.6084/m9.figshare.7992026.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

response accuracies above chance for all pairs by the end of testing. This is likely because the configural character of the REMERGE network prioritizes contextual information ("When BC, choose B"), and only reflects global information through its network propagation (which rapidly washes out when ω is large).

Model-based algorithms were not run in this condition because they do not adjust their behavior on the basis of reward magnitude, only outcome (correct or incorrect). As a consequence, their predicted behavior would be exactly the same as that described in **Figure 7**.

Model-free learning algorithms fail this task dramatically. Not only does overall learning remain flat during training, performance on 11 of the 21 pairs remains below chance even after 210 trials of testing. Because reward does not provide a proxy for stimulus rank in this procedure, an organism that is able to solve this problem cannot be operating on the basis of model-free learning alone (Jensen et al., 2018).

Massed Presentation of a Single Stimulus Pair

Studies of TI almost universally present the various stimulus pairings with equal frequencies, and this procedural assumption underlies many results claiming support for particular models (e.g., Wynne, 1995). However, such models can diverge from the behavior of organisms when a particular pair is presented *en masse* (Lazareva and Wasserman, 2006, 2012; Jensen et al., 2017). These "massed presentation" designs constitute a vital stress-test for models, because feedback in natural environments is almost never counterbalanced.

Actors in this simulation were first trained on 132 trials of adjacent pairs, as described previously. Then, they were trained for an additional 132 trials on only a single pair: FG. After both phases of training, actors completed 210 trials (5 blocks) of all-pairs testing.

Figure 9 depicts the consequences of massed presentation of FG for model-free algorithms. Both *Q*-learning and VTM have their performance disrupted by this manipulation, manifesting in reduced accuracy on other pairs that include F, but otherwise leaving their representations intact. VTM's disruption was smaller than that of *Q*-learning, because F's association with the valueless stimulus G dampens its expected value, even though F was correct for 132 consecutive trials. RL-REMERGE, however, was completely unaffected by massed trials, since the extended training on FG had no impact on the excitatory/inhibitory connections of other pairs in the Conjunctive layer.

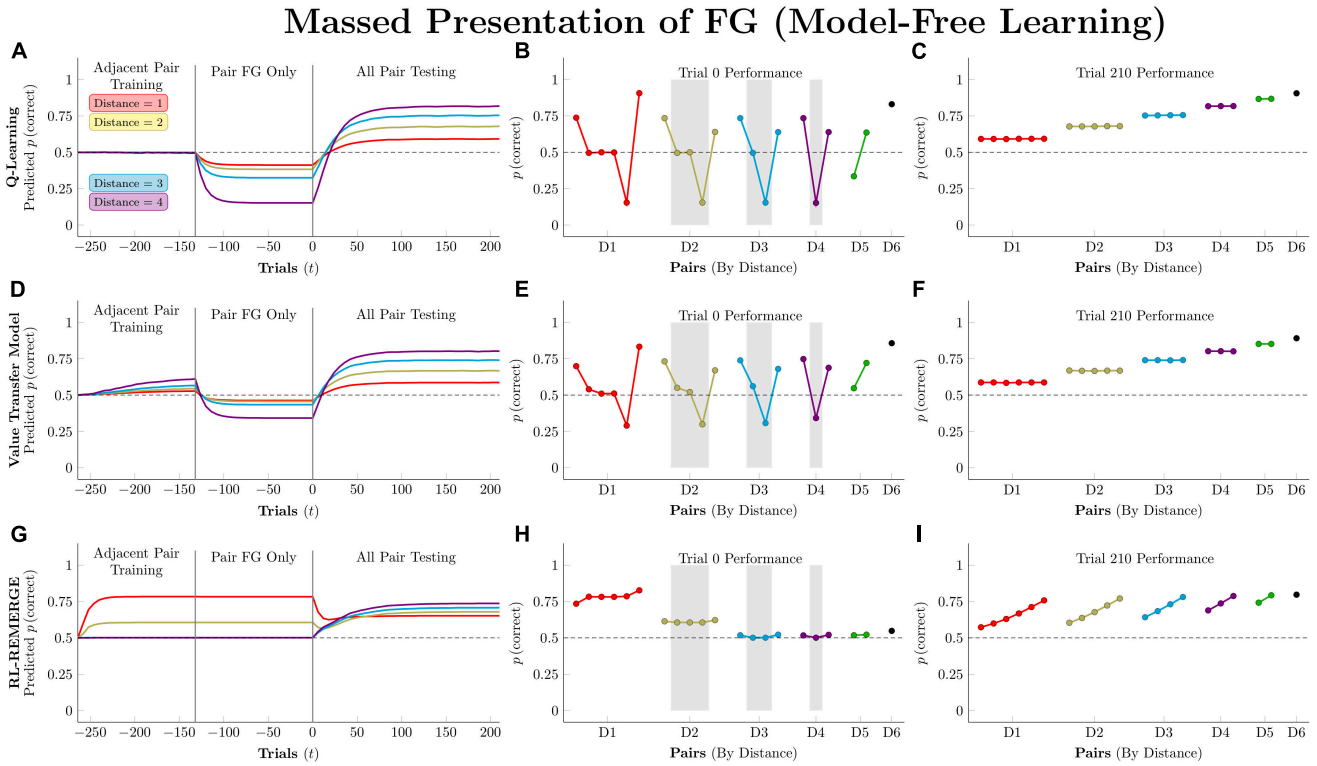


FIGURE 9 | Performance of model-free algorithms on a 7-item TI transfer task (adjacent pair training, all pair testing) with an intervening period during which only the pair FG was presented. **(A)** Mean performance of *Q*-learning on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of *Q*-learning for each of the 21 pairs at transfer, sorted by symbolic distance. Critical pairs are shaded in gray. **(C)** Mean performance of *Q*-learning for each of the 21 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Value Transfer Model (VTM). **(G–I)** As above, but reporting the simulated performance of RL-REMERGE. Reproduced with permission (DOI: 10.6084/m9.figshare.7992029.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

Figure 10 depicts model-based performance for this manipulation. Both RL-Elo and SMC have their performance disrupted in a manner almost identical to the disruption experienced by *Q*-learning and VTM, and for the same reason: all four algorithms only apply informative feedback to the items present in the current trial. SMC's resistance to this disruption is mainly due to its high level of accuracy overall.

Betasort's displayed a clear distance effect by the end of the first phase of training, but all pairs (other than those that include G) drop to chance levels during FG training. However, within 20 trials of all-pairs testing, the original order sprang back into place. This effect is due to the combination of three factors: The relatively low value of φ (which causes U and L to shrink rapidly, without ever equaling zero), the algorithm's implicit updating (which preserves items in a particular order by updating all positions, even when U and L are small), and the Jeffrey's prior that adds 0.5 to the values of U and L in Equation 6. In effect, Betasort has preserved the *relative* order through FG training, but has allowed the *absolute* distance between items to become small as erroneous responses to G gradually push the values of $\frac{U}{U+L}$ for all other stimuli up to nearly 1.0. As items are bunched up against the same axis, the influence of the prior grows stronger, tending performance toward chance levels. However, as soon as pairs other than FG are introduced, the stimuli push themselves back apart (all while preserving their relative order) and the prior's influence returns to background levels.

Reversal of Stimulus Order at Test

We ran simulations in which algorithms were first trained on all 21 stimulus pairs for 210 trials (5 blocks, counterbalanced for position). Then, the stimulus order was reversed, becoming GFEDCBA. This new ordering was presented for another 210 trials.

Figure 11 depicts performance by the model-free algorithms. Because all pairs were being presented, all three successfully determined the ordering. Because the actors have been given no indication of the reversal at trial zero, their performance is a mirror image of their training performance, consistently below chance. By the end of testing, all three have recovered the overall pattern of accuracy seen during training, but they differ in terms of the number of trials needed to do so. *Q*-learning was the slowest of the three, with the order flipped following trial 27, on average. RL-REMERGE was only slightly faster, reversing its representation of the order after 26 trials. By this test, VTM was the nimblest, successfully reversing its order after 17 trials.

Figure 12 plots the effects of order reversal for the three model-based algorithms. Here, a much more dramatic

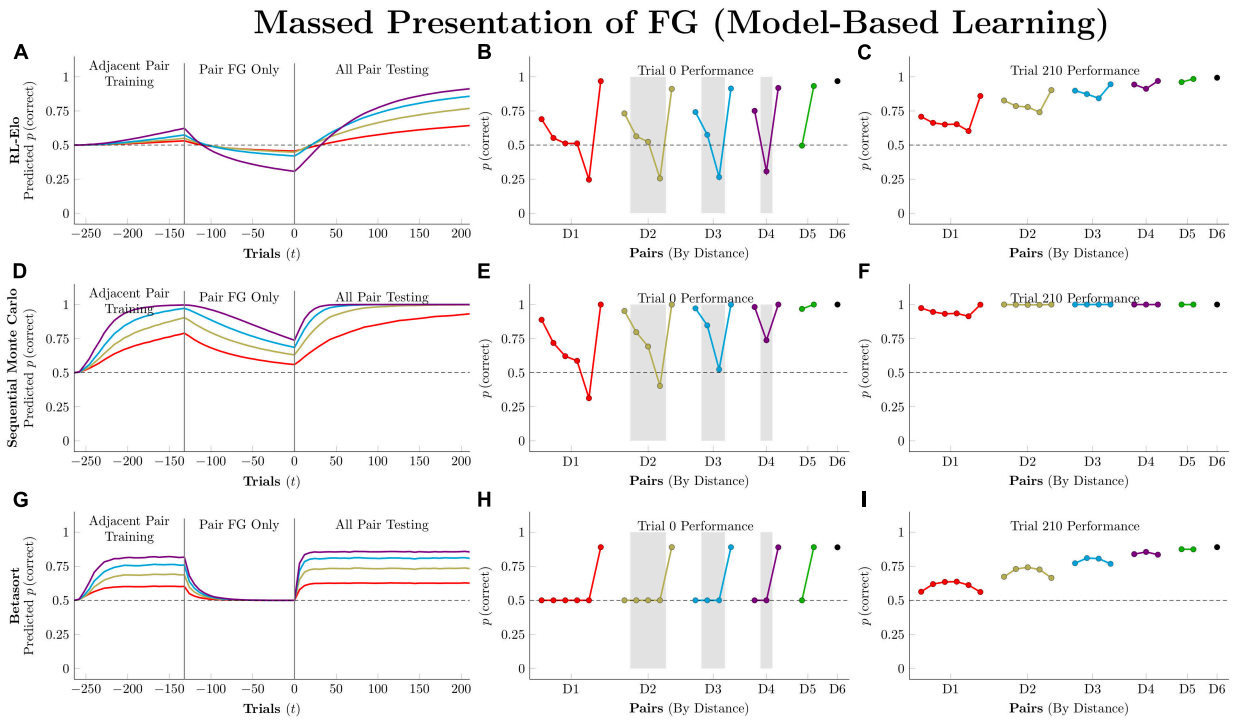


FIGURE 10 | Performance of model-based algorithms on a 7-item TI transfer task (adjacent pair training, all pair testing) with an intervening period during which only the pair FG was presented. **(A)** Mean performance of RL-Elo on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of RL-Elo for each of the 21 pairs at transfer, sorted by symbolic distance. Critical pairs are shaded in gray. **(C)** Mean performance of RL-Elo for each of the 21 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Sequential Monte Carlo (SMC) algorithm. **(G–I)** As above, but reporting the simulated performance of Betasort. Reproduced with permission (DOI: 10.6084/m9.figshare.7992032.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

disruption is seen. Although RL-Elo recovers its original level of performance, it takes an average of 74 trials to return the critical pairs to their correct order. SMC is even more disrupted – its near-ceiling performance translated to near-floor performance at the start of testing, and 210 trials was not sufficient training to recover its baseline performance, only managing to order the critical pairs correctly after 114 trials. In both cases, the culprit is a parameter (δ in the case of RL-Elo and σ in the case of SMC) that can only make small adjustments to their representations.

Contrastingly, the Betasort algorithm reorders the pairs correctly after just 9 trials, and recovers its original ceiling levels of accuracy by the end of the first block of testing. This owes to the low value of φ, which rapidly discards representation information that produces consistently incorrect responses.

Linking Lists by Training a Single Stimulus Pair

One of the most dramatic series of results in the serial learning literature are the "list linking" studies reported by Treichler and colleagues. Monkeys learned the order of multiple separate lists using only adjacent pairs (e.g., ABCDE and FGHIJ). When monkeys were subsequently trained that E > F (in effect "linking" the end of one list to the start of the next), their performance at test was consistent with an inference that the two sublists should be combined into a single 10-item list, ABCDEFGHIJ (Treichler and Van Tilburg, 1996). List linking effects of this kind have since proved to be surprisingly robust over lists of fairly dramatic lengths, reaching well into the double digits (Treichler et al., 2007). However, if monkeys were not given training on the linking pair EF, their performance suggested that they assumed that item ranks were transferable to new pairings (Treichler and Raghanti, 2010). For example, after learning the lists ABCDE and FGHIJ, a monkey would be expected to favor G over D (because the former had a rank of 2 and the latter had a rank of 4) *unless* they were also trained that E > F, in which case the preference would reverse (because all items in the first list are implied by the EF pairing to have lower rank than any items in the second list).

The principle that rank should be transferable across lists is consistent with model-based accounts of serial learning, and is supported by a variety of experimental results (Chen et al., 1997; Merritt and Terrace, 2011; Kao et al., 2018). Insofar as each model makes inferences about position, then it stands to reason that those positions be comparable. However, no computational model of serial learning has yet been able to explain Treichler's list linking results.

To simulate such an experiment, we trained the adjacent pairs of two five item lists (ABCDE and FGHIJ) for 17 blocks (4 pairs each, counterbalanced for position, yielding 16 trials per block, and 272 trials total). In the Linking condition, this training was

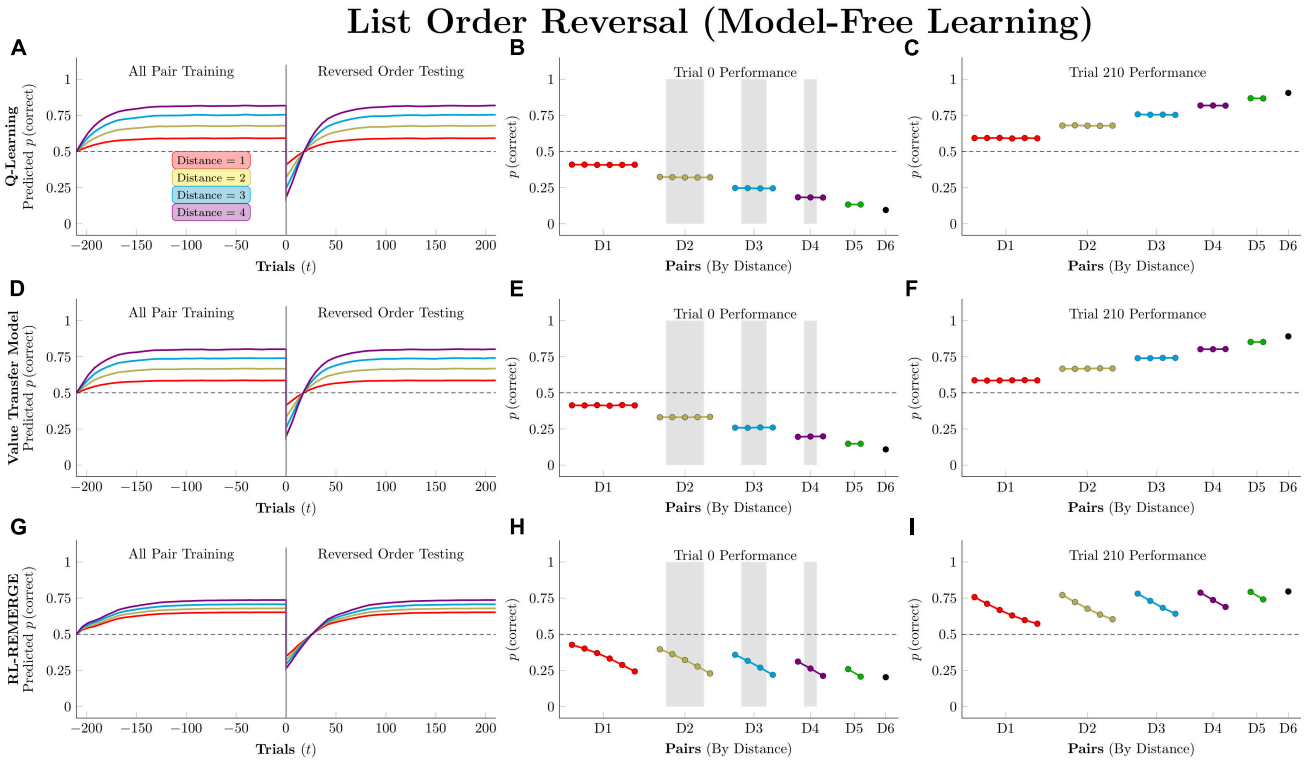


FIGURE 11 | Performance of model-free algorithms on a 7-item pairwise serial learning task, presenting all 21 pairs during training and then reversing the order of the stimuli during testing. **(A)** Mean performance of *Q*-learning on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of *Q*-learning for each of the 21 pairs at transfer, sorted by symbolic distance. Critical pairs are shaded in gray. **(C)** Mean performance of *Q*-learning for each of the 21 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Value Transfer Model (VTM). **(G–I)** As above, but reporting the simulated performance of RL-REMERGE. Reproduced with permission (DOI: 10.6084/m9.figshare.7992038.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

followed by 34 trials of training only the pair EF. In the No Linking condition, these extra 34 trials of training were skipped. Finally, the actors were trained on all 45 pairs from the resulting 10-item linked list, ABCDEFGHIJ for 5 blocks (i.e., 450 trials, given stimulus counterbalancing).

Figure 13 plots the effects of this simulated experiment. *Q*-learning's overall response accuracy for critical pairs is strictly at chance levels, but performance for individual pairs was a mixture of trials that were either very low or very high in accuracy. EF training had the effect of flipping preference for some of these pairs, but these effects canceled out, leaving the Linking and No Linking conditions no different on average. Although this effect was also visible for VTM, the net effect was that performance on some critical pairs was above chance at the end of training, while other critical pairs were below chance. RL-REMERGE, by contrast, displayed almost no difference between the two conditions at any stage of training. The only substantive difference was that the pairs EF, DF, and EG were all above chance in the Linking condition. Although this is not a dramatic effect, it is the only result we found that yields consistently-above-chance predictions for a list linking manipulation.

Figure 14 plots the effects of list linking on the performance of the model-based algorithms. RL-Elo displays a pattern of behavior similar to that of VTM, performing better than chance on some pairs at test, but worse than chance on others. It experienced no benefit from the extra training on EF. Contrastingly, although SMC initially showed a similar pattern, it showed visible benefits from EF training, improving its performance for nearly all critical pairs. Although this was not enough to bring every critical pair above chance, it nevertheless constituted a substantial gain. We interpret this as being a consequence of its crowdsourcing: Since the relative positions of the two lists were allowed to drift with respect to one another, some of the 10,000 candidate orderings were closer to a full separation of the two lists, and these "partial solutions" received a big boost during EF training because they were the ones that tended to view E as having a lower position than F. Finally, Betasort's performance during training was more in keeping with the patterns displayed by VTM and RL-Elo. Despite this, EF training had a visible effect, converging the critical pairs toward chance levels. However, this effect was not due to the algorithm converging on a better solution. It was instead the effect of the recall parameter φ relaxing the model for an additional 34 trials.

Although no algorithms solved the list linking problem, the two that came closest were RL-REMERGE (which learned some information from EF training and never displayed dramatic

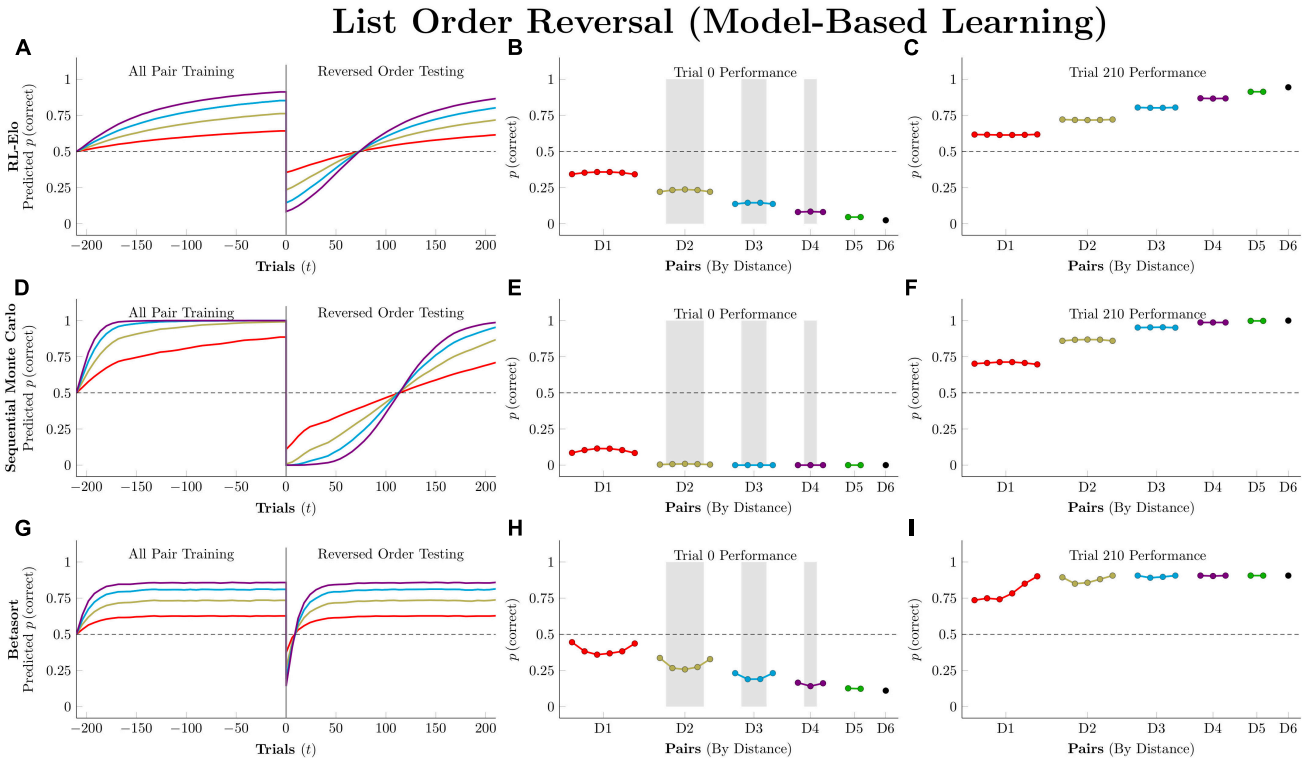


FIGURE 12 | Performance of model-based algorithms on a 7-item pairwise serial learning task, presenting all 21 pairs during training and then reversing the order of the stimuli during testing. **(A)** Mean performance of RL-Elo on non-terminal pairs throughout training and testing, sorted by symbolic distance. **(B)** Mean performance of RL-Elo for each of the 21 pairs at transfer, sorted by symbolic distance. Critical pairs are shaded in gray. **(C)** Mean performance of RL-Elo for each of the 21 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Sequential Monte Carlo (SMC) algorithm. **(G–I)** As above, but reporting the simulated performance of Betasort. Reproduced with permission (DOI: 10.6084/m9.figshare.7992041.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

below-chance performance) and SMC (which appeared to stumble upon a successful subset of models among its current 10,000 candidate orderings).

MODEL COMPARISON AND ENSEMBLE MODELING

Our emphasis thus far has been to qualitatively describe the performance of the algorithms, in order to call attention to the characteristics of behavior that arise consequent to each algorithm's value updating function and action policy. Judging whether an algorithm provides the "best fit" is not merely a matter specifying a particular experimental condition, but also of identifying the particular learning epoch that is of experimental interest. Even if no single model gives a comprehensive account of observed behavior, identifying particular epochs that resemble the performance of one algorithm or another can potentially revealing underlying mechanisms.

Using the full Hubble dataset as the empirical ground truth, a score was computed each algorithm using the Bayesian information criterion (*BIC*) (Schwarz, 1978). Since the absolute scale of BIC scores is arbitrary, we also computed ΔBIC, in which each score has the minimum observed score subtracted from its value. Since lower scores correspond with better fit, the best-fitting model overall receives a ΔBIC of 0.0. These scores are presented in **Figure 15A**. In this comprehensive scoring, the RL-Elo model provided the best fit by a substantial margin, followed by symmetric-updating *Q*-learning, and then by Betasort. Based on **Figures 6**, **7**, it appears as though *Q*-learning beat out Betasort because, although Betasort displayed a symbolic distance effect at transfer, it did not display a terminal item effect. RL-Elo proved the best-fitting model because its behavior displayed both of these phenomena.

Figure 15B plots the values of ΔBIC for each algorithm based only on a three-trial moving window of data (153 trials drawn from Hubble's 51 sessions). ΔBIC can in turn be used as a simple approximation of the weights used by an ensemble model, under the framework of Bayesian model averaging (Raftery, 1996), with each weight given by $\exp(-0.5 \cdot \Delta BIC)$ and the proportional support for each model obtained by dividing each weight by the sum of all weights. These measures of proportional support are plotted in **Figure 15C** as a cumulative plot, such that the relative contribution of each model is given by its vertical cross-section at trial *t*.

Figures 15B,C show that RL-Elo (in green) usually gave the best description of behavior at any given point during learning with an average contribution of 69.4% across all trials.

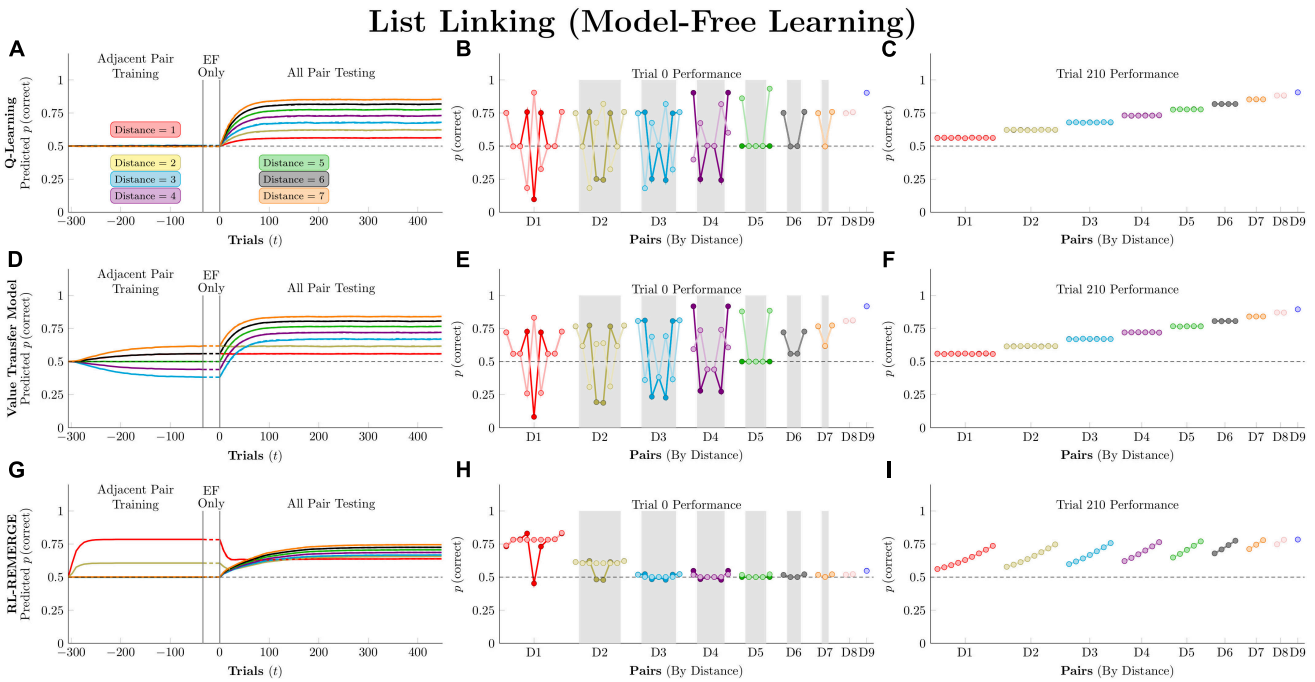


FIGURE 13 | Performance of model-free algorithms on a 10-item list-linking procedure. During training, two 5-item lists were trained (adjacent pairs only). In the Linking condition, actors received training on the linking pair EF. In the No Linking condition, no extra training on EF was provided. Actors were then tested on the 45 possible pairings in the resulting 10-item list. **(A)** Mean performance of *Q*-learning on non-terminal pairs throughout training and testing, sorted by symbolic distance. After training, dashed lines represent the Linking condition, while solid lines represent the No Linking condition. **(B)** Mean performance of *Q*-learning for each of the 45 pairs at transfer, sorted by symbolic distance. Light-colored points represent the Linking condition, while dark-colored points represent the No Linking condition. Critical pairs are shaded in gray. **(C)** Mean performance of *Q*-learning for each of the 45 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Value Transfer Model (VTM). **(G–I)** As above, but reporting the simulated performance of RL-REMERGE. Reproduced with permission (DOI: 10.6084/m9.figshare.7992044.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

Betasort provided a supporting role, averaging a contribution of 22.2% across all trials, while *Q*-learning displayed brief bursts of relevance (averaging a contribution of 3.8% for symmetric updating and 2.7% for asymmetric updating). However, despite RL-Elo generally being dominant in the ensemble, it was not the dominant model in all epochs of learning. During the early stages of learning, Betasort often gave a better description of behavior. Later in training, model-free algorithms based on associative strength (i.e., *Q*-learning and VTM) became more compelling candidates. Betasort was the preferred model during the first block of testing (i.e., trials 0–20, which included the first time each of the non-adjacent pairs was presented). For the remainder of testing however, RL-Elo dominated the ensemble model, with only occasional periods of advantage for Betasort. Throughout both training and testing, RL-REMERGE and SMC poorly described behavior, resulting in negligible contributions to the ensemble model.

Despite the ensemble model being overwhelmingly determined by the model-based algorithms, the period of relevance of model-free learning during the later stages of training could be seen as evidence of a shift from one style of learning to another. It is now widely held that organisms likely use multiple simultaneous systems for solving complex problem, but disagreement remains over whether these strategies work in parallel or cue one another in a flexible sequential fashion (Daw, 2012). The shift in the relevance of models over the course of training is suggestive of such a "two-step" approach.

It is worth emphasizing that although SMC fails utterly as a model of Hubble's behavior, it does not fail to perform TIs. **Figure 7** demonstrates unambiguously that SMC infers the ordering of the stimulus pairs, performs above chance on all stimulus pairs, and displays both the symbolic distance effect and the terminal item effect. Its abysmal ΔBIC scores (which mostly disappear off the top of the scale in **Figure 15B**) reflect performance that is much too good to be a plausible model of TI learning in monkeys. This is an example of a wider pattern in the machine learning literature, in which many algorithms that perform their assigned tasks too well are poor models for neural processes. Simulation and comparison to empirical data provides a way of falsifying such models (Palminteri et al., 2017).

By contrast, RL-Elo is a consistently compelling model because, in addition to its behavior possessing the same qualitative features as SMC, it also approximates the overall error rate observed in Hubble's data. That said, it is also important to note that RL-Elo's performance is disrupted by

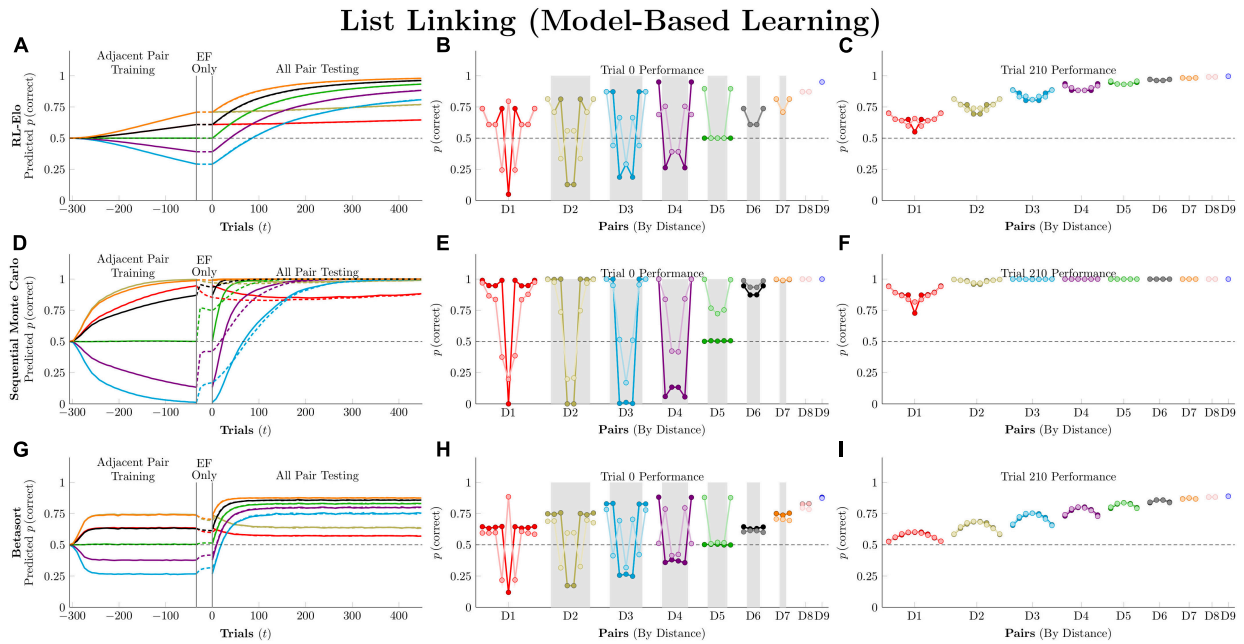


FIGURE 14 | Performance of model-based algorithms on a 10-item list-linking procedure. During training, two 5-item lists were trained (adjacent pairs only). In the Linking condition, actors received training on the linking pair EF. In the No Linking condition, no extra training on EF was provided. Actors were then tested on the 45 possible pairings in the resulting 10-item list. **(A)** Mean performance of RL-Elo on non-terminal pairs throughout training and testing, sorted by symbolic distance. After training, dashed lines represent the Linking condition, while solid lines represent the No Linking condition. **(B)** Mean performance of RL-Elo for each of the 45 pairs at transfer, sorted by symbolic distance. Light-colored points represent the Linking condition, while dark-colored points represent the No Linking condition. Critical pairs are shaded in gray. **(C)** Mean performance of RL-Elo for each of the 45 pairs at the end of training. **(D–F)** As above, but reporting simulated performance of the Sequential Monte Carlo (SMC) algorithm. **(G–I)** As above, but reporting the simulated performance of Betasort. Reproduced with permission (DOI: 10.6084/m9.figshare.7992047.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

"massed trial" manipulations (as shown in **Figure 10**), which the empirical literature has found do not undermine serial order representations in animals.

A NOTE ON COMPUTATIONAL COMPLEXITY

Several factors are considered when comparing the complexity of algorithms. The most straightforward are its memory demands: How many values must the algorithm keep track of at each stage of its operation? Another obvious measure is its operational speed: How long does the algorithm take to complete a series of decisions given reasonable inputs? Finally, it is common to discuss how an algorithm's runtime scales as a function of its inputs: As the number of input items increases, how is runtime expected to grow?

By these criteria, four of the algorithms in this manuscript are clear winners: Q-learning, VTM, RL-Elo, and Betasort. The memory demands of each scales linearly with list length n: Q-learning, VTM, and RL-Elo each keep track of n values in memory, while Betasort keeps track of $4n$ values. They are also computationally rapid. We used a 2016 Macbook Pro to simulate 42,000 trials of a TI task using all pairs from a 7-item list, and this took 0.96 s for Q-learning, 1.26 s for VTM, 1.25 s for RL-Elo, and 1.68 s for Betasort. Finally, all four algorithms are expected have order n complexity, scaling their runtime chiefly as a function of the list length.

RL-REMERGE and SMC perform comparatively poorly. In the case of RL-REMERGE, the Conjunctive layer ensures that its memory and runtime demand grow on the order of n^2. Even at reasonable list lengths, its iterative internal use of renormalization to discover a steady state is computationally intensive (42,000 simulated trials took 2879.46 s). While SMC scales more reasonably, with growth of order n, its 10,000 hypotheses constitute a substantial opportunity cost, both in terms of memory and processing. These are reflected in its costly runtime: 42,000 simulated trials took SMC 653.72 s.

However, these measures are somewhat unfair to both because these algorithms represent distributed solutions to the TI problem. Renormalization of neural networks, for example, is merely a mathematical trick to emulate the behavior of a distributed physical system, so RL-REMERGE's n^2 growth is mainly a problem of memory. Provided one is willing to allocate the neurons needed, processing time would be distributed accordingly. In the same vein, SMC can be understood as 10,000 parallel implementations of RL-Elo. Provided one is willing to distribute the problem in this way (and make use of many processor cores to do so), SMC's performance should be very similar to RL-Elo (although the computational

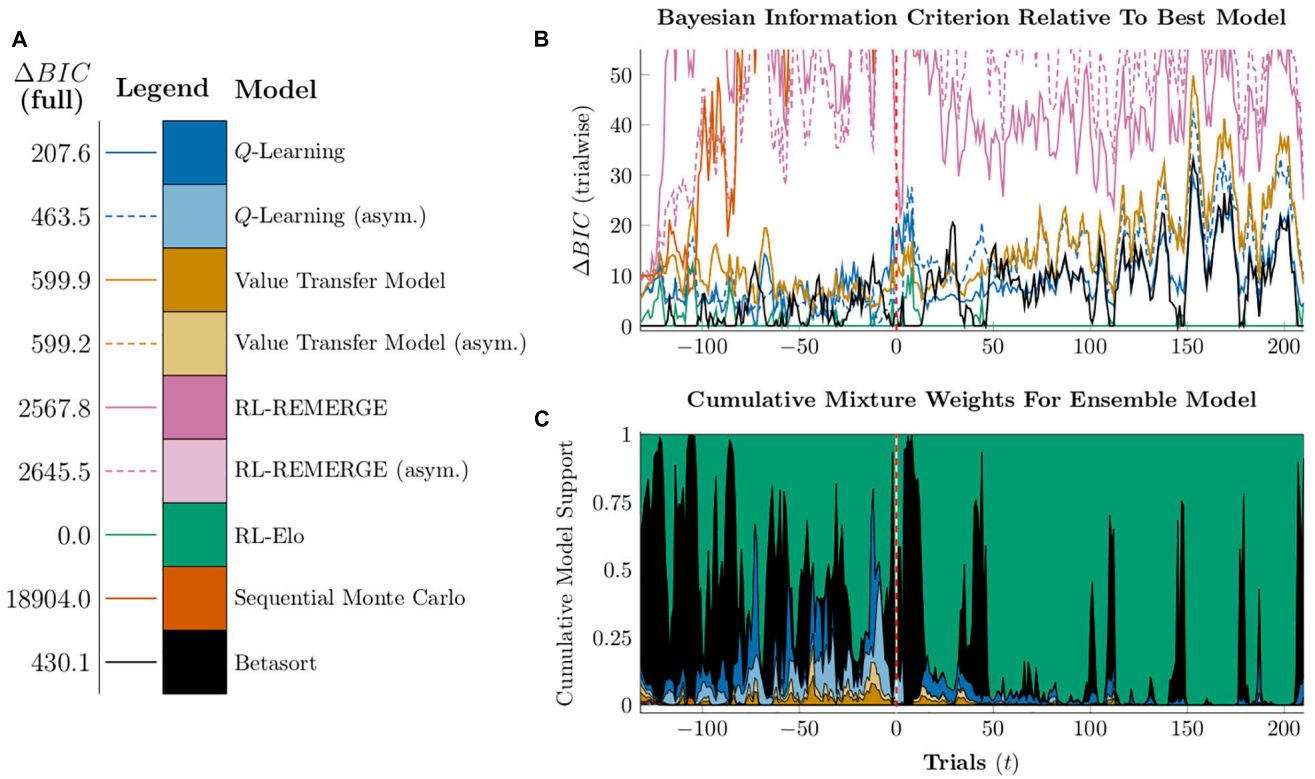


FIGURE 15 | Model comparison of nine learning algorithms performing transitive inference. **(A)** Comparison for model-free algorithms (symmetric and asymmetric updating) and model-based algorithm using the full Hubble dataset, as well as figure legend for remaining panels. ΔBIC refers to the Bayesian information criterion computed for each model, minus the value computed for the best model. **(B)** Value of ΔBIC for each model, computed trialwise using a 3-trial moving window. Trial zero corresponds with the start of all-pairs testing, denoted by a red-and-white dashed line. **(C)** Cumulative plot of the weights for an ensemble model with contributions from all nine algorithms. Weights for Bayesian averaging were approximated using $\exp(-0.5 \cdot \Delta BIC_i)$ for each algorithm i. Reproduced with permission (DOI: 10.6084/m9.figshare.8851700.v1). Copyright 2019, Jensen, Terrace, and Ferrera.

burden of periodic resampling of the list orders is less easily resolved).

Our view is that although complexity should be taken seriously when comparing computational models, it should also be recognized that these algorithms are primarily high-level descriptions of behavior. The Betasort algorithm should not, for example, be viewed as a proposal that organisms are actually computing and drawing random samples from beta distributions. Rather, the beta distribution is the maximum entropy distribution for the parameter uncertainty of Bernoulli processes, and as such there are many physical systems for which it will provide a good approximation. Any of the algorithms in this manuscript could be seen as a higher-level description of some distributed process, and RL-REMERGE and SMC appear less efficient because they explicitly distribute their processing.

With these considerations in mind, we consider RL-REMERGE to be the least plausible of the algorithms, since it requires implementing nodes for every "context." In standard TI experiments, this takes the form of stimulus pairs, but it is easy to imagine a task in which subjects have to choose from among three or four stimuli, causing the Conjunctive layers to grow even faster since it would need a node for every stimulus combination. Contrastingly, the other five algorithms are all able to scale linearly as a function of list length, regardless of the number of stimuli presented in any given trial.

CONCLUSION

The algorithms in this manuscript represent a cross-section of different computational methods for explaining TI performance. These methods succeed to varying degrees in discovering the structure of ordered sets, even when stimuli are presented pairwise without explicit spatial or temporal cues. Our goal was to gain insight about how each handles challenges that arise in natural settings, such as unequal rewards and base rates of presentation, as well as established experimental manipulation, such as list linking. In every case other than *Q*-learning (which was included to provide a model-free baseline), an argument can be made that each of these algorithms can, to some extent, "solve" TI problems. The patterns of error vary dramatically from one algorithm to the next, however, as do the effects of variations in the procedure.

While we can still learn a great deal from traditional TI procedures, the ability to implement theories of TI as computational models demands a shift in the experimental

approach used to study serial learning. In the past, experimentalists have largely avoided making dramatic changes to experimental procedures because their theories were not sufficiently clear about what behavior would be expected under these new designs. By making computational simulation a part of both theorizing and of experimental design, future experiments can much more rapidly identify areas of discrepancy between theory and behavior, which in turn lead to both stronger theories and to more informative experiments.

AUTHOR CONTRIBUTIONS

GJ performed the statistical analysis and wrote the first draft of the manuscript. GJ and VF contributed code to the study. All authors contributed to conception and design of the study, manuscript revision, and read and approved the submitted version.

FUNDING

This work was supported by US National Institute of Mental Health, grant numbers NIH-MH081153 and NIH-MH111703 awarded to VF and HT.

ACKNOWLEDGMENTS

We thank Ning Qian for comments on an earlier version of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.00878/full#supplementary-material

REFERENCES

Acuna, B., Sanes, J. N., and Donoghue, J. P. (2002). Cognitive mechanisms of transitive inference. *Exp. Brain Res.* 146, 1–10. doi: 10.1007/s00221-002-1092-y

André, J. M., Cordero, K. A., and Gould, T. J. (2012). Comparison of the performance of DBA/2 and C57BL/6 mice in transitive inference and foreground and background contextual fear conditioning. *Behav. Neurosci.* 126, 249–257. doi: 10.1037/a0027048

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., et al. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* 100, 490–509. doi: 10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Deuker, L., Schröder, T. N., and Doeller, C. F. (2016). Grid-cell representations in mental simulation. *eLife* 5:e17089. doi: 10.7554/eLife.17089

Bürkner, P.-C., and Vuorre, M. (2018). Ordinal regression models in psychology: a tutorial. *PsyArXiv*

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *J. Exp. Pedagogy* 1, 93–112. doi: 10.1016/j.jecp.2011.06.004

Bush, R. R., and Mosteller, F. (1951). A mathematical model for simple learning. *Psychol. Rev.* 58, 313–323. doi: 10.1037/h0054388

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32.

Chalmers, E., Luczak, A., and Gruber, A. J. (2016). Computational properties of the hippocampus increase the efficiency of goal-directed foraging through hierarchical reinforcement learning. *Front. Comput. Neurosci.* 10:128. doi: 10.3389/fncom.2016.00128

Chen, S., Swartz, K. B., and Terrace, H. S. (1997). Knowledge of the ordinal position of list items in rhesus monkeys. *Psychol. Sci.* 8, 80–86. doi: 10.1111/j.1467-9280.1997.tb00687.x

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science* 352, 1464–1468. doi: 10.1126/science.aaf0941

Couvillon, P. A., and Bitterman, M. E. (1992). A conventional conditioning analysis of "transitive inference" in pigeons. *J. Exp. Psychol.* 18, 308–310. doi: 10.1037//0097-7403.18.3.308

D'Amato, M. R., and Colombo, M. (1990). The symbolic distance effect in monkeys (*Cebus apella*). *Anim. Learn. Behav.* 18, 133–140. doi: 10.3758/bf03205250

Daniels, C. W., Laude, J. R., and Zentall, T. R. (2014). Six-term transitive inference with pigeons: successive-pair training followed by mixed-pair training. *J. Exp. Anal. Behav.* 101, 26–37. doi: 10.1002/jeab.65

Daw, N. D. (2012). "Model-based reinforcement learning as cognitive search: Neurocomputational theories," in *Cognitive Search: Evolution, Algorithms, and the Brain*, eds P. M. Todd, T. T. Hills, and T. W. Robbins (Cambridge, MA: MIT Press), 195–208.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for bayesian filtering. *Stat. Comput.* 10, 197–208. doi: 10.1109/TPAMI.2008.134

Eichenbaum, H. B., Komorowski, R., MacDonald, C. J., and Kraus, B. J. (2016). "How does the hippocampus support the spatial and temporal attributed of memory?," in *The Neurobiological Basis of Memory: A System, Attribute, and Process Analysis*, eds P. A. Jackson, A. A. Chiba, R. F. Berman, and M. E. Ragozzino (Cham: Springer), 39–56.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. New York, NY: Arco Publishing.

Ferguson, T. D., Williams, C. C., Skelton, R. W., and Krigolson, O. E. (2019). Passively learned spatial navigation cues evoke reinforcement learning reward signals. *Cognition* 189, 65–75. doi: 10.1016/j.cognition.2019.03.015

Gazes, R. P., Chee, N. W., and Hampton, R. R. (2012). Cognitive mechanisms for transitive inference performance in rhesus monkeys: measuring the influence of associative strength and inferred order. *J. Exp. Psychol.* 38, 331–345. doi: 10.1037/a0030306

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. U.S.A.* 108, 15647–15654. doi: 10.1073/pnas.1014269108

Grice, G. R. (1972). Application of a variable criterion model to auditory reaction time as a function of the type of catch trial. *Percept. Psychophys.* 12, 103–107. doi: 10.3758/bf03212853

Jensen, G. (2017). "Serial learning," in *APA Handbook of Comparative Psychology: Vol. 2, Perception, Learning, & Cognition*, eds J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, and T. Zentall (Washington DC: American Psychological Association), 385–409.

Jensen, G., Alkan, Y., Ferrera, V. P., and Terrace, H. S. (2018). Reward associations do not explain transitive inference performance in monkeys. *PeerJ* 6:e26889v2.

Jensen, G., Alkan, Y., Muñoz, F., Ferrera, V. P., and Terrace, H. S. (2017). Transitive inference in humans (*Homo sapiens*) and rhesus macaques (*Macaca mulatta*) after massed training of the last two list items. *J. Comp. Psychol.* 131, 231–245. doi: 10.1037/com0000065

Jensen, G., Altschul, D., Danly, E., and Terrace, H. S. (2013). Transfer of a serial representation between two distinct tasks by rhesus macaques. *PLoS One* 8:e70285. doi: 10.1371/journal.pone.0070285

Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P., and Terrace, H. S. (2015). Implicit value updating explains transitive inference performance: the betasort model. *PLoS Comput. Biol.* 11:e1004523. doi: 10.1371/journal.pcbi.1004523

Justensen, N., Bontrager, P., Togelius, J., and Risi, S. (2019). *Deep Learning for Video Game Playing*. Paris: IEEE Transactions on Games.

Kao, T., Jensen, G., Michaelcheck, C., Ferrera, V. P., and Terrace, H. S. (2018). Absolute and relative knowledge of ordinal position. *PeerJ* 6:e26453v1.

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., and Dayan, P. (2016). Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* 92, 1135–1147. doi: 10.1016/j.neuron.2016.10.052

Kumaran, D., and McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol. Rev.* 119, 573–616. doi: 10.1037/a0028681

Lazareva, O. F., and Wasserman, E. A. (2006). Effects of stimulus orderability and reinforcement history on transitive responding in pigeons. *Behav. Process.* 72, 161–172. doi: 10.1016/j.beproc.2006.01.008

Lazareva, O. F., and Wasserman, E. A. (2012). Transitive inference in pigeons: measuring the associative value of stimuli B and D. *Behav. Process.* 89, 244–255. doi: 10.1016/j.beproc.2011.12.001

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: Wiley.

Macphail, E. M. (1987). The comparative psychology of intelligence. *Behav. Brain Sci.* 10, 645–695.

McGonigle, B. O., and Chalmers, M. (1977). Are monkeys logical? *Nature* 267, 694–696. doi: 10.1038/267694a0

Merritt, D., MacLean, E. L., Jaffe, S., and Brannon, E. M. (2007). A comparative analysis of erial ordering in ring-tailed lemurs (*Lemur catta*). *J. Comp. Psychol.* 121, 363–371. doi: 10.1037/0735-7036.121.4.363

Merritt, D. J., and Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (*Homo sapiens*) and rhesus monkeys (*Macaca mulatta*). *J. Comp. Psychol.* 125, 227–238. doi: 10.1037/a0021572

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Neumann, C., Duboscq, J., Dubuc, C., Ginting, A., Irwan, A. M., Agil, M., et al. (2011). Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating. *Anim. Behav.* 82, 911–921. doi: 10.1016/j.anbehav.2011.07.016

Oliva, A., Fernández-Ruiz, A., Buzsáki, G., and Berényi, A. (2016). Spatial coding and physiological properties of hippocampal neurons in the cornu ammonis subregions. *Hippocampusogie* 26, 1593–1607. doi: 10.1002/hipo.22659

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* 21, 425–433. doi: 10.1016/j.tics.2017.03.011

Piaget, J. (1921). Une forme verbal de la comparison chez l'enfant. *Arch. de Psychol.* 18, 141–172.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 83, 251–266. doi: 10.1093/biomet/83.2.251

Raineri, E., Dabad, M., and Heath, S. (2014). A note on the exact difference between beta distributions in genomic (methulation) studies. *PLoS One* 8:e97349. doi: 10.1371/journal.pone.0097349

Redish, A. D. (1999). *Beyond the Cognitive Map: From Place Cells to Episodic Memory*. Cambridge, MA: MIT Press.

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement," in *Classical Conditioning II: Current Research & Theory*, ed. W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rodriguez, P., and Levy, W. B. (2004). Configural representation in transverse patterning with a hippocampal model. *Neural Netw.* 17, 175–190. doi: 10.1016/j.neunet.2003.06.001

Samothrakis, S., Perez, D., Lucas, S. M., and Rohlfshagen, P. (2016). Predicting dominance ranking for score-based games. *IEEE Trans. Comput. Intell. AI Games* 8, 1–12. doi: 10.1109/tciaig.2014.2346242

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA,: MIT Press.

Terrace, H. S. (2010). The comparative psychology of serially organized behavior. *Comp. Cogn. Behav. Rev.* 5, 23–58. doi: 10.3819/ccbr.2010.50002

Terrace, H. S. (2012). "The comparative psychology of ordinal behavior," in . *Oxford Handbook of Comparative Cognition*, eds T. R. Zentall and E. A. Wasserman (Oxford, UK: Oxford University Press), 615–651.

Treichler, F. R., and Raghanti, M. A. (2010). Serial list combination by monkeys (*Macaca mulatta*): test cues and linking. *Anim. Cogn.* 13, 121–131. doi: 10.1007/s10071-009-0251-y

Treichler, F. R., Raghanti, M. A., and Van Tilburg, D. N. (2003). Linking of serially ordered lists by macaque monkeys (*Macacca mulatta*): list position influences. *J. Comp. Psychol.* 121, 250–259. doi: 10.1037/0735-7036.121.3.250

Treichler, F. R., Raghanti, M. A., and Van Tilburg, D. N. (2007). Serial list linking by macaque monkeys (*Macacca mulatta*): list property limitations. *J. Comp. Psychol.* 121, 250–259. doi: 10.1037/0735-7036.121.3.250

Treichler, F. R., and Van Tilburg, D. N. (1996). Concurrent conditional discrimination tests of transitive inference by macaque monkeys: list linking. *J. Exp. Psychol.* 22, 105–117. doi: 10.1037//0097-7403.22.1.105

van Seijen, H., Fatemi, M., Romoff, J., Laroche, R., Barnes, R., and Tsang, J. (2017). "Hybrid reward architecture for reinforcement learning," in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Reed Hook, NY: Curran Associates, Inc), 5392–5402.

Vasconcelos, M. (2008). Transitive inference in non-human animals: an empirical and theoretical analysis. *Behav. Process.* 78, 313–334. doi: 10.1016/j.beproc.2008.02.017

Vasconcelos, M., and Monteiro, T. (2014). European starlings unriddle the ambiguous-cue problem. *Front. Psychol.* 5:944. doi: 10.3389/fpsyg.2014.00944

von Fersen, L., Wynne, C. D. L., Delius, J. D., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *J. Exp. Psychol.* 17, 334–341. doi: 10.1037/0097-7403.17.3.334

Watkins, C. J. C. H., and Dayan, P. (1992). *Q*-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698

Weaver, J. E., Steirn, J. N., and Zentall, T. R. (1997). Transitive inference in pigeons: control for differential value transfer. *Psychon. Bull. Rev.* 4, 113–117. doi: 10.3758/bf03210782

Widrow, B., and Hoff, M. E. (1960). *Adaptive Switching Circuits: Stanford Electronic Laboratories*. Standford, CA: Standford University. Technical Report, 1553-1.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* 8, 229–256. doi: 10.1007/bf00992696

Wu, X., and Levy, W. B. (1998). "A hippocampal-like neural network model solves the transitive inference problem," in *Computational Neuroscience: Trends in Research*, ed. J. M. Bower (New York, NY: Plenum Press), 567–572. doi: 10.1007/978-1-4615-4831-7_94

Wynne, C. D. L. (1995). Reinforcement accounts for transitive inference in pigeons. *Anim. Learn. Behav.* 23, 207–217. doi: 10.3758/bf03199936

Wynne, C. D. L. (1997). Pigeon transitive inference: tests of simple accounts of a complex performance. *Behav. Process.* 39, 95–112. doi: 10.1016/s0376-6357(96)00048-4

Yi, S. K. M., Steyvers, M., and Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *J. Probl. Solving* 2, 81–100.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Jensen, Terrace and Ferrera. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 20 August 2019
doi: 10.3389/fnins.2019.00821

The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines

David B. Parker[1]* *and Qolamreza R. Razlighi*[1,2,3]

[1] Department of Biomedical Engineering, Columbia University, New York, NY, United States, [2] Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY, United States, [3] Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, United States

Due to the nature of fMRI acquisition protocols, slices cannot be acquired simultaneously, and as a result, are temporally misaligned from each other. To correct from this misalignment, preprocessing pipelines often incorporate slice timing correction (STC). However, evaluating the benefits of STC is challenging because it (1) is dependent on slice acquisition parameters, (2) interacts with head movement in a non-linear fashion, and (3) significantly changes with other preprocessing steps, fMRI experimental design, and fMRI acquisition parameters. Presently, the interaction of STC with various scan conditions has not been extensively examined. Here, we examine the effect of STC when it is applied with various other preprocessing steps such as motion correction (MC), motion parameter residualization (MPR), and spatial smoothing. Using 180 simulated and 30 real fMRI data, we quantitatively demonstrate that the optimal order in which STC should be applied depends on interleave parameters and motion level. We also demonstrate the benefit STC on sub-second-TR scans and for functional connectivity analysis. We conclude that STC is a critical part of the preprocessing pipeline that can be extremely beneficial for fMRI processing. However, its effectiveness interacts with other preprocessing steps and with other scan parameters and conditions which may obscure its significant importance in the fMRI processing pipeline.

Keywords: slice timing correction, fMRI — functional magnetic resonance imaging, motion correction, preprocessing algorithms, interleaved 2D multislice sequence

OPEN ACCESS

Edited by:
Reza Lashgari,
Brain Engineering Research Center, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Anders Eklund,
Linköping University, Sweden
Martin J. McKeown,
The University of British Columbia, Canada

***Correspondence:**
David B. Parker
dbp2123@columbia.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *10 April 2019*
Accepted: *23 July 2019*
Published: *20 August 2019*

Citation:
Parker DB and Razlighi QR (2019) The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines. Front. Neurosci. 13:821. doi: 10.3389/fnins.2019.00821

INTRODUCTION

Most functional magnetic resonance image (fMRI) scans are acquired using echo planar imaging (EPI), which rapidly acquires single or multiple 2D slices and stacks them to create a 3D volume. This process typically takes between 0.5 and 4 s (known as the repetition time, or TR), depending on the fMRI pulse sequence, field of view (FOV), and number of acquired slices (Stehling et al., 1991). Slices can be acquired sequentially or with interleave, in which a number of slices are skipped and acquired later during the TR. Interleaving allows more time for the partially excited spins in the adjacent slices to return to equilibrium, which attenuates the artifacts of radio frequency (RF) pulse excitation leakage. Regardless of the manner in which slices are acquired, they cannot be acquired instantaneously, and so the result is an accumulation of offset delays between the first slice and all remaining slices. In order for an accurate time series analysis to be carried out on the fMRI data, these temporal offsets between slices must be corrected for. Slice timing correction (STC) is the

preprocessing step applied to correct for these slice-dependent delays, achieved by shifting the time series of each slice to temporally align all slices to a reference time-point.

In addition to STC, numerous preprocessing steps are required to prepare fMRI data for statistical analysis. Surprisingly, there exists no consensus on the order in which these preprocessing steps should occur, and the interaction of STC on motion parameter estimates has only recently been investigated (Cox, 1996; Power et al., 2017). It has been shown in previous studies that the order and the steps included in the preprocessing pipeline significantly affect the resulting statistical analysis (Zhang et al., 2009; Carp, 2012; Churchill et al., 2012, 2015). One previous report examined motion correction (MC), physiological noise correction, and motion parameter residualization (MPR) (Churchill et al., 2012), however, in this study, the effect of these steps was only evaluated based on their inclusion or omission from a fixed order in the pipeline, and STC was performed at the beginning of every pipeline. Thus, the contribution of STC on different steps in the pipeline remains unclear. To further contribute to this area of research, we examine whether the order of these steps will significantly impact resulting statistical parametric maps.

In a previous study, we directly compared the performance of different STC methods to each other (Parker et al., 2016). In the current study, we aimed to examine the interaction of these STC methods with other preprocessing steps. We have identified five areas common in fMRI analysis where the role of STC is largely uninvestigated: (1) The order of STC and MC in the preprocessing pipeline without spatial smoothing, (2) The order of STC and MC, and its interaction with MPR, without spatial smoothing, (3) The interaction of spatial smoothing and STC, (4) STC on data sampled with a fast TR, and (5) STC on functional connectivity analysis using independent component analysis (ICA). Using both real and simulated data, we quantitatively examined the effect of STC on data acquired with various levels of motion and slice acquisition sequences. We hypothesized that different combinations of motion and slice acquisition orders would alter the effectiveness of different preprocessing pipelines. We also use different STC techniques to see if the STC algorithm has any effect on pipeline order. We use FSL and SPM's slice timing routines, where FSL implements a low-order Hanning windowed sinc interpolation, while SPM uses a frequency-domain phase shift. Finally, we apply a STC method developed by our group, FilterShift (FS), which implements a moderate-order Kaiser-windowed sinc function.

Pipeline Order Without Spatial Smoothing

Some studies suggest that the optimal order of STC and other preprocessing steps such as spatial realignment depends on both the level of motion and the slice acquisition (Or interleave) order (Sladky et al., 2011). Motion may exacerbate the STC problem, introducing not only shifts in time but also in space. The effects of motion will change based on the slice acquisition orders. Different slice acquisition orders will result in visually different volumes, even if they undergo identical motion. This will lead to inaccurate and inconsistent motion parameter estimates. In these instances, it has been shown that STC before spatial realignment alters the motion parameter estimates and result in less accurate spatial realignment (Power et al., 2017). Likewise, as spatial realignment applies spatial transformations to the data, a voxel's time series after spatial realignment may contain values interpolated from adjacent slices which were sampled at different offsets. Thus, applying a single time shift to a voxel containing data sampled with different delays is intuitively flawed. Therefore, it is important to investigate which order of the pipelines would be optimal in the preprocessing pipeline.

Motion Parameter Residualization

Motion parameter residualization is the practice of including motion parameters in the GLM as nuisance regressors used for residualization which orthogonalizes the BOLD activity from the motion parameters. Because the motion parameters are typically estimated during the realignment step, the effectiveness of motion parameter residualization may change depending on the level of motion, the accuracy of the motion parameter estimation, and the order in which STC and MC are performed. The strategies for MPR vary greatly from study to study. Some groups only include the six rotation and translation parameter estimates, while others include any number of additional derivatives and quadratics in an attempt to capture any non-linear motion related signal fluctuations (Satterthwaite et al., 2013). There is still active research regarding the best kinds of covariates to include in the residualization, and the process is a common addition to modern preprocessing pipelines (Ciric et al., 2017). Regardless of the specific covariates used, the effectiveness of MPR relies on accurate motion parameter estimates. Given the interaction between subject motion and slice timing, it follows that the STC and MC preprocessing steps will have an impact on the motion estimates used for MPR, and indeed the MC itself, as it has been shown in the field (Power et al., 2017). Thus we investigate the effect of MPR on the effectiveness of STC in this paper.

Spatial Smoothing on STC

Spatial smoothing is another preprocessing step that interacts with STC. Spatial smoothing has the effect of distributing a voxel's intensity to any surrounding voxels that fall within the smoothing kernel, thus altering their time series. With a relatively small smoothing kernel, this may impact slices differently, depending on their delay and the type of interleave acquisition. Slices with large delay will have the most error, and will be averaged with adjacent slices with smaller delay. Depending on noise, motion, and slice acquisition order, this can attenuate some of the error in the large delay slice, and create a more accurate time series. Conversely, slices with small delays and errors may be averaged with adjacent slices with large delays and errors, reducing their accuracy. Because spatial smoothing operates across slices with different delays, it is important for us to investigate the interaction of STC, a slice-specific process, and spatial smoothing, a multi-slice process.

STC on Short TRs

As new imaging techniques such as simultaneous multi-slice acquisition (SMS) become more and more common, acquisition times of whole volumes become faster (Barth et al., 2016). It is generally considered that with such short TR, STC is an unnecessary step with little value. This is based on the assumption that the BOLD signal is generally quite slow, and so any signal change due to temporal offsets is likely to also be small. However, the GLM is extremely sensitive to small shifts in signals, and to our knowledge no study has quantitatively demonstrated that STC is futile for short TR acquisitions. With the development of multiband EPI and improvements in the pulse sequence, whole brain volumes can be acquired in just fractions of a second (Feinberg et al., 2010). In our previous study, we showed on simulated data that traditional STC techniques showed little to no benefit with a fast TR, while our method still provided significant improvement (Parker et al., 2016). In this paper, we extend that analysis to real data acquired with a short TR. The Human Connectome Project (HCP) is one such study that employs this acquisition technique to achieve a TR of 0.72 s for task-based and resting-state fMRI. The HCP publishes an extensive pre-processing pipeline for all their data, which ignores STC due to the high sampling at which the volumes are acquired (Glasser et al., 2013). It is argued that the benefit from STC is insignificant at higher sampling rates. Because we have found no thorough investigations into this matter, it is important for us to examine the effect of STC on short-TR data.

STC on Functional Connectivity Analysis

Functional connectivity (FC) analyses are becoming increasingly popular in the current literature (Sala-Llonch et al., 2015). In FC, it is not necessary to use an external regressor for a GLM analysis. There are many methods to extract inherent FC, such as seed based connectivity (Biswal et al., 1995a), ICA (Beckmann and Smith, 2004), and even graph theory (Buckner et al., 2009; Rubinov and Sporns, 2010). These methods rely only on the intrinsic fluctuations in each voxel, and its similarity to other voxels in the brain to extract functionally connected regions. Previous work has already examined the interaction of STC with resting state data connectivity, and have shown no significant effect (Wu et al., 2011). However, this study employed a 6mm spatial smoothing kernel to the data, which will greatly reduce the effectiveness of STC. It's also important to consider that resting state fluctuations (in the default mode network, for example) are very slow, low frequency signals (<0.1 Hz) (Biswal et al., 1995b). It has already been shown that STC has little effect for low-frequency signals, such as block design tasks (Sladky et al., 2011), and we expect the same principle to carry over to resting state fMRI (rsfMRI). Functional connectivity analysis can be extracted from resting-state fMRI data as well as task-based fMRI data (Biswal et al., 1995a). In fact, recently task-based functional connectivity networks have gained much more popularity in the field (Di et al., 2013; Krienen et al., 2014; Stern et al., 2014). Our group recently demonstrated that one of the well-recognized functional connectivity networks, the default mode network, can be extracted from both task-based and resting-state fMRI data (Razlighi, 2018). Regardless of whether the scan is "task" or "resting," the process of extracting independent components remains identical, however, given the relatively faster temporal characteristics of task-based BOLD signals (compared to resting-state BOLD signals), we expect STC to have a larger impact on FC when used on task-based fMRI. In our analysis, we used ICA to extract functional components. Because of the popularity of STC, and the missing investigation into task-based FC, we examine the effect of STC on functional connectivity analysis.

MATERIALS AND METHODS

Simulated Data

It is difficult to evaluate the performance of preprocessing methods on real fMRI data because the true underlying BOLD signal is always unknown. Because of this, we created simulated datasets with a known, underlying BOLD signal for a quantitative comparison. Data was simulated using the same procedure described in Parker et al. (2016). Briefly, our simulated fMRI scans used real subject brain morphology by temporally averaging all volumes from a motion corrected real subject fMRI scan. We used the same subject's structural segmentation from FreeSurfer (Fischl et al., 2002, 2004), and inter-modal rigid-body registration with FSL (Greve and Fischl, 2009) to obtain anatomical ROI masks in the fMRI space. We simulated neuronal activity consisting of sequences of 20 boxcar pulses with jittered onsets and randomly generated durations for each ROI. These stimuli were convolved with the canonical HRF to generate the hemodynamic response. Both HRF and neuronal stimuli were sampled at high frequency to facilitate in simulating motion and slice sampling. Cardiac and respiratory artifacts were simulated by adding a single sinusoid at f_c = 1.23 Hz for cardiac and another sinusoid at f_r = 0.25 Hz for respiratory noise. The magnitude of the cardiac artifacts is modulated based on the power of the cardiac artifact found in a real fMRI dataset. To do this, the Circle of Willis was manually identified, where the cardiac signal dominates the fMRI time-series. Next, the frequency range of the aliased cardiac signal is detected from the signal at the Circle of Willis. Then, the power spectrum of every voxel was examined in the detected frequency range, and the amplitude of the simulated physiological artifact is scaled with the power computed from the detected frequency range. Finally, a low thermal noise level was added to the simulation, consisting of white noise that made up 5% of the signal's energy. The hemodynamic signal was scaled so that the standard deviation equals 4% of the mean signal's magnitude, comparable to a robust signal in the visual cortex. This scaled signal was added to the mean value image. Because we know exactly where the signal is located, we can compare voxels directly from these regions without the need for spatial normalization.

We simultaneously simulated the effect of head motion and slice timing as well as their interaction. To do this, we use motion parameters estimated from real subject scans. We up-sampled these motion parameters from real subject movement inside the scanner (which includes rotations and translations along all three axes) by the number of slices in the simulated scans, using

spline interpolation. For instance, if there were 40 slices in each volume, we up-sampled the motion parameters by 40 times the original fMRI sampling frequency (1/TR). We then transformed the brain volume 40 times according to the upsampled motion parameters between each pair of consecutively acquired fMRI volumes. This gave us the position of the brain volume at the instance of each slice acquisition. Then we sampled the slice from the transformed volumes according to the timing of each slice. The high temporal-resolution BOLD signal present in each voxel was also sampled according to the timing of each slice, which simulated slice timing, creating a lag or delay in sampling between slices. A more detailed description about this simulation can be found in Parker et al. (2016).

We generated 20 simulated scans from a single subject's morphology. Each subject was given a unique BOLD time series for every region of the brain. For each time series, a simulated scan was created with three levels of motion: high, medium, and low. The motion levels were classified by a set of real subjects' mean frame-wise displacement (mFWD) inside the scanner (Low: mFWD $<$ 0.1 mm, med: 0.25mm $<$ mFWD $<$ 0.4 mm, high: 0.6 mm $<$ mFWD $<$ 0.7 mm) (Power et al., 2012). Simulated data with each motion level was then sampled with the three interleave schema; sequential (Interleave 1), even odd (Interleave 2), and every 6th (Interleave 6). Thus, each subject's fMRI scan was simulated a total of nine times, each with unique interleave and motion parameters. Synthesized fMRI scans consisted of 10 min of scanning with an in-plane acquisition matrix of 112 $\times$ 112, and 37 slices. The voxel size was set to 2 mm $\times$ 2 mm $\times$ 3 mm. The TR was equal to 2 s for all experiments. Thus, with twenty BOLD time series each simulated under nine different conditions, we created 180 unique simulated fMRI scans.

Real Data

We chose a task with high temporal sensitivity and a robust response for this study. Thirty right-handed healthy subjects (17/13 young/old; proportion female/male: 0.53/0.61, age mean $\pm$ std: 25.5/64.9 $\pm$ 2.4/2.2 years) were presented with event related visual (flashing checker boards) stimuli with random onset and duration while undergoing fMRI. To ensure attention to the stimuli, subjects responded with a button press at the conclusion of each visual stimulus. Functional images were acquired using a 3.0 Tesla Achieva Philips scanner with a field echo echo-planar imaging (FE-EPI) sequence [TE/TR = 20 ms/2000 ms; flip angle = 72°; 112 $\times$ 112 matrix size; in-plane voxel size = 2.0 mm $\times$ 2.0 mm; slice thickness = 3.0 mm (no gap); 41 transverse slices per volume, 6:1 Philips interleaved, in ascending order]. Participants were scanned for 5.5 min, with at least 37 events of visual and auditory stimuli. Subjects were stratified based on their mFWD over the entire scan period. Ten low motion (mFWD $<$ 0.14 mm), ten medium motion (0.14 mm $\leq$ mFWD $<$ 0.2mm), and ten high motion (mFWD $\geq$ 0.2 mm) subjects were selected for each group. We used this data for both GLM analysis and the model-free ICA analysis commonly used in FC analysis of fMRI data.

In addition, we used the "*100-unrelated*" dataset from the Human Connectome Project (HCP), which consisted of 100 healthy and young participants (age range = 22–35 years, m/f = 52/48) to evaluate the benefit of STC on data with short TR and multiband acquisition (Van Essen et al., 2013). HCP data were acquired on a customized Siemens 3T Skyra scanner with a multiband EPI sequence [TE/TR = 33.1ms/720 ms; flip angle = 52°; FOV = 208 $\times$ 180 mm; voxel size = 2.0 $\times$ 2.0 $\times$ 2.0 mm; 72 axial slices; multiband factor = 8]. We used the N-back working memory scan, and modeled all correct responses as an event-related regressor.

fMRI Data Statistical Analysis

Assessing the effectiveness of STC methods is extremely challenging due to its interaction with motion (Kim et al., 1999), and the fact that its improvement is slice-dependent. The benefit of STC could appear absent if the majority of activation falls on a slice with low temporal delay due to the fact that very little temporal shifting is encountered in these slices. This is analogous to examining the effect of motion correction on data with extremely low motion. In most typical STC techniques, one slice is chosen as the "reference slice," and all other slices are shifted temporally to match the time at which the reference slice was acquired. The effect we see due to STC is directly proportional to the size of this temporal shift. In fact, if the activation falls on the reference slice, then no shifting occurs, and we will see absolutely no difference between STC and uncorrected data. This could be mistaken as evidence that STC may not be required in the preprocessing pipeline. Furthermore, depending on an individual's brain size, position and head orientation in scanner, the same brain region may fall on different slices with different acquisition delays from subject to subject. This makes it extremely difficult to compare STC methods on a given region across subjects, as shown in our previous publication (Parker et al., 2016). Therefore, it is important to evaluate the effectiveness of STC on slices with the maximal temporal delay across subjects. For all experiments in this study, we choose the first slice acquired at the beginning of the TR as our reference slice. Therefore, a "large delay slice" refers to a slice acquired toward the end of the TR, while a "small delay slice" refers to a slice acquired toward the beginning of the TR. This classification obviously depends on which slice is used for the reference slice.

Assuming all subjects are scanned with identical scan parameters, the easiest way to evaluate STC is to identify one high-delay slice in native space that intersects a region of activation in each subject. Higher-level cognitive tasks and contrast maps can sometimes result in only small regions of activation which vary spatially from subject to subject, making these kinds of studies inappropriate for investigating STC. Instead, a task with a well-known, robust, and spatially large activation should be used to ensure that every subject has some activation on a chosen high-delay slice. In our evaluation, we generated simulated data with the same subject brain morphology to control for brain shape differences. For real data, we used only the voxels in subjects' native space that are located in slices with moderate offset delay and have significant activation for an attended flashing checkerboard visual sensory task.

Both simulated and real data underwent various preprocessing pipelines consisting of the following modules in different orders: rigid body spatial realignment (motion correction or MC) was applied with FSL (*mcflirt,* Jenkinson et al., 2002) using rigid-body registration of all volumes to the middle one. STC (temporal realignment) was performed using our in-house method FS, as well as the FSL and SPM default techniques. All STC methods temporally aligned the data to the first slice, acquired at the beginning of the TR. FS is described in our previous publication and summarized here (Parker et al., 2016). Our method relies on the Nyquist–Shannon sampling theorem, which states that a signal sampled at twice its Nyquist frequency can be optimally reconstructed by zero-padding and low pass filtering. Using this, as well as an understanding of digital filtering and the properties of our BOLD signal, we optimized the STC algorithm by upsampling the data, moderately increasing the filter order and addressing initialization artifacts with a circular padding scheme. Finally, an optimal window function is selected for our low pass filter (LPF) – a Kaiser windowed sinc, for its tradeoff between a stable pass band, steep transition width, and suppression of the stop band. The cutoff frequency is adjusted to remove high-frequency noise above 0.21 Hz, which is the highest frequency present in the canonical HRF. The upsampled data is then resampled at a desired offset, to correct for the slice timing artifact. As we have shown in our previous work (Parker et al., 2016), every interpolation scheme can be represented and implemented as low-pass filtering with a specific kernel. This might raise the question, why the proposed FS method (which is essentially another low-pass filter) outperforms the interpolation-based methods that were being used in FSL and SPM. While this has not been discussed extensively in the fMRI field, we need to emphasize that the filter type, order, cut-off frequency, window type, transition to pass-band ratio, zero-padding scheme, and most importantly its implementation are all extremely important factors to be considered in any digital filter design, and could substantially affect the performance of the filter. In that sense, one might consider FS theoretically as a modified implementation of the low-pass filter that is often used in fMRI pre-processing pipeline.

We also created a "gold standard" method to compare STC techniques by constructing a slice-dependent shifted regressor for each slice. These shifted regressors account for the slice dependent acquisition offset delay. This method only applied a LPF to the data, as most STC techniques inherently have a small amount of low pass filtering. For a reliable comparison, the same filter parameters used in our FS method were used for the Shifted Regressor (SR) technique. In theory, this method should produce the best results, if no 3D processing algorithm (e.g., 3D smoothing, or 3D spatial transformation) has been applied in the pre-processing, as those methods combine data from different slices with different offset delays, which would alter the signals of each voxel.

In order to focus purely on the effect of different STC methods, only the specified preprocessing steps are applied to the data. We developed a standard generalized linear model (GLM) in Python and used it to model observed fMRI data Y at each voxel as a linear combination of regressors X which were created by convolving the double gamma HRF with the stimulus timing function. We used a standard GLM model shown below:

$$Y = X \cdot \beta + e$$

where β coefficients were obtained using the ordinary least square estimate and given by,

$$\beta = (X^T X)^{-1} X^T Y$$

Standard GLM statistical inference was performed to obtain the t-statistics and significance level of activation for each voxel independently. In our previous publication, we compared the t-statistics from the GLM directly. In this paper, we compare the standardized beta value, described as:

$$\beta_s = \beta \frac{\sigma_x}{\sigma_y}$$

Where σ_x is the standard deviation of the regressor X, σ_y is the standard deviation of the time series Y. If there is only one regressor in the matrix X, then this value is mathematically equivalent to the *Pearson* correlation coefficient of the regressor with the time series. The correlation provides a more intuitive measure that is normalized, and there for more comparable across different analyses. We assume that a pipeline is better if it is able to construct a time series that is more correlated to the regressor. While there are many methods for fMRI quality assessment, the *Pearson* correlation coefficient is a good choice for extracting the relationship between two variables (Tegeler et al., 1999; Miller et al., 2002; Wagner et al., 2005; Caceres et al., 2009). We then convert the correlation to a z score using Fisher's-z transform for statistical comparison. We refer to this as the voxel's z-score, and it is used as the basis of comparison throughout the paper.

Voxel Selection

Only voxels in a slice with high acquisition delay in either real or simulated data were considered for this selection. For real data, we selected slice 17, (1.46 s delay from the beginning of the TR), and for simulated data we selected slice 18 (1.78 s delay from the beginning of the TR). Previous studies have demonstrated that the benefit of STC is directly proportional to the temporal offset from the selected reference slice, and have shown that parameter estimates have reduced variance from slice to slice, and remain unbiased after STC (Sladky et al., 2011). In a previous publication, we have shown that the difference between STC and uncorrected data is minimal in slices with low delay (Parker et al., 2016). This paper is primarily concerned with the effects of STC on fMRI data in different preprocessing pipelines. In order to evaluate this, we must examine voxels that are affected by the STC process. Just as the effects of MC can only be seen on volumes with motion present, the effects of STC can only be seen on slices with temporal delay. So the evaluation of the STC methods and their interaction with other processing steps would be meaningless if it is done on slices with minimal or no time delay, just as evaluation motion correction technique on scans without motion would give invalid results. Because of this, we focused only on high delay slices to examine the effect of STC and its interaction

with preprocessing steps where it's most detectable, rather than comparing the effect of STC across all slices.

We selected the left superior frontal (LSF) region as our region of interest (ROI) for simulated data since the superior frontal region is one of the few regions that spans over 20 slices, thus including slices with various acquisition delays. Therefore, we can guarantee that some part of the region will fall in a high-delay slice regardless of motion and subject morphology. It also spans from the center of the brain to the frontal region, which will capture many different kinds of motion artifacts, as the same motion may produce different, even opposite signal changes in different regions of the brain (Power et al., 2012). Only voxels in the LSF ROI were used in all simulated data analysis, and the known underlying BOLD signal assigned to the LSF ROI was used as our regressor.

For real data, we created an ROI for each subject that included voxels where we expected significant visual activation. This ROI was created by transferring a group level activation mask for the visual stimulus back into each subject's native space. This group level activation mask was created by running a full default FSL first level analysis, which consists of the following steps: (a) spatial realignment, (b) STC, (c) 3D smoothing with FWHM = 5 mm, (d) intensity normalization (e) temporal filtering (125 s cutoff) (f) GLM with prewhitening, and a second level analysis including: (a) spatial normalization, (b) full Bayesian linear model (Beckmann et al., 2003; Woolrich et al., 2004), and (c) cluster-wise multiple comparison correction (z threshold 2.3, cluster significance threshold $p = 0.05$). Only voxels in the activated regions were used in our real data analysis. Region masks for real and simulated data can be seen in **Figure 1**. The large ROI from the group level is only used to confine voxel selection to an area that is neurophysiologically task-relevant, to avoid false positives. For instance, we want to prevent selecting a false-positive voxel in the motor cortex for visual stimulation. One could just replace these ROIs with an anatomical mask of the visual cortex without having a significant alteration on the results. For real data, the average size of this region across all subject's native space was 8704.3 voxels (104.6 cm^3), with a standard deviation of 974.6 voxels. When confined to a slice of high delay, the average size of this mask was 514.0 voxels (6.2 cm^3), with a standard deviation of 83.4 voxels. Despite the fact that this ROI was identified using a pipeline that performs MC before STC, the spatial smoothing used in the group level analysis makes the resulting mask large enough so that it will not bias the results to favor MC before STC pipelines. The large ROI is necessary, as individual differences in head position and brain function do not guarantee that all subjects will have significant activity in a single anatomical region if it's too small. While it's possible that this ROI may cover different anatomical regions over different subjects, the purpose of the group level analysis was to identify regions functionally and neuro-physiologically associated with the task. Since the function of the brain is what is associated with the BOLD signal, this allows us to identify the regions that are most likely to have the same underlying BOLD signal. This region spans over multiple slices, giving us a good chance at identifying significant voxels within the functional region, on a slice of high temporal delay.

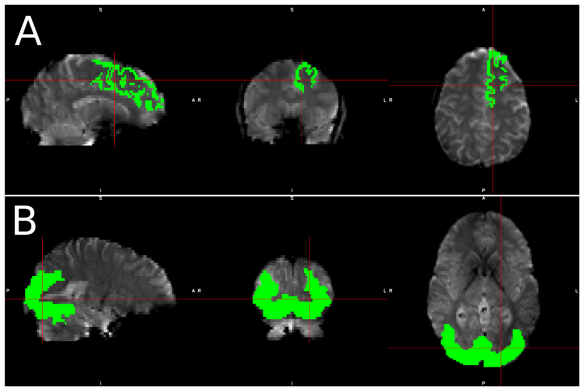


FIGURE 1 | Visualization of spatial ROI's used in analysis. Spatial maps for ROI's used in **(A)** simulated data, which was identical for each simulated subject, and **(B)** Real data, in which the ROI varied slightly from subject to subject due to individual subject morphometry.

Voxel selection was done exclusively from within the generated ROI mask and selected slice with high temporal acquisition delay. We used the SR technique to obtain the significantly activated voxels within the ROI without applying any spatial smoothing. In addition, the SR method does not require an extra STC step, thus making the choice of STC before or after re-alignment irrelevant. For any given subject, 20 voxels that had the highest t-statistics from the SR pipeline were identified. The z-score of the regressor to the time series of these 20 voxels were used to compare all STC methods and scan parameters. We then use a one tail pair-wise t-test on these values to compare the effectiveness of each pipeline. While we are using 30 fMRI scans from real and 20 from simulated data, each scan provided us with 20 voxels, each of which provides a unique fMRI measurement from different location in the brain. In other words, at each voxel location, we can establish an independent but similar comparison between the STC methods. Because of this, we consider each voxel a degree of freedom, giving us 199 DoF for real data (10 subjects for each motion level), and 399 for simulated data. The SR method theoretically does the best job identifying the regions that truly match the expected signal. By using the voxels with the highest statistics from the gold standard method, we identify the 20 voxels we are most certain are true positives, thus minimizing the possibility of type I error confounding our analysis, and allowing us to focus mostly on the effects of interpolation error from STC. It is important to note that these are not necessarily the 20 highest voxels for the other STC methods. Furthermore, from subject to subject, the location of these voxels is not constrained, as not every subject will exhibit the strongest response in the same location. Within each subject, we are comparing the same group of voxels across all different STC methods, so it is appropriate to use a repeated measures t-test to see if the preprocessing pipeline had any significant effect on the GLM, as done previously in the literature (Sladky et al., 2011). Our processing pipeline is illustrated in **Figure 2**.

The low number of voxels used in the comparison was chosen because certain real subjects only had 20 significant voxels

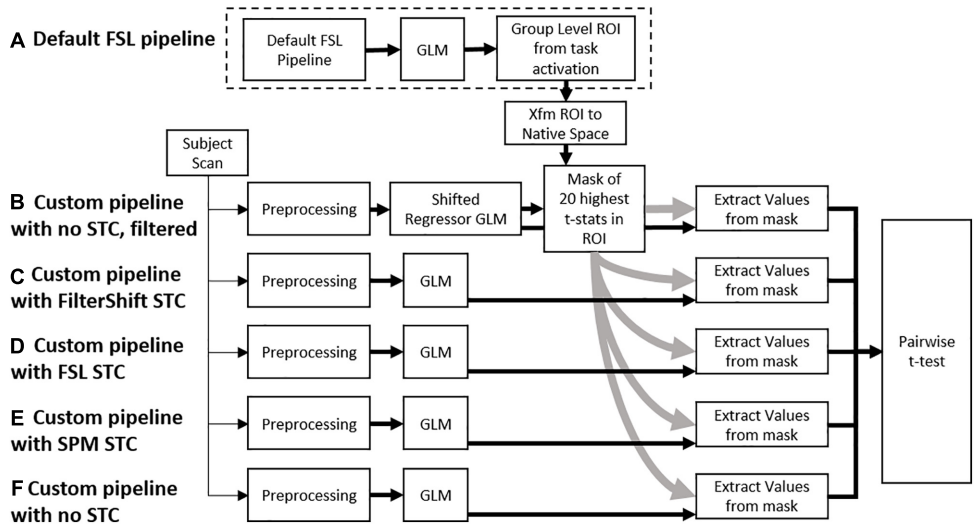


FIGURE 2 | A flowchart of the processing pipeline for real data. **(A)** A default group level GLM is run on all subjects to obtain a region of activation, which is binarized into a mask and transformed back into each subject's native space. This mask is used as an ROI to examine *t*-statistics from five different processing pipelines: **(B)** No STC, but the data is temporally lowpass filtered, and slice-dependent regressors shifted to account for the acquisition offset are used. **(C)** Our in house FS STC method is used. **(D)** FSL STC is used. **(E)** SPM STC is used. **(F)** No STC is used, and regressors are not shifted to account for slice offset. The statistics from pipeline b are used to identify 20 voxels with the highest t-statistics in the ROI. The standardized beta values from these 20 voxels are then extracted from the other pipelines, and the values are compared in a pairwise *t*-test.

present both in the region of interest, and also in the high-delay slice. The reason this number is low is because we do not perform spatial smoothing, which greatly increases the number of significant voxels, at the cost of a lower maximum *t*-value. In order to perform a repeated measures *t*-test, we need to compare values from the same voxels across multiple datasets. To ensure that the voxels we were comparing were in fact statistically significant in every comparison, we use the highest 20 voxels.

Pipeline Order Without Smoothing

We created four different preprocessing pipelines to study the interaction of motion and the order of preprocessing steps. For simulated data, each slice acquisition order and motion level was processed with the following pipelines: (1) STC after motion correction ("After MC"). (2) STC before motion correction ("Before MC"). (3) STC only. (4) Motion correction only. (5) No correction. In this experiment, motion correction was carried out with FSL's *mcflirt* spatial realignment tool using rigid body registration to the middle volume. For real data, subjects were divided into low, medium, and high motion groups, and the same five pipelines were run on each subject. For both simulated and real data, the top 20 voxels selected for statistical analysis in the respective ROI were identified using the "unsmoothed/SR" pipeline as the reference image. The LPF for the SR technique was applied before MC. For simulated data, we also extracted the HRF from these voxels using basic matrix deconvolution and compared it to the true HRF used to generate the simulation. We also computed the dice overlap of the location of each method's top 20 voxels, compared to the top 20 voxels found using the SR method.

Prewhitening is an important step in the preprocessing pipeline. To assess the interaction between prewhitening and STC, we re-ran the real subject pipeline-order analysis with AFNI's prewhitening routine, which is a voxelwise auto-regressive-moving-average ARMA(1,1) model, and compared the results to the non-prewhitened data. The interaction with STC has already been studied, and was found to be limited (Olszowy et al., 2019). To verify these results, we assessed this interaction with STC, both with and without MC.

We also use a common reliability measure to see if STC, and pipeline order, have any effect on the reliability of the data. For this experiment, we use the real, low-motion subjects to minimize confounds and focus on the effect of STC on reliability. We split each subject's time series into two halves, and recomputed the GLM on each half with the appropriate segment of the regressor. A brain mask, created using FSL's BET, is used to extract the beta values from the brain for each half of the time series. The correlation between the beta values from each half is then calculated. High correlation indicates good similarity between the analyses run on each half of the scan.

Motion Parameter Residualization

We also examined the effect of pipeline order while including realignment parameter estimates as nuisance regressors, also known as MPR. Six motion parameters are estimated during the rigid body realignment stage described earlier, and included as nuisance regressors along with their temporal derivatives (rotation and translation for the x, y, and z axis) (Power et al., 2014). These parameters were acquired from FSL's *mcflirt* spatial realignment tool, using rigid body registration to the middle volume. In cases with "No MC," motion parameters were

estimated after STC, but no realignment was done. Depending on the level of motion, more accurate motion parameters may be extracted before or after STC. This was run on a subset of 5 low, medium, and high motion subjects for both real and simulated data. As in the previous section, we used the "unsmoothed/SR" pipeline as the reference image for statistical analysis.

Spatial Smoothing on STC

Slice timing correction is particularly sensitive to any sort of image processing/analysis that spans over multiple slices, the most common of such processes is 3D spatial smoothing. Smoothing is generally done to enhance the robustness of the statistical analysis results and reduce false positive rates (Mikl et al., 2008). Our goal in this experiment is to investigate the effect of 3D smoothing on the effectiveness of STC. Four different Gaussian non-linear smoothing kernels with FWHM = 3.5, 5, 6.5, and 8 *mm* were used to smooth simulated and real data using FSL's smoothing method. Smoothing was applied by spatially convolving each 3D volume in the time-series with the appropriate 3D Gaussian kernel. For real data, the effect of smoothing was examined on high, medium, and low motion scans. For simulated data, low motion scans were used, and the effect of smoothing was examined for sequential, interleave 2, and interleave 6 acquisition methods. Each fMRI scan was first motion corrected and slice timing corrected before the various smoothing kernels were applied. We compared the performance of each STC method across the different smoothing kernels, where higher z-scores are considered better performance. For each motion/interleave case for real/simulated data respectively, the corresponding "unsmoothed/SR" image was used as the reference image for voxel selection.

STC on Short TRs

To examine the benefit of STC on real data with high temporal resolution, we modified the HCP volumetric preprocessing pipeline described in Glasser et al. (2013) to output the results in native space, and removed any spatial smoothing to avoid the complications involved in assessing STC benefit, as described above. We refer to this custom pipeline as the "modified HCP pipeline." We selected the N-back working memory task, described elsewhere (Barch et al., 2013), as our experimental data. As in the previous experiments, a default FSL first and second level analysis was run to obtain a group level activation mask of an event-related regressor modeling activation from all 2-back correct events. A python-based GLM was then run on each subject's native space data with (a) the modified HCP pipeline with the gold standard SR, (b) the modified HCP pipeline with added FS STC, and (c) the modified HCP pipeline with no STC. We used the gold standard SR results as our reference image for statistical analysis.

Effect of STC on Functional Connectivity Analysis

To examine the effect of STC on the extraction of functionally connected regions, we ran FSL's MELODIC ICA analysis tool. MELODIC uses probabilistic ICA to extract spatially independent components along with their time series. MELODIC attempts to extract an optimal number of spatial/temporal components, so that when recombined they approximate the original data as closely as possible without overfitting (Smith et al., 2004).

For simulated data, we used low motion subjects with no added noise to attempt to minimize any confounds present in the analysis. We select data with a TR of 2 and interleave 6, which matches the parameters of our real data, for a more accurate comparison. Only STC was performed on the data, and no additional processing was done prior to running MELODIC ICA extraction. We manually identified a single IC which has the maximum spatial overlap with the task-related co-activation pattern found with the GLM in Section "Pipeline Order Without Smoothing," and ran FSL's dual regression to assess the statistical fit of this IC's time series on the voxels present in the spatial IC (Beckmann et al., 2009). This provided us with z-statistics describing the significance of the IC's time series for any given voxel. For this experiment, the z-statistics come directly from the statistical parametric map, rather than the z-score of a correlation as in the other analyses. Because there is no gold standard for ICA, we cannot use the same methods for comparing 20 voxels as described before. Instead, we simply extract significant z statistics generated from the dual regression that fall within a given ROI from both high and low delay slices. For real data, this ROI is the group level task activation map described earlier, and for simulated data this ROI is the LSF anatomical region where the signal was simulated in. We then examined the z-statistics within the mask separately for each slice, and together as a whole. We hypothesize that the uncorrected data will have lower z-statistics in high-delay slices, while STC data will have uniform z-statistic values over slices of all delays.

RESULTS

Pipeline Order Without Spatial Smoothing

Table 1 shows the mean z-scores extracted from the spatial realignment/STC order experiment for simulated data described in Section "Voxel Selection." These values are plotted in **Supplementary Figure S1**. The voxels with the top 20 t-statistics from the SR method are identified, and the z-scores are compared across different motion levels, interleave types, and STC methods. Within each method, the z-scores are extracted from three different pipelines: STC "before MC," STC "after MC," and only STC ("no MC"). The "Uncorrected" condition refers to scans with no STC, and only "MC" or "No MC," where "No MC" refers to an unprocessed scan.

From this table, we see that the results vary based on the level of motion, the slice acquisition orders, and the methods of STC. Significant differences are indicated with a "$*$" for $p < 0.01$, and a "$\ddagger$" for $p < 0.001$. All t-tests were calculated with respect to the STC "Before MC" case within each method/motion/interleave block, which is highlighted in gray, and the pipeline with the highest average z-score value is bolded within each STC/Interleave/Motion category. For FS, STC "After

TABLE 1 | Effect of the order in which STC and MC are applied in the preprocessing pipeline in simulated data.

	Sequential			**Interleave 2**			**Interleave 6**		
	Low motion	***Medium motion***	***High motion***	***Low motion***	***Medium motion***	***High motion***	***Low motion***	***Medium motion***	***High motion***
FilterShift									
Before MC	1.50 ± 0.24	0.62 ± 0.15	0.39 ± 0.14	1.50 ± 0.24	0.62 ± 0.15	0.39 ± 0.15	1.51 ± 0.24	0.62 ± 0.15	0.39 ± 0.14
After MC	**1.57 ± 0.27**‡	**0.64 ± 0.19**	**0.61 ± 0.17**‡	**1.56 ± 0.26**‡	**0.63 ± 0.18**	**0.56 ± 0.16**‡	**1.55 ± 0.26**‡	**0.63 ± 0.19**	**0.56 ± 0.17**‡
No MC	1.20 ± 0.26‡	0.38 ± 0.17‡	0.15 ± 0.14	1.20 ± 0.26‡	0.38 ± 0.17‡	0.16 ± 0.14‡	1.21 ± 0.26‡	0.38 ± 0.17‡	0.16 ± 0.13‡
FSL									
Before MC	1.52 ± 0.24	0.64 ± 0.19	0.57 ± 0.20	1.23 ± 0.14	0.60 ± 0.16	0.55 ± 0.18	**1.06 ± 0.11**	**0.59 ± 0.15**	**0.54 ± 0.17**
After MC	1.52 ± 0.24	0.64 ± 0.18	**0.60 ± 0.17**	**1.24 ± 0.15***	0.60 ± 0.16	**0.54 ± 0.15**	1.04 ± 0.11‡	0.58 ± 0.14	0.51 ± 0.14
No MC	1.18 ± 0.24‡	0.38 ± 0.17‡	0.15 ± 0.14‡	1.03 ± 0.17‡	0.37 ± 0.16‡	0.16 ± 0.14‡	0.92 ± 0.14‡	0.36 ± 0.15‡	0.16 ± 0.13‡
SPM									
Before MC	**1.18 ± 0.13**	0.60 ± 0.16	0.53 ± 0.18	1.11 ± 0.12	0.58 ± 0.15	**0.54 ± 0.17**	**1.05 ± 0.11**	**0.58 ± 0.15**	**0.53 ± 0.17**
After MC	1.17 ± 0.13	**0.60 ± 0.15**	**0.56 ± 0.15**	1.11 ± 0.12	0.58 ± 0.14	0.52 ± 0.14	1.03 ± 0.11‡	0.57 ± 0.14*	0.51 ± 0.14
No MC	0.98 ± 0.16‡	0.36 ± 0.16‡	0.14 ± 0.14‡	0.95 ± 0.15‡	0.36 ± 0.15‡	0.15 ± 0.13‡	0.91 ± 0.13‡	0.36 ± 0.15‡	0.16 ± 0.13‡
Uncorrected									
MC	**0.95 ± 0.08**	**0.54 ± 0.12**	**0.52 ± 0.13**	**0.85 ± 0.08**	**0.51 ± 0.12**	**0.50 ± 0.12**	**0.77 ± 0.07**	**0.47 ± 0.11**	**0.47 ± 0.12**
No MC	0.84 ± 0.11‡	0.34 ± 0.14‡	0.14 ± 0.14‡	0.76 ± 0.10‡	0.32 ± 0.13‡	0.15 ± 0.13‡	0.70 ± 0.09‡	0.31 ± 0.12‡	0.15 ± 0.12‡

Mean z-score of top 20 voxels identified in the Shifted Regressor "Before MC" case, extracted across various preprocessing pipelines for five simulated subjects. For the Shifted Regressor data, "Before MC" refers to when the low pass filter was applied to the data in the preprocessing pipeline. Pairwise t-tests were used to determine significant differences between pipelines for a given STC technique, relative to the reference case highlighted in gray (STC before MC). Significant differences are indicated with a "" for $p < 0.01$, and a "‡" $p < 0.001$. Bold numbers indicate the highest mean z-score of the three pipelines for a given interleave, motion level, and STC technique. Highlighted rows indicate the data used as reference in the t-test.*

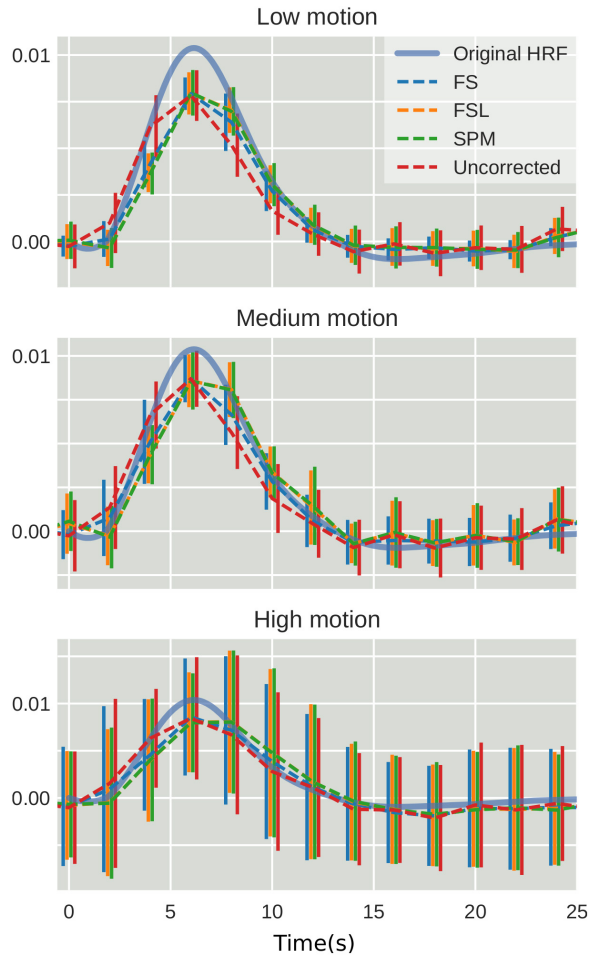


FIGURE 3 | The effect of STC on the accuracy of BOLD response. Normalized HRF's extracted from the top 20 voxels identified using the shifted regressor method for low, medium, and high motion in simulated data. HRF's were extracted using FIR deconvolution for each STC method (dashed lines). The reference HRF used in the simulation is plotted as a solid blue line. For each motion level and STC method, HRF's are combined across all pipeline orders.

MC" significantly outperformed "before MC" across all interleave values for both low and high motion ($p < 0.001$), but not for medium motion, where there were no significant differences for the order of STC and MC. In all cases, both "After MC" and "before MC" outperformed "No MC."

For FSL, pipeline order had no effect in sequential data. For interleave 2, "After MC" is significantly higher than "before MC" for low motion data ($p < 0.01$), but not for medium and high motion subjects. For interleave 6, we now see that "before MC" performs significantly better than "After MC" for low motion subjects ($p < 0.001$). For medium and high motion subjects, "before MC" does have a higher mean z-score than "After MC," but it is not statistically significant ($p > 0.01$).

For the SPM corrected data, we see many similar trends. Pipeline order is largely irrelevant for both sequential and interleave 2 acquired data, with no significant differences between "After MC' and "Before MC." For interleave 6, we see that "Before MC" performs significantly better than "After MC" for low motion ($p < 0.001$), and for medium motion ($p < 0.01$), but not for high motion. Finally, for uncorrected data, "MC" always performed significantly better than "No MC" for all data sets ($p < 0.001$).

Interestingly, the effects of STC and spatial realignment seem to add non-linearly. For example, in the medium motion case with interleave 6, the application of MC to the data without STC raises the average z-score from 0.31 to 0.47 (52% increase). The addition of FSL's STC without MC raises the z-score from 0.31 to 0.36 (16% increase). However, the addition of both MC and FSL's STC raises the z-score from 0.31 to 0.58 (87% increase). This is greater than the summation of just MC or just STC. This highlights the importance of applying both STC and MC to fMRI data in the pre-processing stream.

To examine the effect of STC on estimated BOLD response parameters, the HRF was extracted using FIR deconvolution (Glover, 1999) from the 20 voxels identified in the previous analysis for simulated data, using the "STC after MC" pipeline, as it performed best for simulated data. A more accurate HRF indicates that the time series better matches the expected stimulus response. We computed the sum of squared error (SSE) between the extracted HRF and the true underlying HRF, both normalized. The HRF was normalized because each voxel was assigned a BOLD signal that was 4% of the mean intensity of that voxel. Thus, for every voxel, the magnitude of the extracted HRF should be different, meaning that absolute error does not fully describe the goodness of fit, as the size of the original signal must be accounted for. The extracted HRFs can be seen in **Figure 3**. For each STC method, the HRFs from all pipelines are included in the plots. It can be seen that the error of the estimates increases as motion increases. The SSE are visualized separately using a box-plot in **Figure 4** for each of the STC techniques. For low motion, FS has a mean SSE of 0.15, while FSL, SPM, and uncorrected data have a mean of 0.24, 0.29, and 0.39 respectively. The error increases for medium and high motion, reaching a mean of 0.62 for FS, 0.67 for FSL, 0.72 for SPM, and 0.73 for uncorrected data in the high motion case. As shown, increasing the level of head motion increases the SSE and reduces the difference between the utilized STC techniques. Note, the line in the boxplot indicates the median.

Figure 5 shows a violin plot of real data and the effect of STC/MC order. For low motion, the only significant difference was found with the FS STC method, where STC after MC created significantly lower z-scores than STC before MC ($p < 0.001$), and STC only ($p < 0.01$). For all other STC methods in low motion subjects, there were no significant differences between any other pipelines, even for uncorrected data, which compares MC in the absence of STC. For medium motion, pipeline order had a significant impact for all STC techniques. For FS, FSL, and SPM, STC and MC together in any order performed significantly better than No MC ($p < 0.001$), and STC before MC performed worse than STC after MC ($p < 0.01$ for FS and SPM, $p < 0.001$ for FSL). In the medium motion case, MC significantly improved uncorrected data ($p < 0.001$). For high motion subjects, there is again no significant difference between pipelines applying

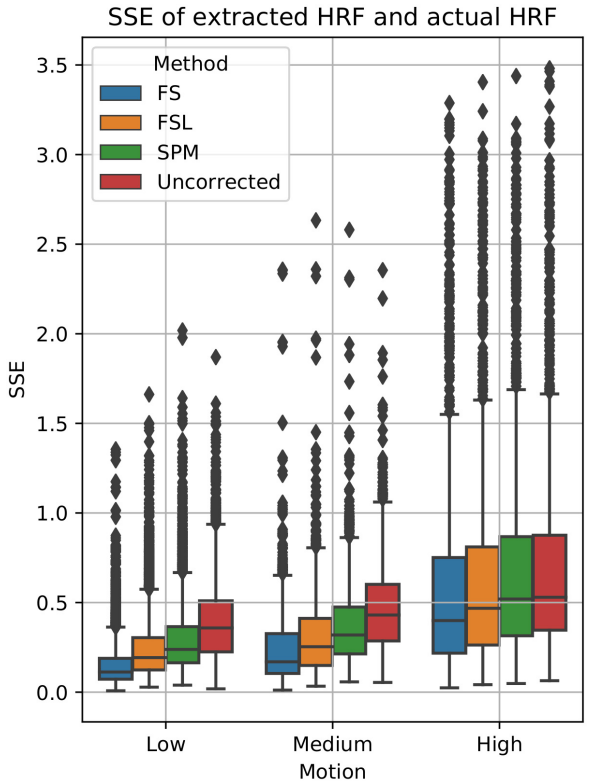


FIGURE 4 | Quantification of BOLD HRF extraction error. SSE of normalized HRF's deconvolved from STC data, compared to the canonical HRF used in the simulation for low, medium, and high motion. For each motion level and STC method, HRF SSE's are combined across all pipeline orders.

STC before or after MC. However, the presence of MC again improves z-scores significantly across all methods ($p < 0.001$). For uncorrected data, MC again proved significantly beneficial ($p < 0.01$). Generally, for subjects with high and low motion, the order of the preprocessing pipeline seem to have no effect, while for the medium motion cases, MC before STC shows small but statistically significant improvement over STC before MC for FSL and SPM. Real data also highlights the importance of performing both MC and STC. For medium and high motion subjects, STC and MC performed together in *any* order ("Before MC" and "After MC" columns) always results in higher statistics than just STC alone ("No MC" column, $p < 0.001$).

In order to further probe these results, we calculated the dice overlap of the top 20 voxels from each method in the high delay slice, compared to the top 20 voxels identified using the SR method. This was done for real and simulated data, and for each different pre-processing pipeline. **Figure 6** shows the resulting Dice overlap coefficient of simulated data, and **Figure 7** shows the same for real data. As shown, applying MC after STC (the "Before MC" column) typically lowers the dice overlap of the highest voxels relative to those of the "gold standard" method, which worsens as motion increases. In real data, we see a similar trend, with slightly lower dice overlap when MC is applied after STC. FS method has a lower dice overlap than other STC methods in real data for medium and high motion cases when MC is not applied.

To compare the interaction of STC and PW, we examine the z-scores of the pipeline with STC only, with and without PW. For low motion subjects, prewhitening did not have a significant effect on the z-scores for STC data ($p > 0.05$). The t-statistic maps from the prewhitened data had a mean correlation coefficient of 0.90 with the t-statistics maps from the non-prewhitened data across all subjects, while beta maps had a mean correlation coefficient of 0.83. All z-score averages from the prewhitened data were within 0.04 units of the corresponding value in the non-prewhitened data. This agrees with the literature in the field, that has found only small interactions between STC and prewhitening (Olszowy et al., 2019). Increasing the motion level, or including MC into the pipeline introduced more differences in z-scores between prewhitened and non-prewhitened data, indicating a possible interaction with prewhitening and motion. This relationship is beyond the scope of this manuscript, and can be addressed in future work.

To examine the effect of STC on reliability, the correlations of the low motion subjects' beta maps are shown in **Supplementary Figure S2**. When examining the reliability, we found no significant differences between pipeline orders within each STC method with a two sample t-test ($p > 0.1$). We also ran a repeated measures t-test on the uncorrected data, and the STC only data for each method, which also showed no significant differences ($p > 0.1$). This shows that STC does not lower the reliability of the fMRI data.

Motion Parameter Residualization

Residualizing motion parameters affected the results from the section "Pipeline Order Without Spatial Smoothing," as shown in **Table 2**. **Table 2** shows the summarized statistics for simulated data, and these results are plotted in **Supplementary Figure S3**. It is clear from **Table 2** that the order of STC and MC are greatly affected by MPR. For all STC methods, across all levels of motion and all interleave methods, STC after MC provided the best z-scores ($p < 0.001$). For uncorrected data, MC only proved significantly better than no MC for low motion sequential data ($p < 0.01$), and high motion data for all interleaves ($p < 0.01$).

For real data, the effect of including motion parameters was small, and seemed to change based on motion level. These values are plotted in **Supplementary Figure S4**. For low and medium motion, MPR reduced the average z-scores across all methods by 0.03 and 0.04 respectively. For high motion, the average z-scores decreased by only 0.003. For low motion, the effect of combining STC and MC is enhanced, as STC "Before MC" is now significantly better than "No MC" for FS, FSL, and SPM ($p < 0.001$). For SPM, STC "After MC" also became significantly different from no MC ($p < 0.01$). For medium motion, the significant differences in pipeline order disappear for FSL and SPM. For high motion, the same significant differences present in the data without MPR were also found in this data set. Additionally, for FS, the pipeline order now significantly effected z-scores, with STC after MC resulting in higher z-scores ($p < 0.01$).

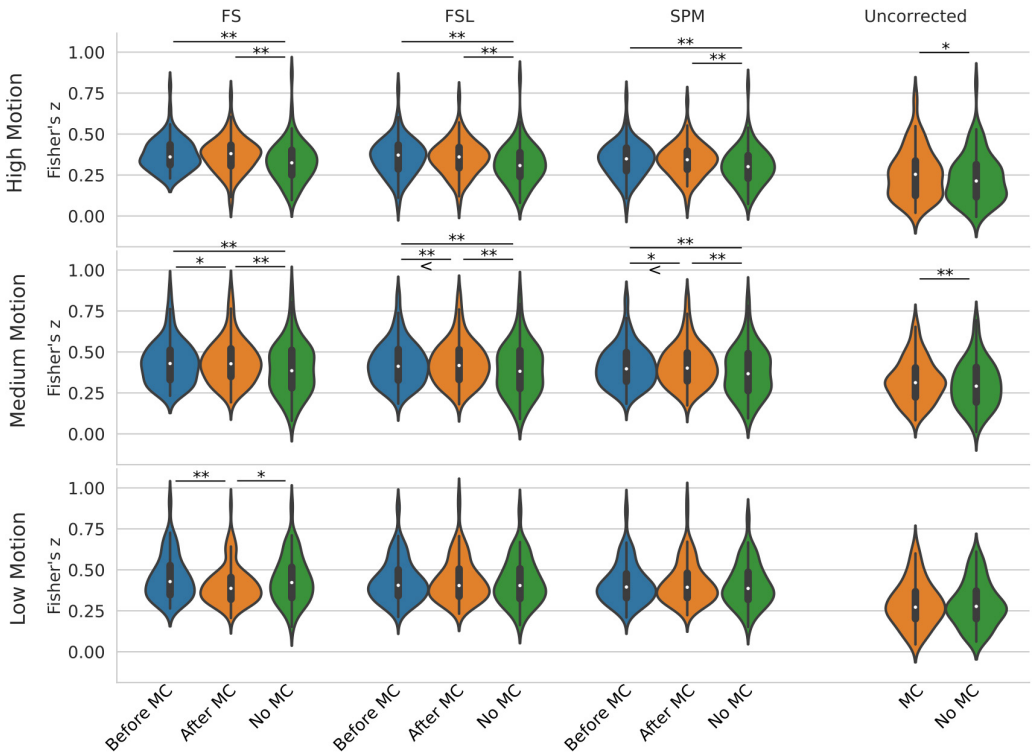


FIGURE 5 | Effect of the order in which STC and MC are applied in the preprocessing pipeline in real data. *z*-scores from the top 20 voxels from a visual ROI of 30 real subjects' time series using different preprocessing pipelines, with no smoothing. FS, FSL, and SPM slice timing correction were applied before motion correction, after motion correction, or without any motion correction. For the "Uncorrected" case, only motion correction was applied in the preprocessing pipeline. '<' and '>' Symbols indicate which mean is greater for cases that are not easily distinguishable.

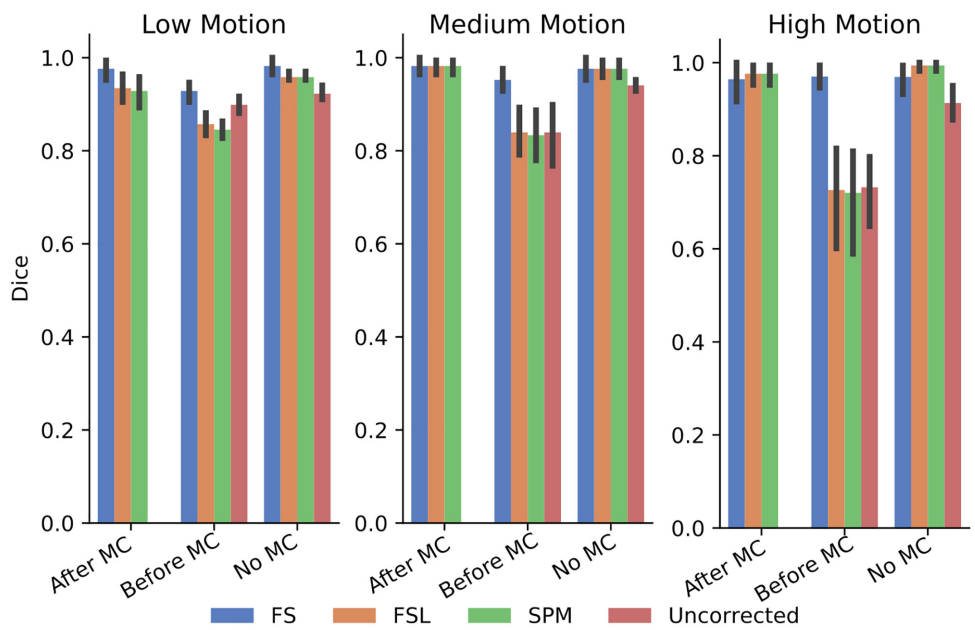


FIGURE 6 | The similarity of the selected voxels in our evaluations of simulated data. Dice overlap in simulated data of top 20 voxels for each STC in various pipelines, compared to the top 20 voxels of the shifted regressor method within that pipeline. Uncorrected data in the "No MC" column refers to data that has had no MC or STC.

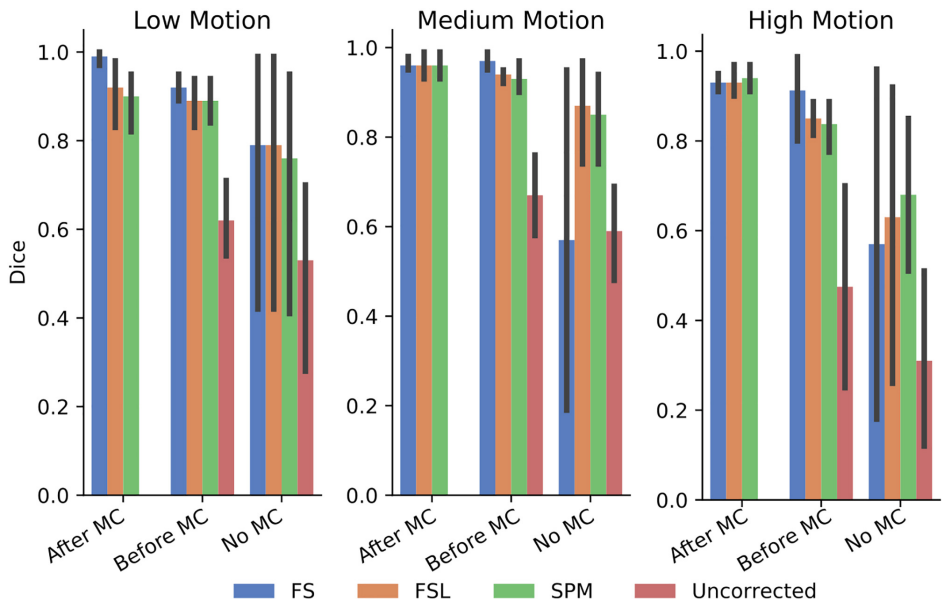


FIGURE 7 | The similarity of the selected voxels in our evaluations of real data, Dice overlap in real data of top 20 voxels for each STC in various pipelines, compared to the top 20 voxels of the shifted regressor method within that pipeline. Uncorrected data in the "No MC" column refers to data that has had no MC or STC.

Spatial Smoothing on STC

We compared the effect of spatial smoothing on STC using both real and simulated data. **Figure 8** shows the results across simulated subjects with low motion for various types of interleave acquisition. For simulated data, increasing the kernel size reduces the mean z-scores monotonically as kernel size increases for all STC data under all interleave scenarios. For all interleave parameters and STC methods, each level of smoothing was significantly different from the following level ($p < 0.001$). Interestingly, mean z-scores from uncorrected data only decreased monotonically for sequential acquisition and interleave 2 acquisition. For uncorrected data with interleave 6, the mean z-score is 19% higher than the unsmoothed case at FWHB, and 6% higher than the unsmoothed case at FWHM of 5 mm. Beyond 5 mm, the smoothed uncorrected data no longer performs better than the unsmoothed data, dropping monotonically as the FWHM increases. It is also important to note the effect of smoothing on the SR method. For interleave 2 and 6, the average z-score for SR was lower than the average z-score for FS in all smoothed datasets. This effect is amplified as kernel size increases, and as interleave parameter increases. In other words, SR method can only be considered as a gold standard technique when there is no spatial smoothing applied, and any smoothing with kernel size larger than 3.5 mm will deteriorate its performance to the level where other methods could outperform it.

For real data, the effect of spatial smoothing on the time series z-scores are shown in **Figure 9**. For low motion, FS, FSL, and SPM do not significantly change from the unsmoothed case compared to any other kernel size. The only significant differences between consecutive smoothing kernels are found between 6 and 8 mm for FS, FSL, and SPM. The SR method, on the other hand, has significantly lower z-scores between smoothing kernels 3.5 and 5 mm ($p < 0.01$). The mean z-scores continue to decrease significantly with kernel size 6.5 and 8 mm ($p < 0.001$). For uncorrected data, we see the same phenomenon from the simulated data, where the smoothed z-scores are significantly larger as the kernel size increases from 3.5 to 5 mm ($p < 0.01$), from 5 to 6.5 mm ($p < 0.001$), and from 6.5 to 8 mm ($p < 0.001$). This increase in z-score of the uncorrected data was expected, since smoothing blends the high delay voxels with low delay ones, which essentially performs similar to linear interpolation. For medium and high motion subjects, we see the same general decreasing trend in z-scores as larger smoothing kernels are applied to STC data. We no longer see the same initial increase in z-score in the uncorrected data. For all STC methods, the unsmoothed data has significantly higher z-scores than the smoothed data for kernel sizes greater than or equal to 5 mm ($p < 0.001$). For the SR method, this difference becomes significant at even just 3.5 mm ($p < 0.01$). For low and medium motion cases, the average z-score for SR data smoothed with an 8 mm kernel is even lower than any other STC method, and in fact even lower than uncorrected data in the medium motion case.

STC on Short TRs

We compared the z-score in the native space generated from HCP data processed with three different pipelines with no spatial smoothing. We have shown previously on simulated data that short TR data benefits greatly from the ideal LPF we use in FS. When performed on HCP data, we found our STC method significantly increased the mean z-score in the top 20 voxels by 13% ($p < 0.001$), shown in **Figure 10**. The SR method was further able to

TABLE 2 | Effect of the order in which STC and MC are applied in the preprocessing pipeline in simulated data, with MPR.

	Sequential			**Interleave 2**			**Interleave 6**		
	Low motion	***Medium motion***	***High motion***	***Low motion***	***Medium motion***	***High motion***	***Low motion***	***Medium motion***	***High motion***
FilterShift									
Before MC	1.00 ± 0.44	0.39 ± 0.25	0.14 ± 0.09	0.94 ± 0.41	0.39 ± 0.26	0.14 ± 0.09	0.78 ± 0.31	0.35 ± 0.22	0.13 ± 0.09
After MC	**1.33 ± 0.47**‡	**0.65 ± 0.37**‡	**0.50 ± 0.18**‡	**1.25 ± 0.47**‡	**0.63 ± 0.36**‡	**0.48 ± 0.17**‡	**0.97 ± 0.30**‡	**0.56 ± 0.28**‡	**0.43 ± 0.15**‡
No MC	1.01 ± 0.44*	0.39 ± 0.25	0.14 ± 0.11	0.95 ± 0.41*	0.39 ± 0.26	0.13 ± 0.11	0.78 ± 0.31	0.36 ± 0.22	0.12 ± 0.10
FSL									
Before MC	1.00 ± 0.43	0.39 ± 0.24	0.15 ± 0.10	0.87 ± 0.34	0.38 ± 0.22	0.14 ± 0.09	0.79 ± 0.29	0.35 ± 0.18	0.14 ± 0.09
After MC	**1.31 ± 0.46**‡	**0.65 ± 0.35**‡	**0.50 ± 0.18**‡	**1.08 ± 0.33**‡	**0.58 ± 0.25**‡	**0.47 ± 0.16**‡	**0.94 ± 0.25**‡	**0.53 ± 0.19**‡	**0.43 ± 0.14**‡
No MC	1.01 ± 0.44	0.39 ± 0.25	0.14 ± 0.11	0.87 ± 0.34	0.38 ± 0.21	0.14 ± 0.11	0.79 ± 0.29	0.36 ± 0.18	0.14 ± 0.10
SPM									
Before MC	0.84 ± 0.32	0.35 ± 0.19	0.14 ± 0.09	0.81 ± 0.30	0.37 ± 0.21	0.14 ± 0.09	0.78 ± 0.29	0.35 ± 0.18	0.14 ± 0.09
After MC	**1.03 ± 0.30**‡	**0.56 ± 0.24**‡	**0.47 ± 0.16**‡	**0.97 ± 0.27**‡	**0.54 ± 0.21**‡	**0.45 ± 0.15**‡	**0.93 ± 0.24**‡	**0.52 ± 0.19**‡	**0.43 ± 0.14**‡
No MC	0.75 ± 0.46*	0.35 ± 0.19	0.14 ± 0.11	0.81 ± 0.30	0.36 ± 0.19	0.14 ± 0.10	0.78 ± 0.28	0.35 ± 0.18	0.13 ± 0.10
Uncorrected									
MC	**0.74 ± 0.27**	0.33 ± 0.16	**0.14 ± 0.09**	0.65 ± 0.23	0.31 ± 0.16	**0.14 ± 0.08**	0.59 ± 0.20	0.29 ± 0.14	**0.13 ± 0.08**
No MC	0.63 ± 0.38‡	0.33 ± 0.16	0.13 ± 0.10	0.65 ± 0.23	0.31 ± 0.16	0.12 ± 0.09*	0.59 ± 0.20	0.29 ± 0.14*	0.11 ± 0.09‡

Mean z-score of top 20 voxels identified in the Shifted Regressor "Before MC" case, extracted across various preprocessing pipelines for five simulated subjects with motion parameters included as nuisance regressors in the GLM. For the Shifted Regressor data, "Before MC" refers to when the low pass filter was applied to the data in the preprocessing pipeline. Pairwise t-tests were used to determine significant differences between pipelines for a given STC technique, relative to the reference case highlighted in gray (STC before MC). Significant differences are indicated with a "" for $p < 0.01$, and a "‡" $p < 0.001$. Bold numbers indicate the highest mean z-score of the three pipelines for a given interleave, motion level, and STC technique. Highlighted rows indicate the data used as reference in the t-test.*

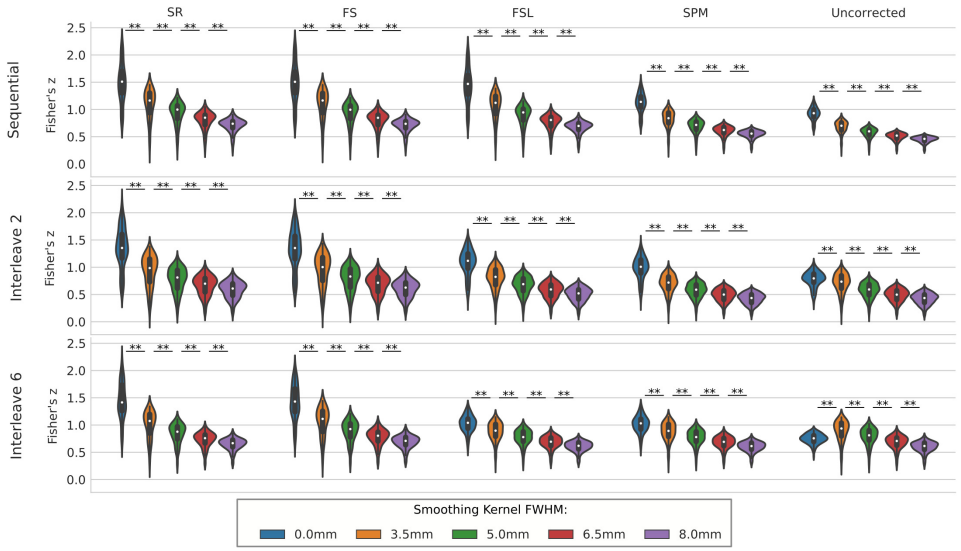


FIGURE 8 | The effect of spatial smoothing on simulated data with STC. Violin plots show the resulting *z*-scores from the voxels with the top 20 t-statistics for 20 simulated data sets processed with smoothing kernels of 0, 3.5, 5, 6.5, and 8 mm. We examined the effect of smoothing for sequential acquisition, even-odd (interleave 2) and every 6th (interleave 6) acquisition (rows). For each interleave/smoothing condition, statistics were compared across 5 STC conditions: Shifted Regressor, FS, FSL, and SPM STC methods were examined, as well as uncorrected data. For each interleave, the voxels with the top 20 t-statistics were identified from the unsmoothed Shifted Regressor case. The *z*-score of these 20 voxels are plotted for all other STC methods and all other smoothing conditions. The effect of smoothing is seen as a lowering of the mean *z*-scores, as well as reducing the variance across STC methods due to the distribution of slice-dependent errors. For interleaved data, the Shifted Regressor method, which is supposed to be a gold standard, performs worse than our proposed FS method due to these distributed errors.

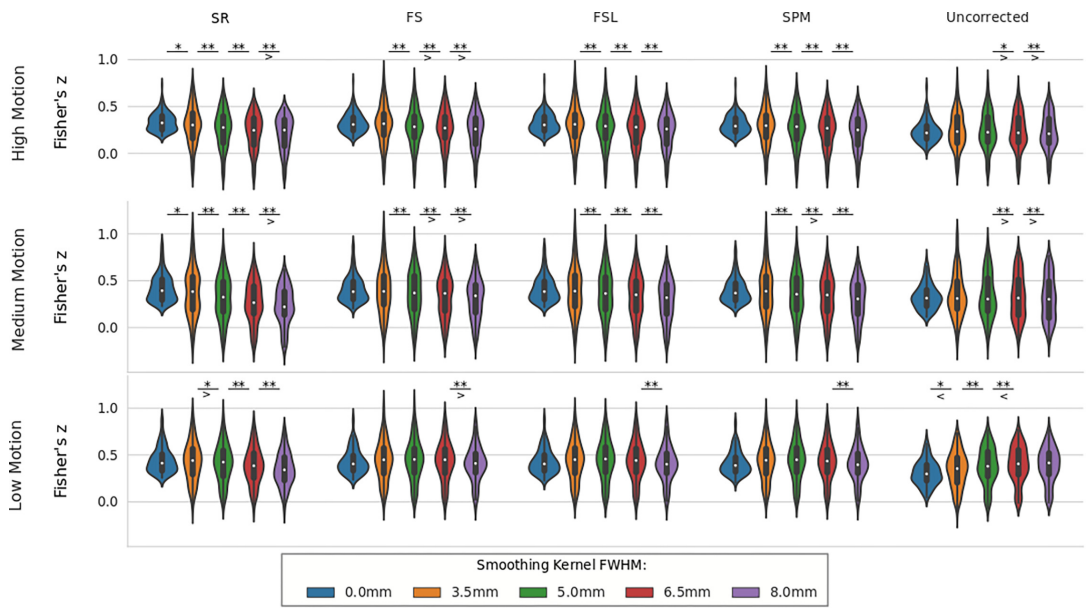


FIGURE 9 | The effect of spatial smoothing on real data with STC. Violin plots show the resulting *z*-scores from the voxels with the top 20 t-statistics for 30 real data sets (10 low motion, 10 medium motion, 10 high motion) processed with smoothing kernels of 0, 3.5, 5, 6.5, and 8 mm. We examined the effect of smoothing for low, medium, and high motion levels (rows). For each motion/smoothing condition, statistics were compared across 5 STC conditions: Shifted Regressor, FS, FSL, and SPM STC methods were examined, as well as uncorrected data. For each motion level, the voxels with the top 20 t-statistics were identified from the unsmoothed Shifted Regressor case. The *z*-score of these 20 voxels are plotted for all other STC methods and all other smoothing conditions. The effect of smoothing is seen as a lowering of the mean *z*-scores, as well as reducing the variance across STC methods due to the distribution of slice-dependent errors. In real data, the "Gold standard" (Shifted Regressor) is outperformed by all other STC methods when large smoothing kernels are used. '<' and '>' Symbols indicate which mean is greater for cases that are not easily distinguishable.

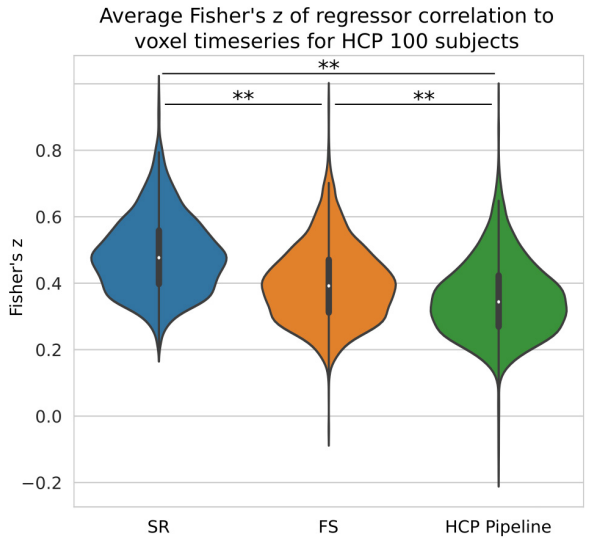


FIGURE 10 | The Benefit of STC on short-TR data. *Z*-scores from voxels with the top 20 t-statistics for 100 HCP unrelated subjects. Violin plots show the combined *z*-scores of all 100 subjects from the 20 voxels with the highest *t*-values in the Shifted Regressor case when processed with the default volumetric HCP pipeline, with our proposed FS method, and with the Shifted Regressor. This shows the clear benefit of applying STC to short TR (0.72 s), multiband data.

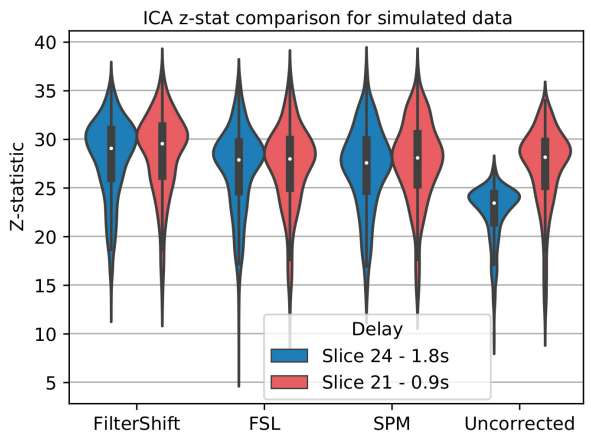


FIGURE 11 | STC on the ability to extract functionally connected regions in simulated data. For each simulated subject, a task IC was extracted with MELODIC, and the IC's time series was regressed onto the subject's data using FSL's dual regression. Z statistics from a low and high delay slice were extracted for every subject and plotted here. For uncorrected data, the extracted IC performs as well as STC data in the low delay slice, but poorly explains the variance in the high delay slice, resulting in a lower z-statistic. For STC data, the average z-statistics are comparable for both high and low delay slices.

increase the *z*-score by 37% from the HCP default pipeline, further highlighting the sensitivity of low TR data to small temporal shifts.

STC on Functional Connectivity Analysis

For simulated data, we extract and plot significant z statistics from a slice with higher temporal delay, and a slice with low temporal delay (slice 20 and 23) for all simulated subjects in **Figure 11**. For the low delay slice, we see that all STC methods perform similarly, and the extracted z-statistics are relatively the same. For the high delay slice, we see that the uncorrected data has lower z-stats than the low delay slice. However, STC improves the statistics for the high delay slice in all cases. This can be seen in **Figure 12**, where ICA dual regression is run on both uncorrected data, shown in A, and STC data shown in B, with the resulting z-statistics shown for six consecutive slices. The IC for the LSF region covers the same spatial region in both analysis, however, after performing dual regression, we see that for uncorrected data, the time series of the IC only strongly predicts the fMRI signal in some of the slices, while for the STC data, the z-statisitcs remain high across all slices. Across all subjects and slices, STC was able to improve the z-statistics of the high delay slice.

For real data, we repeated the analysis done for simulated data, and extracted z-statistics for two slices with low and high delay from all subjects (Slice 15 and 18), and plotted the distributions in **Figure 13**. As seen in this figure, the differences between slices have largely vanished, across all subjects and slices, z-statisitcs increased greatly with FS by 27% compared to FSL, 18% compared to SPM, and 30% compared to uncorrected data.

DISCUSSION

Our previous work has demonstrated the benefit of STC on fMRI data in both real and simulated data which had already been shown in the literature (Henson et al., 1999; Calhoun et al., 2000; Vogt et al., 2009; Sladky et al., 2011). However, because the slice-timing problem is fundamentally intertwined with many confounding factors such as involuntary head motion during scanning, spatial realignment, 3D smoothing, and spatial normalization, it has as yet been difficult to determine whether STC would provide any benefit to the analysis. In the current study, we rigorously examined the impact of these major confounding factors on the effectiveness of STC. We showed that even though the effectiveness of STC changes depending on the preprocessing pipeline, it still remains beneficial to be included as a preprocessing step. While focusing on 20 highly significant voxels does limit our conclusions about STC to only true positives, previous studies have shown that STC does not adversely affect parameter estimate bias, and in fact was shown to suppress bias, so we do not anticipate the need to evaluate STC on voxels outside our ROIs (Sladky et al., 2011). Although we cannot determine from this study if the conclusions reached here generalize over the whole brain, we have isolated the voxels most affected by STC, and so if an effect is not found here, it is unlikely to be found anywhere else. Other studies on STC have shown the benefit of STC in regions distributed across the brain (Henson et al., 1999; Calhoun et al., 2000; Sladky et al., 2011), and so it is likely the observations made here will apply

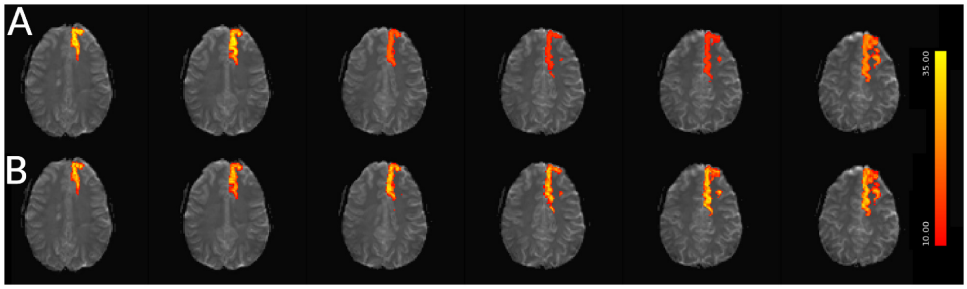


FIGURE 12 | Spatial map of functionally connected regions with and without STC on simulated data. The left Superior Frontal region is extracted using ICA on both STC data and uncorrected data. The ICA from each dataset is then regressed back on the original data for **(A)** uncorrected and **(B)** FSL STC data. In the left-most slices, regression z-statistics are higher in the uncorrected data, however, overall the STC data best describes the variance in the entire region, as seen by the drop in z-statistics in slices 3–5 in uncorrected data, while the STC data's z-statistics remain high across all slices.

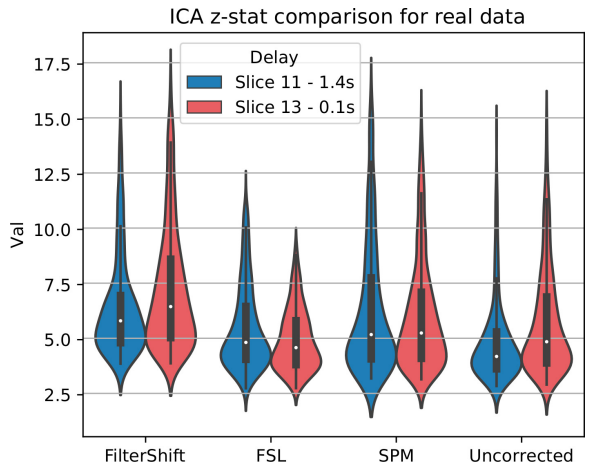


FIGURE 13 | STC on the ability to extract functionally connected regions in real data. For each real subject, a task IC was extracted with MELODIC, and the IC's time series was regressed onto the subject's data using FSL's dual regression. Z statistics from a low and high delay slice were extracted for low motion subjects and plotted here. For uncorrected data, the extracted IC performs as well as STC data in the low delay slice, but poorly explains the variance in the high delay slice, resulting in a lower z-statistic. For STC data, the average z-statistics are comparable for both high and low delay slices.

to at least the entire region of activation, if not the whole brain, though the magnitude of the effect will change with the temporal delay of the slice.

There are additional STC methods not examined in this paper. Some studies attempt to combine both MC and STC using 4D interpolation to account for temporal and spatial misalignment in one step (Roche, 2011). This approach takes into account both the position of the head in space as well as the timing of each slice. Other methods have proposed slice-by-slice realignment and resampling to a standard template (Bannister et al., 2007). This method tracks which slices contribute to each voxel, and interpolation is carried out using the known slice acquisition times. While these methods seem to be addressing the issue of interaction between STC and MC, it is crucial to emphasize that the EPI sequence in which fMRI data are acquired is a steady-state acquisition and any disruption in its equilibrium (i.e., over/under excitation due to motion) will be broadcasted to other slices and volumes and it will take multiple acquisitions until it reaches to its steady state again (Power et al., 2015). Any fMRI volume acquired during this period will be contaminated which essentially deteriorates the effectiveness of any simultaneous STC and MC.

Pipeline Order Without Spatial Smoothing

We analyzed the interaction of STC, MC, and their order in the pre-processing pipeline on different levels of motion. In simulated data, the significance of pipeline order seemed to be more effected by the interleave parameter rather than motion level. Real data seemed to indicate that there was no difference in the order for low and high motion cases, however, it's beneficial to apply motion correction first for the medium motion case for FSL and SPM. Previous work has also found differences in pipeline performance for subjects with different levels of motion (Churchill et al., 2012). As our real data has an interleave parameter of 6, these findings are not replicated in the simulated data, which showed no significant differences between pipeline order. The discrepancies between the real and simulated results may be from the fact that all simulated data used the same subject morphology. Because of this, there is no inter-subject variability across the simulated brains. This essentially makes the comparison one-to one, whereas for real data it is impossible to get to such correspondence in the selected voxels for comparison. Another factor that could contribute to these differences is an incomplete simulation of motion on our part, as motion has been known to create non-linear artifacts in the scanner that are difficult to characterize, but become non-linearly larger with greater head motion (Power et al., 2012). We also don't include spin history effects or B0 inhomogeneity in our simulation, both of which have motion interactions (Drobnjak et al., 2006). Our simulation captured rigid body motion as well as temporal shifts due to slice acquisition. The fact that our simulated results are so different from our real results simply means that there are more artifacts caused by motion than just moving-head artifacts, simulated here, and the interaction of motion and slice timing

may be more complicated than simply linearly adding the effects of both. While this is in some ways a limitation, it also gives us insight to the effectiveness of some pre-processing methods. For example, because motion parameter regression works so well for our simulated data, it is possible that this technique is good at removing variance caused by partial volume effects induced by motion. More research should be done in future to investigate this possibility.

In real data, motion artifacts work in two ways to contaminated fMRI signal: it introduces fluctuations in the signal that would both increase the variance, as well as decrease the *z*-score of the time series from the task-timing, weakening the beta estimate. This may also explain why it is only better to apply motion correction first in medium motion cases for our real data. For low motion, there is very little correction to be made, and so the actual application of MC has only a small effect on the data, and therefore has an insignificant effect on the z-scores. For high motion, the non-linear artifacts from motion may be large enough that traditional rigid body realignment is unable to correct for them. This would mean that regardless of when MC or STC is applied, there are large errors in the data that remain uncorrected, keeping the *z*-scores low. For medium motion, the motion may be large enough to benefit from rigid body realignment, while small enough to not induce significant non-linear artifacts, which would result in the maximum benefit of its application. Regardless of the amount of motion, our results suggest that it is always beneficial to apply both STC and spatial realignment. The benefit of applying them both, in any order, is much greater than the benefit of applying only one. Even for low motion subjects, the average *z*-scores from data with both MC and STC were greater than those from just MC alone.

In general, these results highlight how dependent STC and MC are on each other. In Simulated data, only performing STC was better than only performing MC for low motion subjects, however, in high-motion simulated data we find just the opposite is true, hinting that motion destroys the ability of STC to temporally realign the BOLD signals. One possibility could be that brain regions may shift from one slice to another due to motion. Because each slice is sampled at a different offset, even after spatial realignment, the time series will be made up of temporally non-uniform samples, which current STC routines cannot correct for.

While this doesn't untangle the complicated interactions between STC and motion, it does provide a quantitative measure to aid in pre-processing techniques. Recently, new preprocessing techniques have attempted to reconcile these interactions (Jones et al., 2008; Beall and Lowe, 2014). Jones et al. (2008) uses a method that involves quantifying which native space slice each voxel came from in the registered standard space volume. By taking into account the true sampling delay of each voxel, an accurate voxel-wise estimate of physiological noise can be created to account for such delays. By using a similar technique to identify the true time of acquisition of any given voxel in a registered volume, it may be possible to reduce the slice timing/motion interaction to a non-uniform sampling and reconstruction problem. Future work could involve examining the efficacy of such a model.

The dice overlap for simulated data shown in **Figure 6** indicates that as motion increases, the order in which STC and MC are applied becomes more important with regard to the voxel overlap. This can be seen by comparing the "Before MC" to the "After MC" columns. The more similar these columns are, the less important pipeline order is for spatial overlap. MC before STC has a higher dice overlap than when STC is applied first. One possible reason for this is because the SR has no STC, and the shifts are accounted for in the regression itself. This could be seen as "MC before STC," since the timing is accounted for after MC. Assuming that the SR method does successfully identify the voxels with the highest signal, we can infer that higher overlap with this method is related to a better pipeline. Interestingly, uncorrected data seems to be able to maintain the location of the top 20 voxels even in high motion, however from **Table 1**, we know that the *z*-score of these voxels are significantly reduced. For real data shown in **Figure 7**, we see that our method becomes more reliant on MC for a high spatial overlap as the motion level increases ("No MC" columns compared to any other column). In real data, we see that the effect of the order of STC and MC is smaller. We also see a significant drop in dice overlap when MC is not applied. This further highlights the importance of performing both STC and MC. Further, STC seems to benefit voxel overlap more than MC. The effect of No MC can be seen by comparing the "Uncorrected" bar to any other bar in the "No MC" column. Only MC is applied to "uncorrected data" in the "Before MC" case, so the effect of MC can be seen by comparing the "Uncorrected" bar in the "No MC" column with the "uncorrected" bar in the "Before MC" column. "No MC" means no MC has been applied, and so for uncorrected data, neither STC nor MC is performed in this category.

Prewhitening had no significant effect on STC in real data. In addition, our results are in line with the existing literature. It has been demonstrated in Olszowy et al. (2019) that STC has only a small impact on autocorrelation, and did not significantly influence the results when combined with PW. Combining PW and MC did effect the results, and this result was modulated by motion. It has been shown previously that motion correction can induce temporal correlations between voxels (Power et al., 2012). It's possible that these motion artifacts are also inducing autocorrelation within a voxel. Understanding this phenomenon would be valuable to the field, and is left as an area of future research.

The reliability measures should not be taken as evidence that there are no differences between pipelines. The BOLD signal only makes up about 5% of the variance in fMRI. For a given task, the vast majority of the voxels are only noise, and have no relevant signal in their time series. Applying a time shift on these noisy voxels will have no effect on the reliability. Of the voxels that are activated, only those that fall on high delay slices will see significant changes due to STC. Because of this, the reliability measure will be made up of mostly voxels that have little to no change due to slice timing correction, with only a small portion seeing any real difference. This is why we wouldn't expect to see a difference in the reliability measure, and it is for this reason that we focused our analysis on voxels with high *t*-values in high-delay slices for all investigations in the paper.

Motion Parameter Residualization

Motion parameter residualization lowered the *z*-score slightly in most cases for both real and simulated data, though the effects were larger for simulated data. This indicates that the presence of motion parameter estimates in the GLM likely interfere with accurate beta estimation. This can be caused by colinearity between the motion parameters and the regressor being used in the analysis. When using MPR, in simulated data, STC before MC was extremely detrimental to the beta value estimation. This could be because STC effects the accurate estimation of motion parameters (Power et al., 2017). It is possible that the results for STC before MC would improve if the parameters are estimated first, and MC itself is performed after STC. There are many cases where "No MC" performs almost as well as other pipelines, again highlighting the fact that our simulated data does not add non-linear motion artifacts to the data, therefore linear regression of motion parameters should be sufficient for removing most of the variance. It's also important to note that our analysis was performed on highly significant voxels, which tend to fall inside the ROI, rather than along the edges. It's likely that motion has a larger effect on voxels along the edges of activation, as they will be more susceptible to sampling a new brain region with no task-related signal when the subject moves.

For real data, the low motion subjects saw a decrease in statistics with the addition of motion parameter regression, indicating that the added regressors either failed to model the low motion artifacts, or there simply wasn't enough motion-induced variance to make up for the loss in degrees of freedom. If the motion in the scan is simply too low for an accurate estimate, the realignment parameters would be essentially noise. Real data with MPR and no STC performed significantly worse than all methods with STC for medium and high motion. It's possible that the motion parameters themselves are sensitive to time shifts.

Our results agree with Churchill et al., who found that MPR was generally unhelpful, and in fact harmful for low motion subjects, in the sense that it lowered their chosen measurement of quality in that paper (Churchill et al., 2012). However, it is difficult to compare our motion levels with this study, as FWD is not used in their study, and "high motion" is based only on the extracted yaw parameters.

Spatial Smoothing on STC

For spatial smoothing, the achieved STC gain became very small with a large kernel size ($>$6.5 mm). For real data, differences between STC methods disappeared with a 5 mm smoothing kernel, and at 8 mm STC data is indistinguishable from uncorrected data for low and medium motion subjects. It has been shown that the optimal FWHM varies between 6 and 12 mm, depending on the experimental tasks and the statistics of interest (Mikl et al., 2008). It is important to note that at this level removing the STC step might slightly improve or worsen the results, depending on the location of a given ROI, as shown by the *z*-score of the "uncorrected" column of the low motion subjects. This is rather counter-intuitive and might be considered as evidence that STC is not beneficial at all. However, recent studies such as the HCP has also taken a critical look at smoothing and its effects on neuroimaging data, and has greatly restricted the size of the kernel, as well as the extent of the areas used in smoothing (Glasser et al., 2013), which may be enough to retain the benefits of STC.

As in the real data, mean *z*-score from simulated uncorrected data only decreased monotonically for sequential acquisition and interleave 2 acquisitions. For interleave 6 acquisition, the initial improvement seen in the 3.5, 5, and 6.5 mm kernels are caused by the phenomenon described in the "Introduction," where time series from adjacent slices with less delay actually improve the statistics of the high delay slice. Initially, this benefits the *z*-score, shown in the 3.5 mm kernel case, as it averages the data with voxels in adjacent slices. These voxels are still in the LSF region, but with lower temporal delay. This initial increase is reduced as the kernel size continues to grow, because the LSF region is just a narrow (less than 5 mm) cortical region, which cannot extend beyond two voxels in the x/y plane. Increasing the kernel size beyond 3.5 mm will cause the blending of the voxels outside the LSF that have completely different underlying BOLD signal, which essentially corrupts the time series, reducing the *z*-score. These results do not suggest that spatial smoothing should not be performed, but rather they show how smoothing interacts with, and may obscure, the effects of precise preprocessing steps such as STC.

Finally, it is important to note the failure of slice-based regressors (SR) method, as the gold standard, in the presence of smoothed data. This method performs worse than FS in simulated data with interleave 2 and interleave 6 with the 8 mm kernel, and for real data it performs worse than all other STC methods for all motion levels with the 8mm kernel. This is due to the fact that in our gold standard, the regressors are accounting for a slice dependent delay. Spatial smoothing averages a slice with time series from the surrounding slices that fall within the kernel. Each adjacent slice was sampled with a different temporal delay. Smoothing this time series with data from surrounding slices acquired at different offsets greatly contaminates the fMRI signal, and eliminates any benefit gained by using slice-specific regressors. Additionally, in our gold standard pipeline, the data was filtered with the same low-pass filter as used in the FS method without performing any temporal shifting on the data. If this filtering was not performed, the SR method would likely be even worse than FSL and SPM's STC techniques (which only apply very slight low-pass filtering, inherent to their interpolation kernels). This suggests that slice-based regressions should not be used with data that has undergone spatial smoothing.

STC on Short TRs

Both our simulated and real data are based on event-related design, which requires fast ($<$5 s) and random stimuli duration. Block-designed experiments, on the other hand, have much longer ($>$10 s) stimulus duration, thus making it less susceptible to the acquisition offset delay. It has already been shown in the literature that block design accompanied with moderate 3D smoothing may not benefit substantially from STC (Sladky et al., 2011). Therefore, we decided not to repeat that experiment again

in this work. Nevertheless, extremely high TRs (TR > 5 s) without STC could not be tolerated even for block related designs. On the other hand, it is generally assumed that sub-second TR data will not benefit from STC. We have shown with our results that FS STC does improve *z*-score, even on multi-band data. The increased number of time points only makes the regressors more sensitive to temporal shifts and interpolation errors. Because our method performs a more accurate reconstruction than traditional STC techniques, we are able to significantly improve the *z*-score in these high frequency data sets. Furthermore, a short TR may reduce aliasing of physiological noise into the frequency range of the BOLD signal. Our STC method can potentially filter out more of the high frequency noise, and retain only the range of frequencies in which the BOLD signal is biologically plausible. This significantly improves *z*-score beyond what traditional STC is capable of.

Effect of STC on Functional Connectivity Analysis

We have shown that FC network extraction can benefit from STC. In simulated data, we clearly show the problem faced when extracting task based networks from data without STC. The difference in z-statistics between slices with large temporal offsets is observable in the uncorrected data, but is completely corrected for in STC data. This is shown quantitatively in **Figure 11** and qualitatively in **Figure 12**. From these figures, we can see that the extracted IC's time series "locked" strongly with slices of a certain temporal delay. Even a voxel in the same region with the same time course, but sampled at an offset, will be assigned a weaker weight to that IC. That IC's time series will also account for less variance in the dual regression, resulting in lower z-statistics. The reason we don't see the difference between high and low delay slices in real uncorrected data is partly due to inter-subject variability in region activation. Additionally, there's no guarantee that MELODIC will always extract an IC with a time series aligned to the low-delay slices. If the variability between subjects is enough to randomly distribute the slices with which the IC correlates more strongly to (High delay slices or low delay slices), then we would expect the group level difference to become smaller and smaller as we add subjects to the analysis. For example, one subject's IC might be more correlated with a high delay slice, while the other subject's IC might be more correlated with a low delay slice. These differences would average out, and we wouldn't see the dramatic delay effect that we do in simulated data. It's true that there is also no guarantee that MELODIC would consistently prefer low-delay slices for simulated data, however, given the identical morphology, activation region, and low motion, we effectively minimized all non-signal related sources of variance, so we would expect similar results.

Despite these potential confounds, we are still able to show that STC, specifically FS, increases the statistical fit of the IC's extracted time series to the data in both real and simulated subjects. By removing sources of variance such as temporal offsets, voxels are more strongly correlated with each other, resulting in a better network extraction, as well as a more accurate estimate of the network's time series.

CONCLUSION

In an effort to critically examine the effects of STC in the context of a standard preprocessing pipeline, we have performed a series of experiments in which we manipulate the preprocessing pipeline on data with various levels of motion and with various slice acquisition orders. While no clear rule for the order of pre-processing steps emerged, we were able to reinforce the importance of both STC and MC. Here, we summarize the key findings from our experiments: For all levels of motion and for all slice acquisition orders, STC and MC together significantly improved z-scores in both real and simulated data. MPR altered nearly all simulated cases, as well as all medium and high motion cases for real subjects. For real data, MC before STC was always the optimal pipeline when using MPR, likely due to the fact that a more accurate motion estimate can be obtained if MC is performed first (Power et al., 2017). However, this nuisance regression proved detrimental for all *z*-scores across all scan conditions. We found that on multiband data with extremely short TRs, STC significantly improved results. Importantly, we found that STC of any kind never lowered the resulting *z*-scores. The only case that showed minimal improvement from STC was data filtered with large smoothing kernels. However, in studies using smaller smoothing kernels, such as in surface based analysis, STC will only become more beneficial as the smoothing kernels shrink. Finally, we found that STC improves the coherence across slices of FC networks extracted from task-based fMRI data using ICA. STC improved the fit of the IC to the time series' of voxels in the IC, especially those in high delay slices.

This paper provided an in-depth look at many common situations in which STC may be implemented. We have shown that in all such cases, STC is a valuable addition to the preprocessing pipeline.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

DP ran all the analysis, generated the figures, and wrote the manuscript. QR generated the simulated data set and wrote the manuscript.

FUNDING

This work was supported in part by the NIA K01 AG044467, NIA R01 AG026158, and NIH-NIA R01 AG057962 grants.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.00821/full#supplementary-material

REFERENCES

Bannister, P. R., Michael Brady, J., and Jenkinson, M. (2007). Integrating temporal information with a non-rigid method of motion correction for functional magnetic resonance images. *Image Vis. Comput.* 25, 311–320. doi: 10.1016/j.imavis.2005.10.002

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189. doi: 10.1016/j.neuroimage.2013.05.033

Barth, M., Breuer, F., Koopmans, P. J., Norris, D. G., and Poser, B. A. (2016). Simultaneous multislice (SMS) imaging techniques. *Magn. Reson. Med.* 75, 63–81. doi: 10.1002/mrm.25897

Beall, E. B., and Lowe, M. J. (2014). SimPACE: generating simulated motion corrupted BOLD data with synthetic-navigated acquisition for the development and evaluation of SLOMOCO: a new, highly effective slicewise motion correction. *Neuroimage* 101, 21–34. doi: 10.1016/J.NEUROIMAGE.2014.06.038

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage* 20, 1052–1063. doi: 10.1016/S1053-8119(03)00435-X

Beckmann, C. F., Mackay, C. E., Filippini, N., and Smith, S. M. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage* 47:S148.

Beckmann, C. F., and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* 23, 137–152. doi: 10.1109/TMI.2003.822821

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995a). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995b). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., et al. (2009). Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci.* 29, 1860–1873. doi: 10.1523/JNEUROSCI.5062-08.2009

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., and Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *Neuroimage* 45, 758–768. doi: 10.1016/j.neuroimage.2008.12.035

Calhoun, V., Golay, X., and Pearlson, G. (2000). "Improved fMRI slice timing correction: interpolation errors and wrap around effects," in *Proceedings, ISMRM, 9th Annual. Meeting*, Denver.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63, 289–300. doi: 10.1016/J.NEUROIMAGE.2012.07.004

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., et al. (2012). Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Hum. Brain Mapp.* 33, 609–627. doi: 10.1002/hbm.21238

Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., and Strother, S. C. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. *PLoS One* 10:e0131520. doi: 10.1371/journal.pone.0131520

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* 154, 174–187. doi: 10.1016/J.NEUROIMAGE.2017.03.020

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

Di, X., Gohel, S., Kim, E. H., and Biswal, B. B. (2013). Task vs. Rest—different network configurations between the coactivation and the resting-state brain networks. *Front. Hum. Neurosci.* 7:493. doi: 10.3389/fnhum.2013.00493

Drobnjak, I., Gavaghan, D., Süli, E., Pitt-Francis, J., and Jenkinson, M. (2006). Development of a functional magnetic resonance imaging simulator for modeling realistic rigid-body motion artifacts. *Magn. Reson. Med.* 56, 364–380. doi: 10.1002/mrm.20939

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., et al. (2010). Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. *PLoS One* 5:e15710. doi: 10.1371/journal.pone.0015710

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., et al. (2004). Sequence-independent segmentation of magnetic resonance images. *Neuroimage* 23(Suppl.1), S69–S84. doi: 10.1016/j.neuroimage.2004.07.016

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the Human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI1. *Neuroimage* 9, 416–429. doi: 10.1006/NIMG.1998.0419

Greve, D. N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 48, 63–72. doi: 10.1016/j.neuroimage.2009.06.060

Henson, R., Buechel, C., Josephs, O., and Friston, K. (1999). The slice-timing problem in event-related fMRI. *Neuroimage* 9:125.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/NIMG.2002.1132

Jones, T. B., Bandettini, P. A., and Birn, R. M. (2008). Integration of motion correction and physiological noise regression in fMRI. *Neuroimage* 42, 582–590. doi: 10.1016/j.neuroimage.2008.05.019

Kim, B., Boes, J. L., Bland, P. H., Chenevert, T. L., and Meyer, C. R. (1999). Motion correction in fMRI via registration of individual slices into an anatomical volume. *Magn. Reson. Med.* 41, 964–972. doi: 10.1002/(sici)1522-2594(199905)41:5<964::aid-mrm16>3.3.co;2-4

Krienen, F. M., Yeo, B. T. T., and Buckner, R. L. (2014). Reconfigurable task-dependent functional coupling modes cluster around a core functional architecture. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130526. doi: 10.1098/rstb.2013.0526

Mikl, M., Mareèek, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., et al. (2008). Effects of spatial smoothing on fMRI group inferences. *Magn. Reson. Imaging* 26, 490–503. doi: 10.1016/j.mri.2007.08.006

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., et al. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *J. Cogn. Neurosci.* 14, 1200–1214. doi: 10.1162/089892902760807203

Olszowy, W., Aston, J., Rua, C., and Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. *Nat. Commun.* 10:1220. doi: 10.1038/s41467-019-09230-w

Parker, D., Liu, X., and Razlighi, Q. R. (2016). Optimal slice timing correction and its interaction with fMRI parameters and artifacts. *Med. Image Anal.* 35, 434–445. doi: 10.1016/j.media.2016.08.006

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320–341. doi: 10.1016/J.NEUROIMAGE.2013.08.048

Power, J. D., Plitt, M., Kundu, P., Bandettini, P. A., and Martin, A. (2017). Temporal interpolation alters motion in fMRI scans: magnitudes and consequences for artifact detection. *PLoS One* 12:e0182939. doi: 10.1371/journal.pone.0182939

Power, J. D., Schlaggar, B. L., and Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* 105, 536–551. doi: 10.1016/J.NEUROIMAGE.2014.10.044

Razlighi, Q. R. (2018). Task-evoked negative BOLD response in the default mode network does not alter its functional connectivity. *Front. Comput. Neurosci.* 12:67. doi: 10.3389/fncom.2018.00067

Roche, A. (2011). A four-dimensional registration algorithm with application to joint correction of motion and slice timing in fMRI. *IEEE Trans. Med. Imaging* 30, 1546–1554. doi: 10.1109/TMI.2011.2131152

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/J.NEUROIMAGE.2009.10.003

Sala-Llonch, R., Bartres-Faz, D., and Junque, C. (2015). Reorganization of brain networks in aging: a review of functional connectivity studies. *Front. Psychol.* 6:663. doi: 10.3389/fpsyg.2015.00663

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage* 64, 240–256. doi: 10.1016/J.NEUROIMAGE.2012.08.052

Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., and Windischberger, C. (2011). Slice-timing effects and their correction in functional MRI. *Neuroimage* 58, 588–594. doi: 10.1016/j.neuroimage.2011.06.078

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, S208–S219. doi: 10.1016/J.NEUROIMAGE.2004.07.051

Stehling, M. K., Turner, R., Mansfield, P., and Ljunggren, S. (1991). Echo-planar imaging: magnetic resonance imaging in a fraction of a second. *Science* 254, 43–50. doi: 10.1126/science.1925560

Stern, Y., Habeck, C., Steffener, J., Barulli, D., Gazes, Y., Razlighi, Q., et al. (2014). The reference ability neural network study: motivation, design, and initial feasibility analyses. *Neuroimage* 103, 139–151. doi: 10.1016/J.NEUROIMAGE.2014.09.029

Tegeler, C., Strother, S. C., Anderson, J. R., and Kim, S.-G. (1999). Reproducibility of BOLD-based functional MRI obtained at 4 T. *Hum. Brain Mapp.* 7, 267–283. doi: 10.1002/(sici)1097-0193(1999)7:4<267::aid-hbm5>3.0.co;2-3

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Vogt, K., Ibinson, J., Small, R., and Schmalbrock, P. (2009). "Slice-timing correction affects functional MRI noise, model fit, activation maps, and physiologic noise correction," in *Proceedings 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine*, Honolulu.

Wagner, K., Frings, L., Quiske, A., Unterrainer, J., Schwarzwald, R., Spreer, J., et al. (2005). The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. *Neuroimage* 28, 122–131. doi: 10.1016/J.NEUROIMAGE.2005.06.005

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage* 21, 1732–1747. doi: 10.1016/j.neuroimage.2003.12.023

Wu, C. W., Chen, C.-L., Liu, P.-Y., Chao, Y.-P., Biswal, B. B., and Lin, C.-P. (2011). Empirical evaluations of slice-timing, smoothing, and normalization effects in seed-based, resting-state functional magnetic resonance imaging analyses. *Brain Connect.* 1, 401–410. doi: 10.1089/brain.2011.0018

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., et al. (2009). Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magn. Reson. Imaging* 27, 264–278. doi: 10.1016/J.MRI.2008.05.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Parker and Razlighi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neural Circuits

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fncir.2019.00054

The Concept of Transmission Coefficient Among Different Cerebellar Layers: A Computational Tool for Analyzing Motor Learning

Saeed Solouki[1], *Fariba Bahrami*[1]* *and Mahyar Janahmadi*[2]

[1] *Control and Intelligent Processing Center of Excellence, Human Motor Control and Computational Neuroscience Laboratory, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran,* [2] *Department of Physiology, Neuroscience Research Center, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

High-fidelity regulation of information transmission among cerebellar layers is mainly provided by synaptic plasticity. Therefore, determining the regulatory foundations of synaptic plasticity in the cerebellum and translating them to behavioral output are of great importance. To date, many experimental studies have been carried out in order to clarify the effect of synaptic defects, while targeting a specific signaling pathway in the cerebellar function. However, the contradictory results of these studies at the behavioral level further add to the ambiguity of the problem. Information transmission through firing rate changes in populations of interconnected neurons is one of the most widely accepted principles of neural coding. In this study, while considering the efficacy of synaptic interactions among the cerebellar layers, we propose a firing rate model to realize the concept of transmission coefficient. Thereafter, using a computational approach, we test the effect of different values of transmission coefficient on the gain adaptation of a cerebellar-dependent motor learning task. In conformity with the behavioral data, the proposed model can accurately predict that disruption in different forms of synaptic plasticity does not have the same effect on motor learning. Specifically, impairment in training mechanisms, like in the train-induced LTD in parallel fiber-Purkinje cell synapses, has a significant negative impact on all aspects of learning, including memory formation, transfer, and consolidation, although it does not disrupt basic motor performance. In this regard, the overinduction of parallel fiber-molecular layer interneuron LTP could not prevent motor learning impairment, despite its vital role in preserving the robustness of basic motor performance. In contrast, impairment in plasticity induced by interneurons and background activity of climbing fibers is partly compensable through overinduction of train-induced parallel fiber-Purkinje cell LTD. Additionally, blockade of climbing fiber signaling to the cerebellar cortex, referred to as olivary system lesion, shows the most destructive effect on both motor learning and basic motor performance. Overall, the obtained results from the proposed computational framework are used to provide a map from procedural motor memory formation in the cerebellum. Certainly, the generalization of this concept to other multi-layered networks of the brain requires more physiological and computational researches.

Keywords: transmission coefficient, multi-layer neural network, synaptic plasticity, cerebellar motor learning, optokinetic reflex

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Christian Hansel,
University of Chicago, United States
Alberto Antonietti,
Politecnico di Milano, Italy

***Correspondence:**
Fariba Bahrami
fbahrami@ut.ac.ir

Received: *16 June 2019*
Accepted: *29 July 2019*
Published: *27 August 2019*

Citation:
Solouki S, Bahrami F and Janahmadi M (2019) The Concept of Transmission Coefficient Among Different Cerebellar Layers: A Computational Tool for Analyzing Motor Learning. Front. Neural Circuits 13:54. doi: 10.3389/fncir.2019.00054

INTRODUCTION

The cytoarchitecture of the cerebellar cortex has a uniform character and is roughly divided into three parallel circuit elements (Marzban et al., 2015), which, from the inner to the outer layer, are called the granular, the Purkinje, and the molecular layers (**Figure 1A**). Effective communication between these layers is mainly regulated by synapses (Evans, 2007). Generally, synaptic plasticity has multifaceted roles in the formation, transfer, and consolidation of memory, which are critical for learning, and also for enhancement of signal-to-noise ratio for robust neurotransmission (D'Angelo et al., 2005; Jörntell and Hansel, 2006; Borjkhani et al., 2018). Despite an extensive amount of electrophysiological and behavioral studies on the mechanisms underlying synaptic plasticity (Gao et al., 2012; Freeman, 2015), the precise contribution of each type of plasticity in the evolution of motor memory remains debated (Yuzaki, 2013; D'Angelo et al., 2016), since various parts of the cerebellar circuit are involved in controlling the baseline motor performance, simultaneously integrating multiple cortical and sensory inputs, and modulating the output firing rate of deep nuclei (DN) toward premotor areas (Mauk et al., 2000; Cheron et al., 2013; Badura et al., 2016).

The existence of different forms of synaptic plasticity, besides access to sensorimotor information, has enabled the cerebellum to participate actively in motor learning (De Zeeuw and Ten Brinke, 2015). Motor learning in the cerebellum is mediated by multiple plasticity sites at both cortical and nuclear parts. Naturally, organizing movement patterns in a coordinated manner requires the interaction of different forms of synaptic plasticity in the cerebellar network (Gao et al., 2012). In fact, an individual type of plasticity alone cannot account for wide dynamic ranges of cerebellar learning (Boyden et al., 2004; Solouki and Pooyan, 2016).

As schematically illustrated in **Figure 1B**, the transmission flow of mossy fibers' (MFs) input to DN consists of two distinct parts: (1) the direct pathway that relays MFs sensory information directly to DN and (2) the indirect or cortical pathway involving MFs, granule cells (GCs), parallel fibers (PFs), and Purkinje cells (PCs). In addition to the mainstream of the cortical pathway, PCs receive inhibitory input from molecular layer interneurons (MLIs). The indirect pathway operates as a fast learning module while deeper structures (direct pathway) work as a slow learning module, wherein the motor skill is transferred and consolidated into more persistent memory (Lee et al., 2015; Luque et al., 2016). The concept of memory formation in the cerebellar cortex and its later transfer to the nuclear tract is supported by the experimental evidence that the flocculus shutdown after four sessions of optokinetic reflex (OKR) adaptation does not impair the previously acquired motor memory (Shutoh et al., 2006). Reversible blockade of floccular protein synthesis by actinomycin or anisomycin microinjection produced a similar effect (Okamoto et al., 2011). The classical theory of cerebellar learning states that repetition of short-term training sessions can gradually lead to persistent changes in the synaptic efficacy (Ito, 2000; Strata, 2009; De Zeeuw and Ten Brinke, 2015). It is postulated that long-term depression (LTD) at PF-PC synapses constitutes the major mechanism of motor learning in the cerebellar cortex (Ito, 2001; Gao et al., 2012; Hirano, 2013). The strength of PF-PC synapses is adjusted under the guidance of climbing fibers (CFs). Thus, the olivary system seems to train the synapses by encoding the error signals that drive cerebellar learning (Carrel, 2012). The presence or absence of CF signal during PF activity causes LTD or LTP, respectively, at the PF-PC synapses. Postsynaptic plasticity at PF to MLI synapses is also mediated, through spillover, by CF activation in the opposite direction of PF-PC plasticity (Hirano, 2013). On the other hand, the plasticity at MF-DN is driven by synchronized activities of MFs and PCs (Gosui and Yamazaki, 2016).

Recently, a controversy arose from the observation in several studies, indicating that although LTD was abrogated genetically or pharmacologically, motor learning in non-anesthetized behaving animals remained unchanged (Welsh et al., 2005; Faulstich et al., 2006; Schonewille et al., 2011; Yuzaki, 2013; Inoshita and Hirano, 2018). To solve the observed controversy, some researchers justified the PF-LTD vs. motor learning mismatch by considering the possibility of a functional or structural compensatory mechanism, which has yet to be proven (Schonewille et al., 2011; Gao et al., 2012). Some others proposed long-term potentiation (LTP) as an alternative substrate for motor learning (Schonewille et al., 2010) and reduced the role of LTD to an inhibitor of synaptic saturation or a protector against excitotoxic PC death during the inferior olive (IO) overactivation (Welsh et al., 2005). One of the reasons of ambiguity in determining the role of LTD and LTP in motor learning is that changes in each one alone, regardless of what happens to the other, are being studied. In most genetic studies, where molecules required for LTD induction were manipulated, LTP alterations have not yet been investigated (Yuzaki, 2013). Furthermore, the deleted gene of interest has a global effect on every cell of the brain and not just the cells of interest. Thus, one cannot be sure that the resulting phenotype is exclusively related to LTD. In the case of discrepancies between PF-LTD and motor learning, more specific experimental or computational analysis are warranted before drawing any final conclusions.

In this study, we take the advantages of organizational uniformity and layered structure of the cerebellum to develop a computational model for investigating the highly distributed nature of cerebellar synaptic plasticity during a motor learning process. Assuming the synapse as a basis for neurotransmission (Stevens, 2003), plasticity is considered as a mechanism for regulating the transmission coefficient among neuronal layers. The main idea behind our computational scheme lies in estimating the effect of altered transmission coefficient among cerebellar layers on the equilibrium behavior of the system. In the first part of the paper, we provide an analytical framework to describe the rate of changes at synaptic weights as a function of population activity in cerebellar neurons and determine five different conjunctions to evaluate the following scenarios:

1) **MF-GC:** The importance of MF-GC synapses in enhancing signal transmission and tuning spike time of GCs on the millisecond order has been shown so far (Garrido et al., 2013b; Sgritta et al., 2017). Here, we decrease the transmission

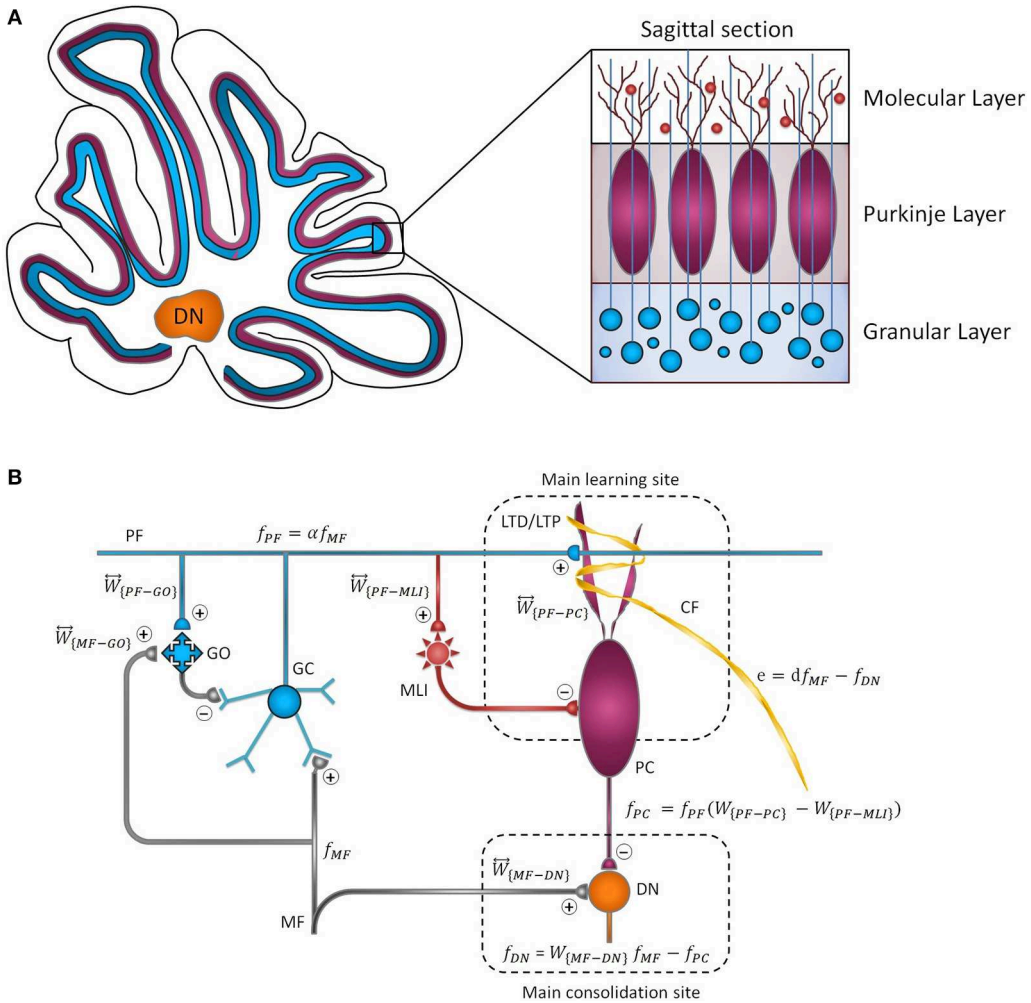


FIGURE 1 | Cytoarchitecture of the cerebellum. **(A)** Cross-sectional views of the cerebellum showing the folding pattern and layered character of the cortex. **(B)** Schematic representation of cerebellar microcircuit. There are two principal afferents to the cerebellar cortex: (1) MFs, which provide sensory information of screen oscillation to GCs, GOs, and DN, and (2) CFs, which convey the retinal image slip information to PCs and drive cortical learning by changing the synaptic efficacy at PF-PC synapses. The ascending axons of the GCs branch in a T-shaped manner to form PFs, which, in turn, make excitatory synaptic contacts with MLIs and PCs. PF-PC and MF-DN synapses are respectively considered as the main sites of learning and consolidation. Synaptic weights are indicated by $\overleftrightarrow{W}$.

coefficient of the input layer among MFs and GCs to elucidate whether this synapse has any role in motor learning and gain adaptation or not.

2) **PF-PC:** We reevaluate the aforementioned dichotomy between PF-LTD and motor learning by considering the intrinsic differences between the spontaneous LTD, which is induced by background activity of CFs and train-induced LTD, which is triggered by use-dependent activity of CFs. Moreover, we use the model to examine the effect of LTD and LTP blockade on motor learning and baseline motor performance. Meanwhile, we monitor the plastic changes at PF to MLI synapse to determine whether the MLI pathway has any backup role for the PF-PC pathway and, if so, to what extent it would contribute in compensating the LTD/LTP defect at PF-PC synapses. Additionally, our model allowed for a temporally specific ablation of LTD, which helps us to clarify different roles played by the early and late phases of LTD in short- and long-term learning.
3) **IO (CF pathway):** Emerging evidence from experimental studies indicates a high level of coordination between olivary system activity and cerebellar learning (Pijpers et al., 2005; Van Der Giessen et al., 2008; Badura et al., 2013; Solouki et al., 2018). However, it is still unclear which phase of learning will be affected by olivary system disruption. Probably, the extent of functional problems caused by IO lesion would be far beyond the defect of PF-LTD because CFs, which are the axons of IO neurons, are involved in adjusting the synaptic plasticity at both molecular and Purkinje layers. To understand the possible effect of olivary system lesion on cerebellar learning, we simulate learning paradigm in the absence of CF signal.
4) **PF-MLI:** Implication of MLIs in motor learning and the possibility of their synergistic interaction with LTD and LTP at

PF-PC synapses are studied from two perspectives. Along with the second scenario, we first examine whether interneurons are able to compensate impairment at PF-PC synapses and, reciprocally, how would impairment at PF-MLI synapses affect PF-PC synaptic plasticity and learning behavior of the model.

5) **PC-DN:** The basis for motor memory consolidation is a controversial topic. One view is that the acquired short-term memory in the cerebellar cortex is transferred to a new location during the learning process, that of so-called saving phenomena (Kellett et al., 2010). In the last scenario, we try to verify the existence of such hierarchical causality by disconnecting the cortical and nuclear parts of the cerebellum.

As a model for cerebellar motor learning, the adaptation of OKR is used. OKR is a kind of cerebellar-dependent oculomotor response that serves to stabilize moving images on the retina (Inoshita and Hirano, 2018). The potential impact of the lesion scenarios on the behavioral trait of the model is addressed by employing the following terms and definitions: (a) basic motor performance, i.e., the initial value of OKR gain before any learning paradigm has taken place; (b) motor learning, which corresponds to adaptation of OKR gain following a learning paradigm; and (c) motor consolidation, which means long-term preservation of the level of adaptation of OKR gain. In the final step, putting together the results of the above scenarios in the execution of a repetitive learning paradigm and tracing their behavioral differences at both phases of short- and long-term learning, we propose a conceptual map of procedural motor memory formation in the cerebellum.

MATERIALS AND METHODS

Definition of Plasticity Rules

Experimental studies in different parts of the brain, including the hippocampus, neocortex, and cerebellum, have revealed activity-dependent processes that can persistently change the efficacies of synapses (Dayan and Abbott, 2001; Gerstner et al., 2014). In this section, we take the advantages of differential equations to characterize the rate of changes of synaptic efficacies as a function of population activity in cerebellar neurons. The transmission delays between the different neural populations are disregarded.

Cerebellar Cortex

Looking at **Figure 1B**, the population firing rate of MF afferents is denoted by $f_{MF} \in R^n$. In the context of OKR adaptation, MFs transmit sensory information of visual stimulus to the granular layer and DN. The MF input is then propagated to GCs with an m by n synaptic connectivity matrix, presumably, $m \gg n$. The GCs are also under the influence of feedforward and feedback inhibition from Golgi cells (GOs). Thus, the output of GCs, denoted by $f_{PF} \in R^m$, can be written as follows:

$$f_{PF} = f_{MF} - W_{\{PF-GO\}} f_{PF} - W_{\{MF-GO\}} f_{MF} \quad (1)$$

$$f_{PF}(1 + W_{\{PF-GO\}}) = f_{MF}(1 - W_{\{MF-GO\}}) \quad (2)$$

$$f_{PF} = \left(\frac{1 - W_{\{PF-GO\}}}{1 + W_{\{MF-GO\}}}\right) f_{MF} = \alpha f_{MF} \quad (3)$$

where $W_{\{MF-GO\}}$ and $W_{\{PF-GO\}}$ are constant representing the synaptic weight of MF-GO and PF-GO, respectively. $\alpha = \left(\frac{1-W_{\{PF-GO\}}}{1+W_{\{MF-GO\}}}\right)$ is defined as transmission coefficient between excitatory mossy fiber input and granular layer. It is postulated that MF burst activation not only induces presynaptic LTP but also enhances intrinsic excitability of the GCs. This augmented excitability results from an increased input impedance and reduced spike threshold, which heightens excitatory postsynaptic potential (EPSP) and facilitates spike output (Armano et al., 2000; Seja et al., 2012). Thus, despite the low background firing rate of GCs owing to tonic inhibition by GOs, sensory stimulus can evoke bursting in GCs, such that MFs input is transmitted with high reliability to the Purkinje layer (D'Angelo and De Zeeuw, 2009). Moreover, the combination of feedforward and feedback inhibition by GOs opens up a fine-tuned time window for maximum transmission of sensory information with a high signal to noise ratio (Chadderton et al., 2004; Gao et al., 2012). Therefore, the transmission coefficient of the input layer would be ≤ 1 ($\alpha \leq 1$). Obviously, this value is higher in the normal group than the lesion one. Herein, α is simply assumed to be near 1 in the normal and 0.3 in the lesion condition.

The excitatory output of GCs is sent to the Purkinje layer by PF axons. Also, the output of each PC is assumed to be $f_{PC} \equiv W_{\{PF-PC\}} f_{PF} - W_{\{PF-MLI\}} f_{PF}$, where $W_{\{PF-PC\}} \in R \times R^m$. The synaptic weight of direct connection among MFs and DN is defined by $W_{\{MF-DN\}} \in R \times R^n$. The final output of the nuclear module (DN) toward the oculomotor neurons is given by $f_{DN} = W_{\{MF-DN\}} f_{MF} - f_{PC}$. The GABAergic nature of PC is taken into account in writing f_{DN}.

The model gradually learns to adapt the actual output f_{DN} to the desired motor output df_{MF}. If the desired gain d is larger (smaller) than the initial state, the synaptic weight experiences depression (potentiation) in $W_{\{PF-PC\}}$ and potentiation (depression) in $W_{\{MF-DN\}}$ (Jörntell and Hansel, 2006; Gosui and Yamazaki, 2016). As mentioned in the previous section, CFs pass on the learning error $e \equiv df_{MF} - f_{DN}$ to PCs, which enables supervised learning at PF-PC synapses. The objective function of supervised learning is given as

$$J\left(W_{\{PF-PC\}}\right) = \frac{1}{2} e^2 \quad (4)$$

The least mean square (LMS) method can be used to determine the steepest descent direction of the error function (Kawato, 2009):

$$\Delta W_{\{PF-PC\}} \propto -\frac{\partial J}{\partial W_{\{PF-PC\}}} = -f_{PF} e = -f_{PF}(CF - \overline{CF}) \quad (5)$$

where $(CF - \overline{CF})$ is the difference between the spontaneous and background activity of CFs, assumed to encode the error signal. In the case that both PFs and CFs are activated, f_{PF} and $(CF - \overline{CF})$ are positive, and thus $W_{\{PF-PC\}}$ is negative (LTD). We can write

$$\dot{W}_{\{PF-PC\}} = -\lambda_1 \frac{\partial J}{\partial W_{\{PF-PC\}}} = -\frac{1}{2}\lambda_1 \frac{\partial e^2}{\partial W_{\{PF-PC\}}} = -\lambda_1 f_{PF} e \quad (6)$$

where λ_1 is the learning rate. In contrast, lone activation of PFs in the absence of e, leads to LTP (Lev-Ram et al., 2002; Coesmans

et al., 2004; Hirano, 2013). We model this effect by adding $\lambda_2 f_{PF}$ to Equation (6) (Dayan and Abbott, 2001). Although this subtractive normalization term is used to contract the train-induced LTD, it cannot avoid the saturation of $W_{\{PF-PC\}}$ when the error signal is off. In the absence of error signal, the spontaneous activity of PFs is counterbalanced by background activity of CF (Ito, 2013). Thus, we decrease the magnitude of $W_{\{PF-PC\}}$ by considering the effect of spontaneous LTD, which is denoted by $-\lambda_2 LTD_{spon}$. This subtle definition helps us to examine the defects in train-induced and spontaneous LTD, separately. We also add $\lambda_3 W_{\{PF-PC\}}$ as a synaptic decaying factor (Dayan and Abbott, 2001; Gerstner et al., 2014). Eventually, Equation (6) is turned to

$$\dot{W}_{\{PF-PC\}} = -\overbrace{\lambda_1 \alpha f_{MF}\left(df_{MF} - W_{\{MF-DN\}} f_{MF} - W_{\{PF-MLI\}} \alpha f_{MF} + W_{\{PF-PC\}} \alpha f_{MF}\right)}^{train-induced\ LTD} + \overbrace{\lambda_2 \alpha f_{MF}}^{spontaneus\ LTP} - \overbrace{\lambda_2 LTD_{spon}}^{spontaneus\ LTD} - \lambda_3 W_{\{PF-PC\}} \quad (7)$$

where λ_2 and λ_3 are the recovery rates satisfying $\lambda_1 \gg \lambda_2, \lambda_3$. In the rest or post-training period, the OKR gain, which might have amplified with adaptation, tends back to its initial value. Let us indicate this initial gain by $d = d_0$. In this null condition, the synaptic weights return to their baseline ($W^0_{\{PF-PC\}}$, $W^0_{\{PF-MLI\}}$, $W^0_{\{MF-DN\}}$) and the error signal approaches zero. Thus, by setting $\dot{W}_{\{PF-PC\}} = 0$ in the above equation, we derive

$$\lambda_2(\alpha f_{MF} - LTD_{spon}) = \lambda_1 \alpha f_{MF} \overbrace{\left(d_0 f_{MF} - W^0_{\{MF-DN\}} f_{MF} - W^0_{\{PF-MLI\}} \alpha f_{MF} + W^0_{\{PF-PC\}} \alpha f_{MF}\right)}^{e \to 0} + \lambda_3 W^0_{\{PF-PC\}} \quad (8)$$

As seen, the balance between PF-LTD and PF-LTP induced by the spontaneous activity of CF and PFs is equal to the baseline weight of $W_{\{PF-PC\}}$. Replacing $\lambda_2(\alpha f_{MF} - LTD_{spon})$ with $\lambda_3 W^0_{\{PF-PC\}}$ in Equation (7) and considering a regulatory coefficient (β_1) for manipulating the strength of train-induced LTD results in

$$\dot{W}_{\{PF-PC\}} = -\lambda_1 \beta_1 \alpha f_{MF}\left(df_{MF} - W_{\{MF-DN\}} f_{MF} - W_{\{PF-MLI\}} \alpha f_{MF} + W_{\{PF-PC\}} \alpha f_{MF}\right) - \lambda_3 (W_{\{PF-PC\}} - W^0_{\{PF-PC\}}) \quad (9)$$

The other neurons stimulated by PFs are MLIs. As mentioned in the previous section, PF-MLI synapses are suggested to exhibit bidirectional plasticity in the opposite direction to that induced at PF-PC synapses. Coherent activation of PFs and a CF induces LTP, whereas lone stimulation of PFs induces LTD at PF-MLI synapses. Thus, changing the direction of LTP and LTD in Equation (9) yields the following plasticity rule for PF-MLI synapse.

$$\dot{W}_{\{PF-MLI\}} = +\lambda_1 \beta_3 \alpha f_{MF}\left(df_{MF} - W_{\{MF-DN\}} f_{MF} - W_{\{PF-MLI\}} \alpha f_{MF} + W_{\{PF-PC\}} \alpha f_{MF}\right) - \lambda_3 (W_{\{PF-MLI\}} - W^0_{\{PF-MLI\}}) \quad (10)$$

where β_3 is a regulatory coefficient for manipulating the strength of train-induced MLI-LTP and $W^0_{\{PF-MLI\}}$ is the baseline weight of PF-MLI synapse. Given the indirect relation of MLIs and CFs through spillover, the amount of use-dependent synaptic weight changes at PF-MLI synapses is not as large as that at PF-PC synapses. Thus, the default value of β_3 is smaller than β_1. It should be noted that in the third scenario, when we are examining IO lesion, β_1 *and* β_3 are replaced by β_2.

Cerebellar Nuclei

The activity of DN has critical impact on precise modulation of long-term learning. However, there have been few biological studies that explain the nuclear plasticity changes in the face of upstream signals (D'Angelo et al., 2016). Numerical models predict that the potentiation in DN could be driven by PC output (Pugh and Raman, 2006; Garrido et al., 2013a; Solouki et al., 2019). We also know that MFs are the only excitatory inputs that induce LTP in nucleus synapses. Indeed, large activity of MFs together with the small activity of PCs should reasonably induce LTP at MF-DN synapses (Pugh and Raman, 2006). Taking into account the above facts, we model LTP in an associative manner as a general function of pre- and post-synaptic activities in DN ($H(f_{MF}, f_{DN}, f_{PC})$). It is also observed that the sole activation of MFs induces LTD in $W_{\{MF-DN\}}$ (Pugh and Raman, 2006; Hirano, 2013). A subtractive normalization term ($\lambda_5 f_{MF}$) is used to show this effect (Gerstner et al., 2014). The magnitude of $W_{\{MF-DN\}}$ is also constrained by adding a multiplicative decaying factor ($\lambda_6 W_{\{MF-DN\}}$) (Gerstner et al., 2014). Finally, $\dot{W}_{\{MF-DN\}}$ can be written as

$$\dot{W}_{\{MF-DN\}} = \lambda_4 H(f_{MF}, f_{DN}, f_{PC}) - \lambda_5 f_{MF} - \lambda_6 W_{\{MF-DN\}} \quad (11)$$

Presumably, $W_{\{MF-DN\}}$ represents the slow dynamics and evolves much more slowly than $W_{\{PF-PC\}}$. Therefore, we assume $\lambda_4, \lambda_5, \lambda_6 \ll \lambda_1$.

Numerical simulations have revealed that the synaptic weight of MF-DN is liable to induce a kind of Hebbian plasticity with postsynaptic gaiting mechanism (Medina and Mauk, 2000; Kawato, 2009; Mapelli et al., 2015). This gaiting is mainly regulated by two afferents to DN, namely, f_{PC} and f_{MF}. Logically, LTP occurs when f_{MF} is large and f_{PC} is small. This can be denoted in the following way (Gerstner et al., 2014):

$$\Delta W_{\{MF-DN\}} \propto f_{MF}\left(DN - \overline{DN}\right) \ or\ f_{MF}(f^{\max}_{PC} - f_{PC}) \quad (12)$$

where $\overline{DN}$ is the background firing rate (temporal average) of DN and $f^{\max}_{PC}$ is the maximum firing rate of PC. Let us set $H = f_{MF}(f^{\max}_{PC} - f_{PC})$, then, Equation (11) becomes

$$\dot{W}_{\{MF-DN\}} = \lambda_4 f_{MF}\left(f^{\max}_{PC} - W_{\{PF-PC\}} \alpha f_{MF} + W_{\{PF-MLI\}} \alpha f_{MF}\right) - \lambda_5 f_{MF} - \lambda_6 W_{\{MF-DN\}} \quad (13)$$

With the null condition denoted by $\dot{W}_{\{MF-DN\}} = 0$ and ($W^0_{\{PF-PC\}}$, $W^0_{\{PF-MLI\}}$, $W^0_{\{MF-DN\}}$), the subtractive

normalization can be specified as follows:

$$\lambda_5 f_{MF} = \lambda_4 f_{MF}(f_{PC}^{\max} - W^0_{\{PF-PC\}}\alpha f_{MF} + W^0_{\{PF-MLI\}}\alpha f_{MF}) - \lambda_6 W^0_{\{MF-DN\}} \quad (14)$$

Substituting Equation (14) into Equation (13), besides inserting a regulatory coefficient (γ) for manipulating the strength of corticonuclear tract connectivity in the first term of Equation (13), yields.

$$\dot{W}_{\{MF-DN\}} = \lambda_4\gamma\left(W_{\{PF-MLI\}} - W_{\{PF-PC\}} + W^0_{\{PF-PC\}} - W^0_{\{PF-MLI\}}\right)\alpha {f_{MF}}^2 + \lambda_6(W^0_{\{MF-DN\}} - W_{\{MF-DN\}}) \quad (15)$$

LEARNING PROTOCOL AND SIMULATION SETUP

In this study, we utilized a typical learning protocol proposed by Shutoh et al. (2006). They recorded horizontal OKR gain of fixed head mice in response to sustained oscillation of checked-pattern screen. The behavioral data together with the simulation results for the normal case are depicted in **Figure 2**. Learning simulation is conducted in a course of five sessions; each contains 1 h of training and 23 h of post-training off-task period. Long-term (slow) adaptation is estimated by gain differences before training on the first day and after training on each day. Likewise, short-term (fast) adaptation is measured by gain differences before and after 1-h daily training. Other experimental data from Katoh et al. (1998) and Shutoh et al. (2006) are used to verify the results of the model in a lesion case. The former investigated the role of chronic olivary system lesions on short- and long-term adaptation of OKR and the latter investigated the effect of reversible flocculus lesion on the transsynaptic shift in memory trace from cerebellar cortex to DN by injecting lidocaine into the cerebellar cortex of mice immediately after training on the 4th day.

Simulation of lesion scenarios in the current study is carried out by manipulating the embedded coefficients in the updating rules (α, β, and γ) and modifying the corresponding baseline values of synaptic weighs ($W^0_{\{PF-PC\}}$, $W^0_{\{PF-MLI\}}$, and $W^0_{\{MF-DN\}}$). In the first case, we partially attenuate MF-GC transmission by reducing α from 1 to 0.3. In the second case, we implement a multi-stage plan to investigate the role of a) train-induced LTD, b) spontaneous LTD, c) spontaneous LTP, and d) learning and recovery rate at PF-PC synapse, separately. Partial and complete blockade of train-induced PF-PC LTD is done by reducing β_1 from 1 to 0.3 and 0, respectively. According to Equation (7), spontaneous PF-PC LTD deficit disturbs the baseline balance of $W_{\{PF-PC\}}$ in favor of spontaneous LTP and increases $W_{\{PF-PC\}}$. From Equations (12) and (15), the growth of $W_{\{PF-PC\}}$ and subsequent increased activity of PCs takes $W_{\{MF-DN\}}(t)$ to the resting state. Moreover, from Equation (10), the increase of $W_{\{PF-PC\}}$ and the decrease of $W_{\{MF-DN\}}$ enlarge $W_{\{PF-MLI\}}$. Therefore, to simulate the impairment of spontaneous PF-PC LTD, we set $W^0_{\{PF-PC\}} = W^0_{\{PF-MLI\}} = 1.1$ and $W^0_{\{MF-DN\}} = 0$, which were set at 1 in the normal condition. In contrast to spontaneous LTD, impairment of spontaneous PF-PC LTP decreases the synaptic weight of $W_{\{PF-PC\}}$ toward LTD until it stops at 0 (Equation 7). Under this condition, the baseline values of $W_{\{MF-DN\}}$ and $W_{\{PF-MLI\}}$ shift to a lower level (Equations 10 and 15). In practice, vanishing all PF-PC synapses reduces the working frequency and basal intrinsic excitability of PCs, which in turn affects the activity of MLIs and DN (Schonewille et al., 2010). To simulate the impairment of spontaneous PF-PC LTP, we set $W^0_{\{PF-PC\}} = \beta_1 = \lambda_3 = 0$, $\beta_3 = 0.48$, and $W^0_{\{PF-MLI\}} = W^0_{\{MF-DN\}} = 0.5$. Finally, we carry out simulation with a different set of learning and recovery rate to determine the impact of early and late phases of LTD induction on short- and long-term learning and to check the robustness of suggested plasticity rule against parameter variations. In the third case, we blockade CF signaling to cerebellar cortex by replacing β_1 and β_3 with β_2. Partial and complete blockade of CF signal is simulated by setting β_2 at 0.3 and 0, respectively. In the fourth case, we block either plasticity or baseline activity of PF-MLI synapses. In the case of no PF-MLI plasticity, $W_{\{PF-MLI\}}$ remains constant at its baseline level, that is, $W_{\{PF-MLI\}}(t) = W^0_{\{PF-MLI\}}$ for any t. Elimination of PF-MLI baseline activity is also made by setting $W^0_{\{PF-MLI\}} = 0$. This manipulation reinforces the inhibition from PCs to DN and subsequently reduces $W_{\{MF-DN\}}(t)$ so that $W^0_{\{MF-DN\}} = 0$. The overactivation of PCs in the absence of PF-MLI baseline is compensated by decreasing the baseline of $W_{\{PF-PC\}}$ from 1 to 0.4 and setting $\beta_1 = 0.1$. Furthermore, to avoid LTP saturation at PF-MLI synapses, β_3 is set at 0.2. In the last scenario, we block PC-DN transmission by setting $\gamma = 0$ immediately after training on the fourth session of learning.

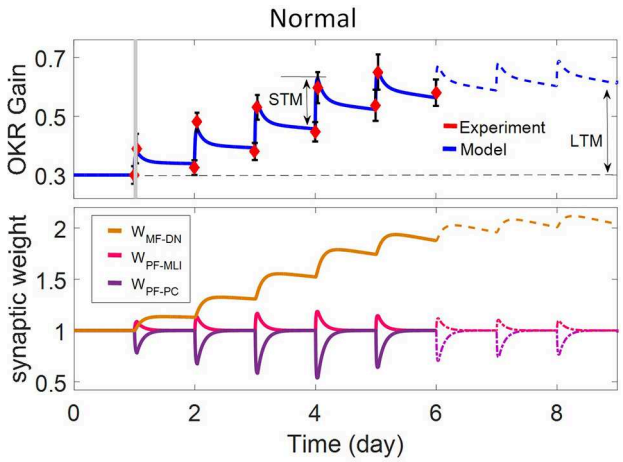


FIGURE 2 | Simulation of OKR adaptation in normal condition. OKR gain changes **(top)** and evolution of synaptic weights at PF-PC, PF-MLI, and MF-DN **(bottom)** during eight sessions of training. The dashed black line indicates the initial level of OKR gain before any training has taken place. The amounts of short- and long-term learning for the 4th and 8th days are indicated by STM and LTM labels. The shaded region specifies the 1-h training period on the first day. The learning curve of the model for the last 3 days, in the absence of experimental data, is plotted by a dashed line.

We adjust the time constants (τ) of induction and recovery periods, which are equal to the inverse of learning rates (λ), based on the electrophysiological study done by Le et al. (2010). They demonstrated that the induction of LTD or LTP at PF-PC synapses occurs within 5–30 min and the recovery from training takes more than 1 day. It is also reported that OKR adaptation is capable to induce a 30–40% decrease in the amplitude of PF excitatory postsynaptic current (PF-EPSC) or the rising slope of PF-EPSP in PCs (Schonewille et al., 2011; Inoshita and Hirano, 2018). Based on the above evidence, a relatively specific and limited space was searched to find the most appropriate set of learning rates for fitting simulation results with the experimental data. Specifically, the values of λ_2 and λ_5 are simply determined by substituting the initial conditions into Equations (7) and (13), respectively. In addition, the baseline values of $W_{\{PF-PC\}}$ and $W_{\{MF-DN\}}$ ($W^0_{\{PF-PC\}}, W^0_{\{MF-DN\}}$) are achieved by the counterbalance between PF-LTD and PF-LTP from Equations (8) and (14).

Based on the fact that the final output of the cerebellum to the oculomotor neurons is provided by DN (**Figure 1B**), a portion of its modular activity, mediated by the upstream synaptic weights $W_{\{PF-PC\}}$, $W_{\{PF-MLI\}}$, and $W_{\{MF-DN\}}$, is associated with the OKR adaptation (Mapelli et al., 2015; Gosui and Yamazaki, 2016). Therefore, a normalization factor is used to transform the actual gain from cerebellar space $\left(W_{\{MF-DN\}} - W_{\{PF-PC\}}\alpha + W_{\{PF-MLI\}}\alpha\right)$ to the real world.

$$G_{OKR} = g_{OKR0}\left(W_{\{MF-DN\}} - W_{\{PF-PC\}}\alpha + W_{\{PF-MLI\}}\alpha\right) \quad (16)$$

where g_{OKR0} is a scaling factor used to adjust the initial amplitude of OKR. Regarding the initial state of the behavioral data of normal case (**Figure 2**), we set $g_{OKR0} = 0.3$. Further details about the model parameters of the simulated scenarios can be found in **Table 1**.

RESULTS

The dynamics of the present model related to the normal and lesion scenarios are plotted in six panels (**Figures 2–5**). Each panel consists of the learning curve(s) of OKR gain at the top and its corresponding time-dependent changes in weights of PF-PC ($W_{\{PF-PC\}}$), PF-MLI ($W_{\{PF-MLI\}}$), and MF-DN ($W_{\{MF-DN\}}$) synapses at the bottom view. Initially, we assess the model capability to reproduce the observed OKR gain adaptation in wild-type mouse (**Figure 2**). It can be seen that the daily repetition of training sessions increased the overall gain step by step from 0.3 to 0.56 after 5 days. When the simulation proceeded for 3 more days, OKR gain further increased at a slower pace until it reached an equilibrium point. The amplitude of OKR increased rapidly during 1 h of sustained daily training and diminished throughout post training or rest period, but did not come back to its previous level. Obviously, the rapid changes are related to the fast dynamic ($W_{\{PF-PC\}}$) and the gradual low changes at the end of each day relative to the start of the same day are attributed to slow dynamic ($W_{\{MF-DN\}}$). The results also support the idea of distributed and synergistic interaction of different forms of plasticity in the cerebellum during a motor learning task (Boyden et al., 2004; Gao et al., 2012; Garrido et al., 2013a; Solouki and Pooyan, 2016). It can be seen that adaptation to a larger gain effectively increases the value of $W_{\{MF-DN\}}$, while it does not change the final value of $W_{\{PF-PC\}}$. Facing the error signal during the training condition ($d > d_0$), $W_{\{PF-PC\}}$ undergoes LTD and then recovers to its initial value in the dark condition. It seems that what has been learned by $W_{\{PF-PC\}}$ during training is partially transferred to $W_{\{MF-DN\}}$. Therefore, $W_{\{PF-PC\}}$ and $W_{\{MF-DN\}}$ are suggested to be responsible for short-term memory formation (STM) and long-term consolidation (LTM), respectively.

$W_{\{PF-MLI\}}$ also changes in sync with $W_{\{PF-PC\}}$ but in the opposite direction. The GABAergic nature of interneurons converses the PF-MLI plasticity to the same direction as the PF-PC plasticity. This increases the possibility of $W_{\{PF-MLI\}}$ participation in STM and compensation of $W_{\{PF-PC\}}$, which is investigated in the next scenarios. Additionally, the depth of modulation in $W_{\{PF-PC\}}$ reaches its maximum level after 3 days. Theoretically, this seems unusual since |e| is the highest on the first day. In practice, however, the long-term gain of OKR increases in an S shape (Shutoh et al., 2006; Takeuchi et al., 2008; Inoshita and Hirano, 2018). We show this effect by adding a regulatory factor to the first term of Equation (9) just for the first 3 days. To simulate OKR adaptation in the lesion conditions, we modified simulation settings including the transmission coefficients highlighted in **Figures 3–5** and baseline values of synaptic weights for compensation, which may occur in behaving animals (see details in **Table 1**).

Complete interruption of the information flow from the cerebellar input layer to the output layer cannot address the reason of the enormous amount of GCs. Therefore, to better understand the evolutionary conservation of this abundant number of cerebellar GCs, a more desirable strategy would be to suppress the neurotransmission from most, but not all, GCs. Moreover, since an experimental elimination of GCs can cause morphological processes in their target neurons (Galliano et al., 2013), it is likely profitable to investigate the functional outcome of inactivating the majority of GCs computationally rather than experimentally. In the first lesion scenario, we have attenuated transmission at MF-GC synapses by reducing α from 1 to 0.3. Based on the model formulation (Equations 9, 10, 15, and 16), the impact of α reduction is not limited to the input layer but propagated in the plasticity rules of the next layers. As a consequence of this manipulation, the overall gain of OKR and plasticity of the synapses significantly declined (**Figure 3**). To provide a tighter comparison of normal and lesion scenarios, we have depicted short- and long-term gain increment of the fifth session of training besides the percentage of synaptic weight changes of the same day in **Figure 6**. Positive and negative long-term gain increments are respectively, considered as the occurrence of motor learning and disruption of basic motor performance. Considering **Figure 6**, it can be seen that even 70% attenuation in MF-GC transmission did not disrupt basic motor performance although motor learning and memory consolidation were compromised. This result is in line with a previous study from Galliano et al. (2013) who noted a minority of functionally intact GCs is adequate for the maintenance of

TABLE 1 | Summary of model parameters for the simulated scenarios.

Model parameters	Simulation scenario									
	Normal	**Attenuation of MF-GC transmission**	**Blockade of train-induced PF-PC LTD**	**Blockade of spontaneous PF-PC LTD**	**Blockade of spontaneous PF-PC LTP**	**Alteration the rate of LTD induction at PF-PC synapse**	**Blockade of CF signal to cerebellar cortex**	**Blockade of PF-MLI plasticity**	**Blockade of PF-MLI baseline activity**	**Blockade of PC-DN transmission**
α	1	0.3	1	1	1	1	1	1	1	1
β_1	1	1	0.3, 0[a]	1	0	1	–	1	0.1	1
β_2	–	–	–	–	–	–	0.3, 0[a]	–	–	–
β_3	0.5	0.5	0.5	0.5	0.48	0.5	–	0	0.2	0.5
γ	1	1	1	1	1	1	1	1	1	0[b]
$W^0_{\{PF-PC\}}$	1	1	1	1.1	0	1	1	1	0.4	1
$W^0_{\{PF-MLI\}}$	1	1	1	1.1	0.5	1	1	1	0	1
$W^0_{\{MF-Dn\}}$	1	1	1	0	0.5	1	1	1	0	1
τ_1^c	0.33	0.33	0.33	0.33	0.33	2	0.33	0.33	0.33	0.33
τ_3^c	2.5	2.5	2.5	2.5	$\gg$2.5[d]	1.5	2.5	2.5	2.5	2.5
τ_4^c	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5

[a] *The coefficients of 0.3 and 0 are applied in the cases of partial and complete blockade, respectively.*

[b] *The value of γ is set at 0 immediately after training on the fourth session.*

[c] *Time constant (τ_i) is equal to the inverse of learning rate (λ_i) and expressed in hours. τ_1 and τ_3 represent time constant of $W_{\{PF-PC\}}$ during learning and recovery period, respectively. The values of τ_4 and τ_6, which represent the time constant of $W_{\{MF-DN\}}$ during and after training, are considered equal.*

[d] *The value of τ_3 can be arbitrarily so large that $\lambda_3 = 0$.*

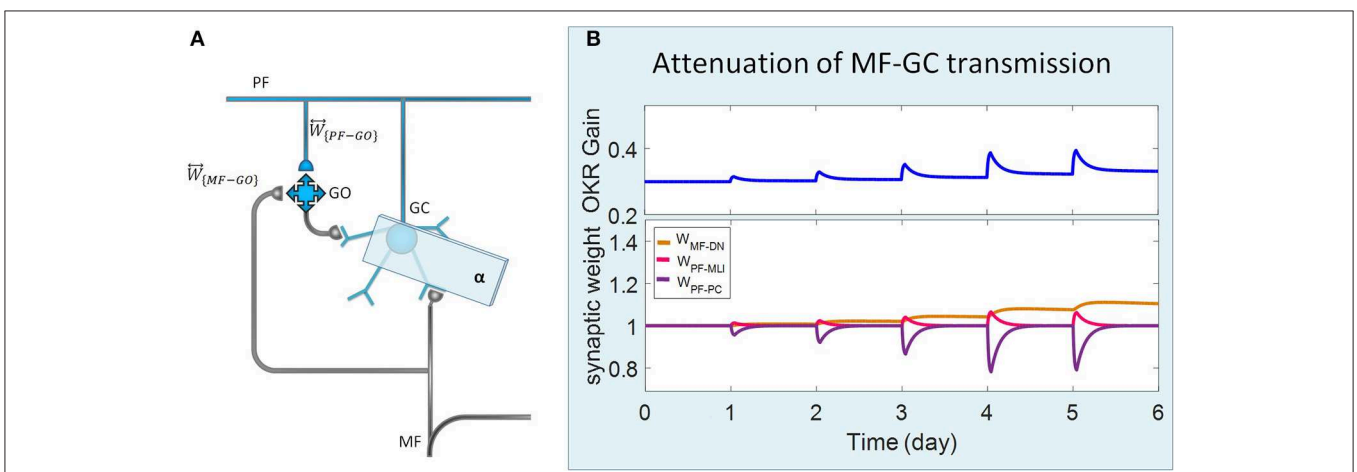


FIGURE 3 | Lesion in the transmission of the granular layer. **(A)** Schematic representation of lesion location in the granular layer. **(B)** Simulated OKR adaptation **(top)** and evolution of synaptic weights **(bottom)** during five sessions of training, while the transmission coefficient of MF-GC synapses (α) is attenuated from 1 to 0.3.

basic motor performance, whereas formation and consolidation of sophisticated memories require higher numbers of normal GCs regulating PC firing.

LTD at PF-PC synapses is believed to be the main cellular mechanism for cerebellar motor learning (Kano et al., 1994; Ito, 2001; Belmeguenai et al., 2008; Hirano, 2013; Yuzaki, 2013; Kuwabara et al., 2014; Freeman, 2015). However, the necessity of LTD for the acquisition of motor learning is challenged by the demonstration of normal motor learning in LTD-deficient mice (Welsh et al., 2005; Faulstich et al., 2006; Schonewille et al., 2011). In response to the dichotomy that has arisen between PF-PC LTD and motor learning, some researchers proposed LTP as an alternative substrate for motor learning (Schonewille et al., 2010), some others have spoken about the possibility of a compensatory pathway (Lemon and Edgley, 2010; Gao et al., 2012), and some found the existing documentation inadequate and postponed the final conclusion to further *in vivo* experiment and theoretical investigations (Hirano, 2013; Ito et al., 2014). In this regard, we conducted a multi-stage lesion scenario to examine the possible involvement of different factors (including

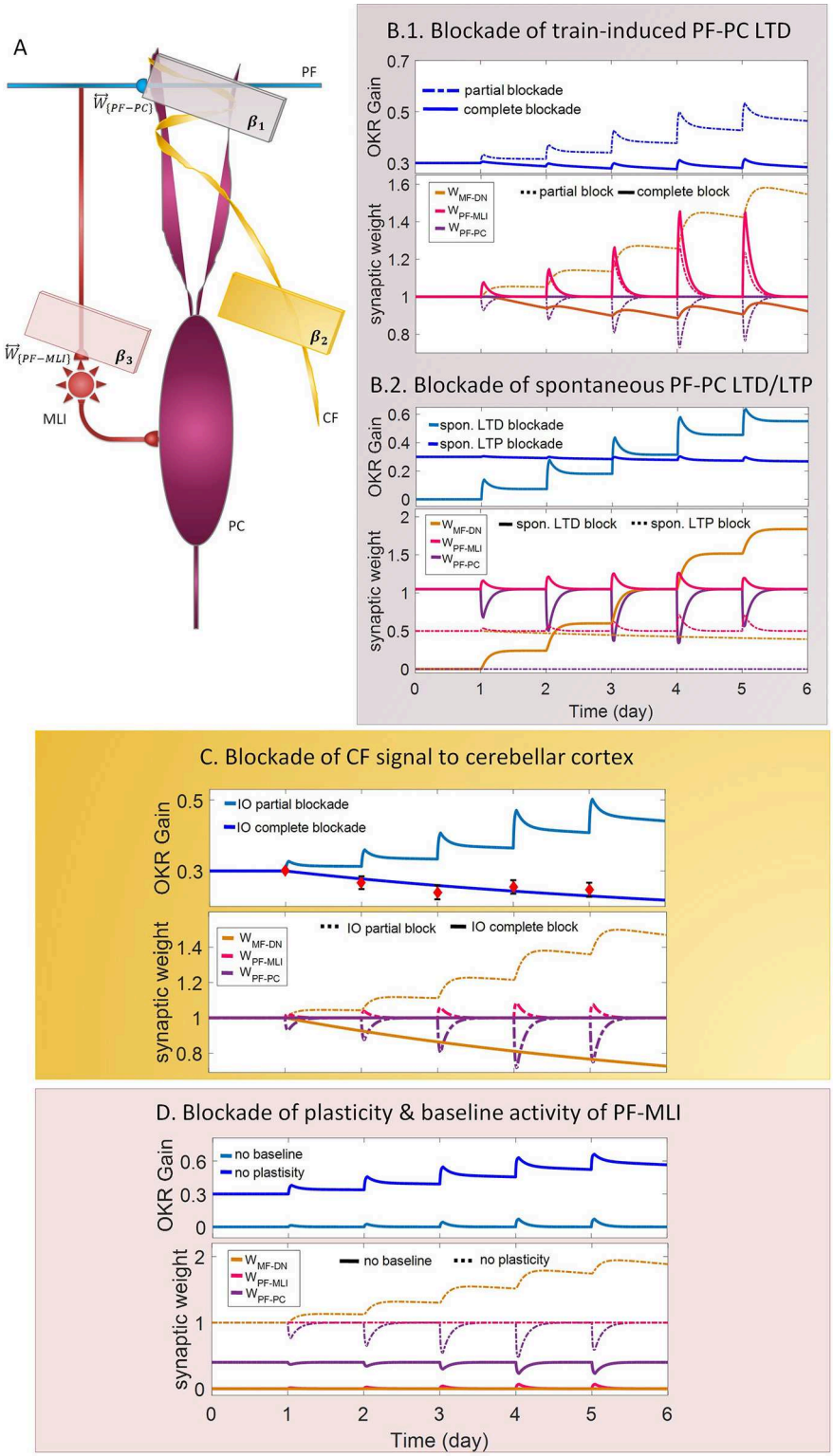


FIGURE 4 | Lesion in the transmission of the molecular and the Purkinje layers. **(A)** Schematic representation of lesion sites and the corresponding transmission coefficients. **(B–D)** Dynamics of the model in gain-up training for 5 days. Each panel consists of the learning curve(s) of OKR gain at the top and its corresponding time-dependent changes in synaptic weights at the bottom. Experimental data in **(C)** are plotted in red dots.

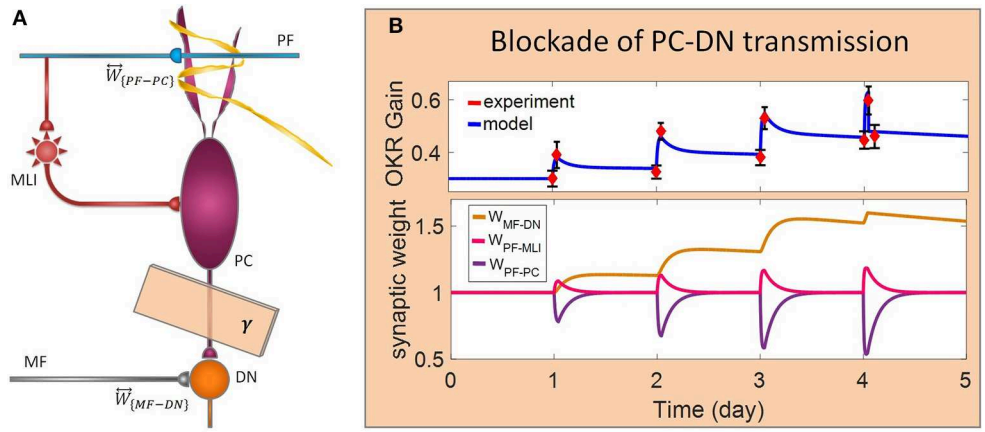


FIGURE 5 | Lesion in the transsynaptic shift of memory trace from cerebellar cortex to DN. **(A)** Schematic representation of lesion location. **(B)** Dynamics of the model in gain-up training for 4 days. Lesion is made by interruption of PC-DN transmission immediately after 1-h training on the 4th day. Experimental data are plotted in red dots.

train-induced LTD, spontaneous LTD, spontaneous LTP, and rate of learning and recovery) in weight changes at PF-PC synapses and motor learning. The impairment of train-induced PF-PC LTD is simulated by partial and complete abrogation of temporal correlation between use-dependent and synchronous activity of CF and PFs (see details in **Table 1**). **Figure 4B.1** compares the OKR gain and synaptic weights in the partial and complete lesion conditions during five sessions of learning. The OKR gain continued to increase at a lower rate than normal in the partial lesion case, whereas it remained largely stopped at the initial state in the complete lesion case. This result indicates that complete blockade of train-induced PF-PC LTD would interrupt motor learning without causing noticeable damage to baseline motor performance. Furthermore, deficiency in train-induced PF-PC LTD led to a 31% increase in PF-MLI synaptic weight in the complete lesion case, which might be interpreted as a confirmation for the compensatory role of PF-MLI pathway for PF-PC synapse. However, despite the major impact of such a compensatory mechanism in upregulating the MF-DN synaptic weight ($W_{\{MF-DN\}}$) and hindering fall from the initial gain state (0.3), it was not that enough to prevent loss of motor learning. Therefore, PF-MLI synapses contribute in providing a more robust basic motor performance rather than compensating motor learning. As shown in **Figure 6**, the blockade of train-induced LTD had a negative impact on both short- and long-term gain increments. Specifically, in the complete lesion case, the slight increase in the short-term gain results from the modulating effect of MLIs on PCs. Therefore, it can be inferred that the molecular layer takes part in short-term memory formation, but not as much as Purkinje layer does. Next, we simulated spontaneous PF-PC LTD deficit by eliminating LTD induced by spontaneous activity of CF and PFs. As a result, the baseline weight of $W_{\{PF-PC\}}$ became larger than normal. The increase in $W_{\{PF-PC\}}$ overactivates PCs and takes $W_{\{MF-DN\}}(t)$ to 0. As shown in **Figure 4B.2**, the initially decreased OKR gain was successfully compensated day by day through overinduction of train-induced LTD. According to **Figure 6**, one of the highest rate of train-induced PF-PC LTD induction with 44% depression in $W_{\{PF-PC\}}$ and the highest rate of short- and long-term gain increment among all the lesion scenarios belong to blockade of spontaneous LTD. This shows the high ability of train-induced PF-PC LTD in compensating spontaneous PF-PC LTD deficit. It seems that the behavioral learning differences between spontaneous and train-induced LTD are the key to explain the discrepancies between cerebellar PF-PC LTD and motor learning. A conventional protocol of PF-PC LTD induction is composed of conjunctive stimulation of PFs and CF at 1 Hz for 5 min or pairing PFs stimulation with direct intracellular depolarization of PC (Hammond et al., 2015); this stimulation corresponds to spontaneous LTD, since both CF and PFs elicit tetanic spikes without temporal modulation, while train-induced LTD is provoked by use-dependent and variable activity of CF and PFs (Schonewille et al., 2011). Then, we simulate spontaneous PF-PC LTP by inactivating LTP induction at PF-PC synapses, which was triggered via the background activity of GCs. This manipulation disturbed the balance of $W_{\{PF-PC\}}$ baseline in favor of spontaneous LTD and decreased $W_{\{PF-PC\}}$ so that it was clamped to 0 and consequently train-induced LTD was also inactivated. In this condition, the OKR gain did not increase after five sessions of learning (**Figure 4B.2**). The long-term gain increment has a slight (non-significant) negative value (**Figure 6**), which indicates the intactness of basic motor performance despite the complete destruction of motor learning and memory consolidation. This result is in line with the experimental observations in calcineurin-deficient mice (Schonewille et al., 2010). The slight formation of STM, in this case, is related to MLI activity.

Next, we carried out the simulation with two different learning and recovery rates for the synaptic weight at PF-PC synapses to specify the impact of early and late phases of LTD on short- and long-term memory and to check the robustness of suggested plasticity rule against parameter variation. As

shown in **Figure 6**, increasing the recovery rate of $W_{\{PF-PC\}}$ reduced long-term gain increment although there was no defect in short-term gain increment. It seems that shortening the duration of late phase of LTD (increasing recovery rate) had a negative impact on transfer and subsequent consolidation of motor learning. On the other hand, decreasing the learning rate of $W_{\{PF-PC\}}$ (lengthening the duration of the early phase of LTD) had a negative impact on short-term memory formation, which is a prerequisite for long-term consolidation. Furthermore, application of different sets of learning or recovery rate did not cause any unrealistic performance overshoot or disruption in the order of LTD/LTP occurrence in $W_{\{PF-PC\}}$ and $W_{\{MF-DN\}}$, which shows the robustness of the suggested plasticity rule against parameter variations.

In the third lesion scenario, we considered partial and complete blockade of CF signal that emanated from IO neurons to the cerebellar cortex by inactivating train-induced plasticity at both PF-PC and PF-MLI synapses. The model ability in reproducing lesioned behavior was successfully verified by experimental data (Katoh et al., 1998; Shutoh et al., 2006) [compare the simulated OKR gain of complete IO lesion with the experiment (**Figure 4C**)]. After lesions were made on the olivary system, the ability to change OKR gain was significantly reduced compared with the normal group. In the complete block case, both short- and long-term gain increment had negative values (**Figure 6**). This suggests that the olivary system (specifically CFs) plays a critical role not only in motor learning but also in basic motor performance. As a confirmation for this view, mutant mice with structural abnormalities in the innervation pattern of PCs by CFs have shown obvious signs of ataxia (Janahmadi et al., 2009; Yuzaki, 2013). Compared with the train-induced PF-PC lesion, complete blockade of CF had a more negative impact on basic motor performance. A possible factor that may have led to such a difference is the lack of train-induced PF-MLI LTP in the CF lesion case. This issue supports this notion that train-induced PF-MLI LTP plays a crucial role in providing a more robust basic motor performance.

Although feedforward inhibition of MLIs onto PCs have been studied extensively, its behavioral relevance has remained enigmatic (Wulff et al., 2009). In the fourth lesion scenario, we investigated how GABAergic neurotransmission from MLIs to PCs regulates adaptation of OKR by conducting the simulation in two modes: (1) no PF-MLI plasticity and (2) no PF-MLI baseline (**Figure 4D**). In the no plasticity mode, we assumed that PF-MLI synaptic weight is fixed at a baseline value ($W^0_{\{PF-MLI\}}$). The resulting OKR gain did not show any considerable difference with that of the normal group. The lack of plasticity in PF-MLI synapse was successfully compensated by about 5% overinduction in train-induced PF-PC LTD. The question that may be posed is that if the lack of plasticity in the PF-MLI synapse does not cause any problems in the learning process, what then is the role of this plasticity? The plasticity that was induced at PF-MLI synapses relies on indirect (spillover) activity of CFs (Gao et al., 2012). Thus, it may have a lower learning capacity and less effective power to modulate the activity of DN and the final output of the cerebellum. However, as predicted earlier, the little memory that forms in this plasticity is used to provide a robust basic motor performance especially when train-induced PF-PC LTD is absent. In the no baseline mode, we set the baseline of $W_{\{PF-MLI\}}$ at 0. Removing the inhibition from PCs increased their excitability and shifted the membrane potential of DN toward a hyperpolarized state. To avoid the overexcitation of PCs and unrealistic variations in PF-MLI plasticity, we decreased the baseline value of $W_{\{PF-PC\}}$ and transmission coefficient of $W_{\{PF-MLI\}}$, respectively. Simulation results showed that the daily increase in OKR gain did not accumulate by the repetition of training sessions (**Figure 4D**). This suggests that blockade of PF-MLI baseline activity impairs memory transfer and consolidation, leaving short-term memory formation intact, consistent with experimental reports (Wulff et al., 2009).

In the last lesion scenario, we blocked PC-DN transmission by setting $\gamma = 0$ immediately after training on the fourth session of learning. An experimental data set (Shutoh et al., 2006) is used to test the ability of the model in reproducing the lesion behavior (**Figure 5**). This manipulation caused almost all that has been learned in the fourth session to be lost and gained back to its previous level. Furthermore, deactivation of the PC-DN pathway eliminates the induction of LTP at the MF-DN synapse, indicating that memory transfer from $W_{\{PF-PC\}}$ to $W_{\{MF-DN\}}$ is disrupted. Normally, the disruption of memory transfer affects long-term learning consolidation. This suggests that the cerebellar cortex plays a critical role not only in short-term memory formation but also in transferring the formed memory to the nuclear region. Furthermore, it can be inferred that OKR gain depends on the modulation strength of PC activity. In an extreme case when all PF–DN synapses die out, PCs cannot modulate their spike discharges, thereby decreasing the OKR gain.

DISCUSSION

In this study, we presented a mechanistic firing rate model of the cerebellar circuit endowed with bidirectional plasticity rules located at PF-PC and PF-MLI synapses and a Hebbian-like plasticity rule with a postsynaptic gating mechanism at MF-DN synapses. Considering the synaptic interaction among the cerebellar layers, the concept of transmission coefficient was proposed as a tool for motor learning analysis. Manipulation of transmission coefficients at five synaptic junctions among cerebellar layers, referred to as lesion scenarios, provided a great potential to answer some of the fundamental questions about the underlying mechanisms involved in cerebellar motor learning and memory formation.

In order to select a modeling strategy in compliance with the objectives of this research, the following points were considered. (*i*) We tried to adhere to the structure and physiology of the cerebellum and, at the same time, preserve the simplicity of the model. This simplicity helped us to provide a panoramic view of the different phases of motor memory formation and to reproduce steady-state changes of learning behavior over a day time scale. (*ii*) The adaptation of OKR

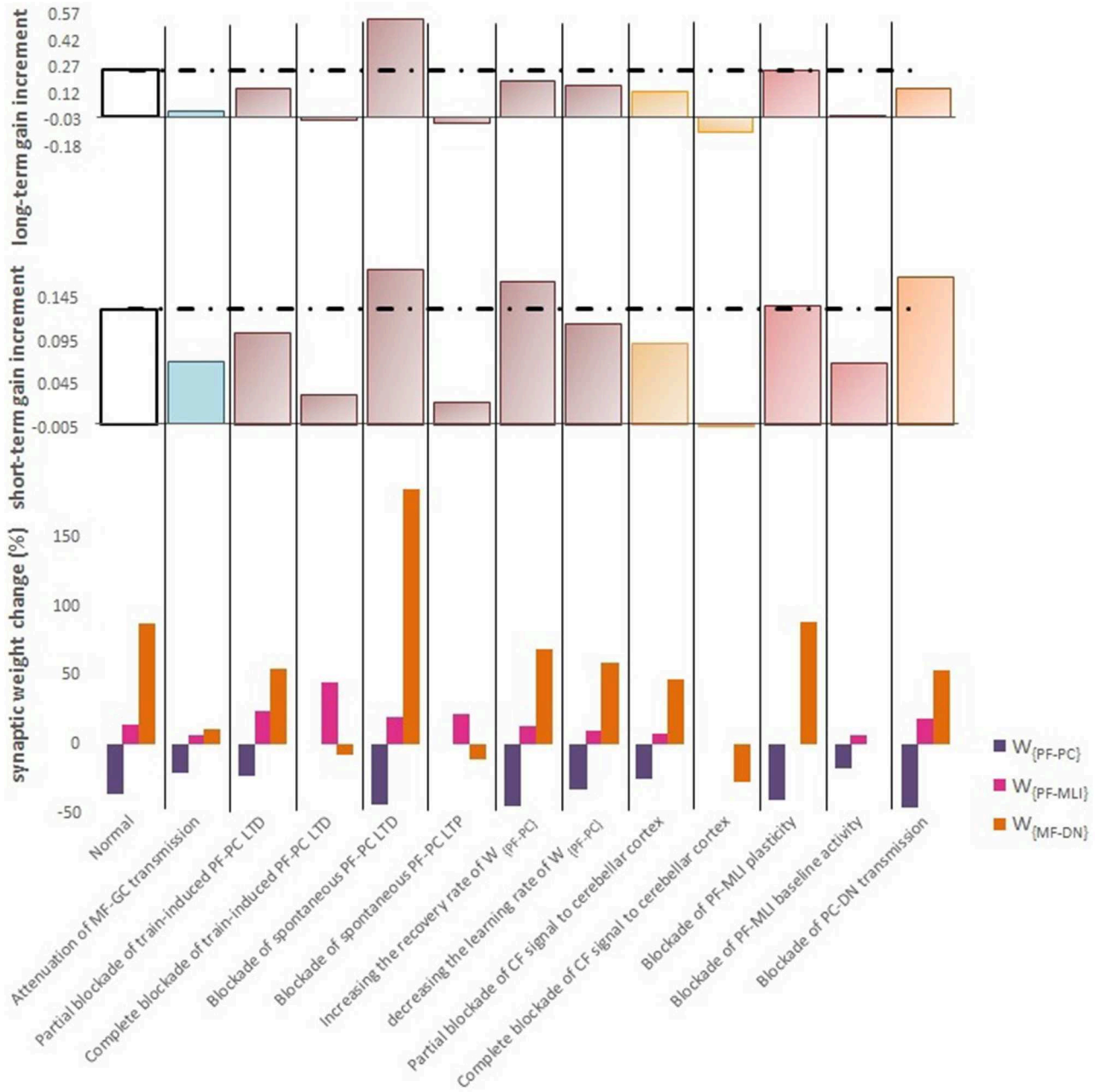


FIGURE 6 | Quantitative data of the model's behavior in confronting different lesion scenarios. **(Top)** Bar graph displaying the long-term or slow adaptation of OKR gain measured by gain differences after training on the 5th day and before training on the first day. Positive and negative long-term gain increments are respectively, considered as the occurrence of motor learning and disruption of basic motor performance. Blockade of CF signaling to cerebellar cortex had the most destructive effect on both motor learning and basic motor performance. Conversely, blockade of PF-MLI plasticity had the least negative impact on the increment of OKR gain. Additionally, blockade of spontaneous PF-PC LTD was completely compensated by overinduction of train-induced LTD. Notably, the more increment in this case does not mean the greater final OKR gain than normal, since the initial state of OKR gain started from 0 rather than 0.3. **(Middle)** Bar graph showing the short-term or fast adaptation of OKR gain measured by gain differences before and after 1 h of training on the 5th day. Dashed lines indicate normal values. **(Bottom)** Bar graph showing the percentage of synaptic weight changes at the fifth session of learning.

gain as a cerebellar-dependent task is used to make a more straightforward relationship between the neural and behavioral domains. This improves the interpretability of behavioral disorders caused by cerebellar malfunction and minimizes the possible involvement of other brain regions in the control of movement. (*iii*) Synaptic plasticity rules were defined as a function of background and use-dependent activities of the neuronal population. Distinguishing between the underlying

factors involved in long-term synaptic weight modification helped us to further elucidate their contribution to the learning process. (*iv*) The compatibility of the simulation results with experiments was examined. The results showed good accordance with the findings of previous studies. (*v*) The functional contribution of interneurons in motor learning was regarded. (*vi*) The possibility of a compensatory mechanism in different lesion scenarios was investigated.

The repetitive alternation between task and off-task period during gain-up training for five consecutive sessions effectively increased the OKR gain by 90%. Meanwhile, the model could show the cooperative action of the main forms of synaptic plasticity at PF-PC, PF-MLI, and MF-DN synapses during learning. Then, we tested the model performance under various lesion conditions. The percentage of changes in synaptic weights and the extent of gain increment relative to the initial state were used as a measure for evaluating and comparing different conditions. The model predicted that train-induced PF-PC LTD is the most effective compensation mechanism since it could successfully cover the lack of spontaneous PF-PC LTD and PF-MLI plasticity. In addition, train-induced PF-PC LTD had a major role in the learning capability of the model since its blockade completely stopped motor learning and memory formation, although it did not disrupt basic motor performance. In this regard, the overinduction of PF-MLI LTP was not able to prevent motor learning impairment, despite its vital role in preserving the robustness of basic motor performance. The pivotal role of PF-PC LTD in cerebellar learning is highlighted in several studies, including those evaluating the animal models of autistic spectral disorder (ASD) (Koekkoek et al., 2005; Baudouin et al., 2012; Piochon et al., 2014). For instance, the incomplete elimination of surplus CFs in patDp/+ mice causes PF-PC LTD saturation and motor learning impairment (Piochon et al., 2014). Similarly, deletion of Fmr1 gene in PCs (L7-Fmr1; a model for syndromic autism) attenuates eyeblink conditioning by altering LTD profile (Koekkoek et al., 2005). ASD models with environmental etiology also support the involvement of synaptic transmission of PCs in motor learning (Wang et al., 2018). Valproic acid (VPA) is an anti-seizure drug that can replicate ASD-like symptoms. Administration of VPA effectively reduces the cell density and dendritic arborization of PCs. Such an anatomical defect is associated with suppressed synaptic transmission of PCs and impaired motor learning in VPA-treated mice (Wang et al., 2018). As another lesion case, blockade of spontaneous PF-PC LTP clamped $W_{\{PF-PC\}}$ to 0 and subsequently disabled motor learning. Of course, as in the previous case, this manipulation did not show abnormal motor performance. Thus, it can be inferred that both train-induced LTD and spontaneous LTP are required for motor learning. If the motor performance is considered as a consequence of postnatal long-term learning (Manto and Jissendi, 2012), then the birth defect in train-induced LTD or spontaneous LTP would cause severe impairment in motor performance. Compared to LTD, fewer studies have focused on the impact of PF-LTP on motor learning (Yuzaki, 2013). L7-PP2B mutant mouse is one of the animal models created for this purpose (Schonewille et al., 2010). Consistent with our results, this mutant showed impaired acquisition of eyeblink conditioning and impaired gain adaptation of vestibulo-ocular–reflex. The basal motor performance remained unaffected. Another point about the

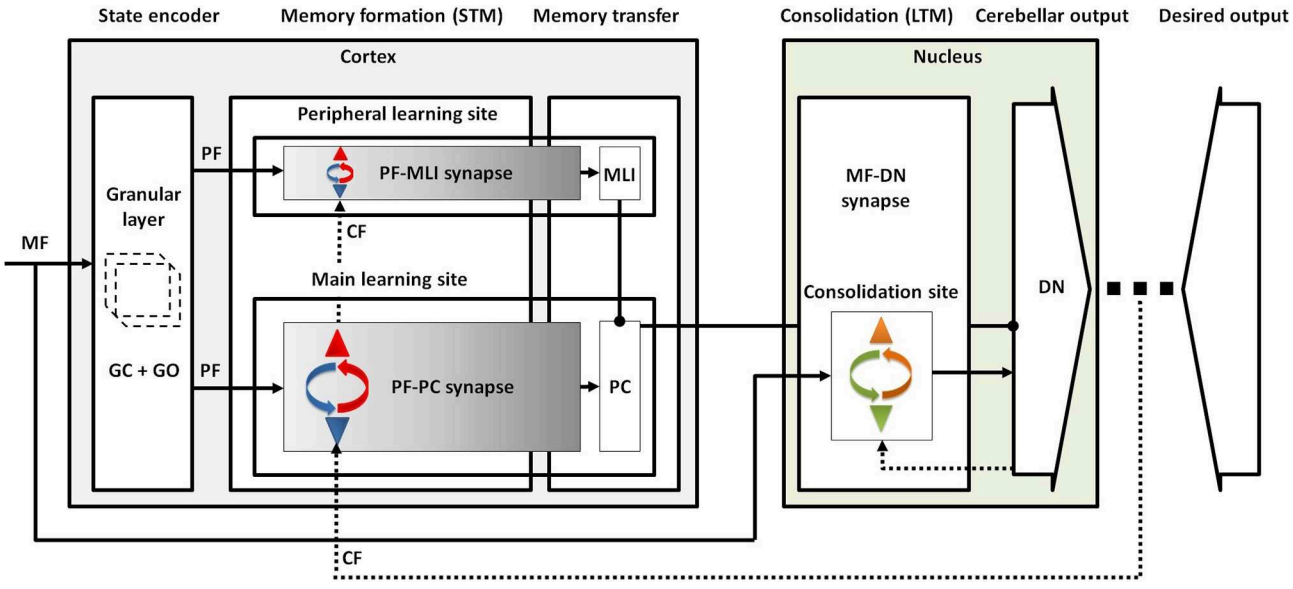


FIGURE 7 | Conceptual map of procedural motor memory formation in the cerebellum. Sensory input is transmitted to the cortical and nuclear modules by MFs. The granular layer, benefiting from abundant clusters of granule and Golgi cells, conveys the encoded sensory information to both molecular and Purkinje layers by PFs. Conjunctive activation of PFs and CF forms short-term memory at the main and peripheral learning sites. Then, the formed memory is transferred to the nuclear module through inhibition exerted by PCs on DN mostly at the late phase of LTD. Finally, the postsynaptic activity of DN drives the formation of long-term memory at the consolidation site.

mutant is that, despite the abrogation of LTP, LTD functions normally. This may appear to be in conflict with our results and the notion of mutual dependence between LTD and LTP suggested by several studies (Baudouin et al., 2012; Yuzaki, 2013). However, our model explained this contradiction by differentiating between train-induced and spontaneous LTD and predicted that the reported LTD, induced by frequency constant stimulus, was spontaneous and not train-induced. Among all the lesion cases, IO lesion, which was referred to as blockade of CF signaling to cerebellar cortex, had the most destructive effect on both motor learning and basic motor performance since it would result in the inactivation of all train-induced mechanisms mediated by CF error signal not just LTD in PF-PC, but also LTP in PF-MLI synapses. Thus, studies on transgenic animals that suffer from IO lesion or CFs abnormalities may not be an exact representation of LTD defect. As another case, the lesion caused by 70% attenuation of MF-GC transmission did not disrupt basic motor performance although it compromised motor learning and memory consolidation. Based on our formulation, manipulating the transmission coefficient in the input layer reduced the plasticity of synaptic weights in the next layers and even the inner nuclear layer. These observations provide important clues from a dynamical viewpoint, as to why vertebrates require an extreme abundance of functional GCs in their life and to what extent this abundance is redundant. The functionality of GABAergic interneurons in regulating the adaptation of OKR was investigated in two different conditions: (1) no PF-MLI plasticity and (2) no PF-MLI baseline. The model predicted that the lack of PF-MLI plasticity is completely compensable through overinduction of train-induced PF-PC LTD. Furthermore, the indirect access of MLIs to instruction signals from CFs and the final output of the cerebellum reduce their learning and modulating capability. The necessity of PF-MLI plasticity was more tangible when the train-induced PF-PC LTD had defected. In such a circumstance, PF-MLI plasticity acted as a reserve mechanism, using its limited short-term memory formation capacity to save the basic motor performance from impairment. On the other hand, lack of PF-MLI baseline impaired memory transfer and consolidation, while leaving short-term memory formation intact. Simulation of the learning process with different sets of learning and recovery rates asserted that the early phase of LTD participates in short-term memory formation and the late phase of LTD is involved in memory transfer. The possibility of a hierarchical causality in transferring the acquired short-term memory across the cortical and its partial consolidation into the nuclear parts of the cerebellum was investigated by cutting off the connection of flocculus with the DN during the training. After lesions were made, OKR gain returned to its pretraining level and MF–DN synaptic weight failed to further increase, whereas PF–PC synaptic weight remained intact, indicating disruption of memory transfer and consequently long-term memory consolidation. The existence of such a biphasic mechanism, with a first rapid phase of memory formation in the cerebellar cortex and a slower phase of memory consolidation in the cerebellar nuclei, is also observed in the spiking models of the cerebellum (Antonietti et al., 2016, 2017; Luque et al., 2016). These models showed that the functional interplay among distributed plasticity sites could facilitate the slow memory consolidation at the nuclear sites, which is important for faster and more stable reacquisition of associative motor tasks. Recently, a distributed spiking model of the cerebellum is used to predict the impact of different pathological damages of the cerebellar cortex on the acquisition of eyeblink conditioning (Geminiani et al., 2018). The pathological cases consistently indicated that (1) an intact functionality of cerebellar cortex is needed to accelerate memory transfer to DN; (2) partial impairment in MF afferents results in an imperfect transmission of information to the granular layer, which in turn influences the activity of PCs and compromises learning; and (3) partial damage in PF-PC LTD does not completely stop learning, but decreases its velocity.

Taking all the obtained results from lesion scenarios together, we proposed a conceptual map for procedural motor memory formation in the cerebellum (**Figure 7**). Accordingly, the sequential processes of cerebellar motor learning occur in the following way. (1) Conjunctive activation of PFs and CF during training sessions causes short-term memory formation at PF-PC synapses (main learning site) and to a lesser extent at PF-MLI synapses (peripheral learning sites). (2) After training, the activity of DN is modulated to a higher level than before training, since the inhibition exerted by PCs is weakened by PF-PC LTD and PF-MLI LTP. (3) This enhanced DN activity resulted in LTP at MF-DN synapses by a Hebbian-like postsynaptic mechanism. (4) Meanwhile, PF-PC synapses (along with PF-MLI synapses) recover from LTD steadily by spontaneous PF-PC LTP, and this causes a slow decrease in DN activity. Thus, the late phase of LTD provides enough time to transfer memory from PF-PC to MF-DN synapse. (5) The recovery from PF-PC LTD and PF-MLI LTP erases the learned short-term memory, whereas (6) the slow decrease of DN activity acts to consolidate learning at MF-DN synapses. When the train-induced PF-PC LTD defects, PF-MLI plasticity acts as a backup mechanism, using its limited short-term memory formation capacity to save the basic motor performance from impairment.

It should be noted that, although long-term synaptic plasticity has been proposed to be the dominant form of plasticity for cerebellar motor learning, various types of plastic mechanisms can occur at multiple synaptic or even non-synaptic sites of the network that provide supplementary and overlapping functions (Gao et al., 2012). For instance, there is accumulating evidence supporting the involvement of short-term and intrinsic plasticity in motor learning and memory formation (Hirano, 2018; Shim et al., 2018). However, the exact functionality of such phenomena remains to be shown. Furthermore, the transient nature of the short-term plasticity and the non-synaptic (non-local) character of the intrinsic plasticity make the tracking of their effects harder. Therefore, analyzing the steady-state behavior of the system, we ignore the short-term plasticity and mainly focus on the effect of long-term synaptic changes on cerebellar learning. We also implicitly take into account the role of the intrinsic plasticity in determining the optimum working frequency and basal intrinsic excitability of PCs by modifying the corresponding baseline values of synaptic weights in the lesion conditions.

The selected modeling strategy has a great ability in displaying the equilibrium behavior of the system, even in the post training period. However, this simplified model, working with analog variables representing the average firing rates of the different neuronal populations, encounters limitations in showing the spatiotemporal evolution of spike patterns of the principal neurons and the diversity of delays imposed by the interneurons. Equipping the model with the spike-timing-dependent plasticity (STDP) mechanisms allows us to match learning with the network temporal dynamic and to simulate more complex phase-varying tasks.

In future work, the impact of different learning paradigms on the efficacy of learning can be evaluated. Also, to gain further insight into the cellular and molecular bases of motor learning, it is necessary to extend the compartments of the model to mimic the non-linear behavior of cerebellar network more precisely. In this study, we considered gain adaptation of OKR as a representative of cerebellar learning. Certainly, the study of model behavior in the implementation of other cerebellar-dependent tasks can help to uncover more features of motor learning in the cerebellum.

Altogether, the proposed model, taking the efficacy of neurotransmission among different layers of the cerebellum into account implicitly, provides a computational basis toward evaluating multiple synaptic plasticity in cerebellar learning and memory.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

AUTHOR CONTRIBUTIONS

SS, FB, and MJ conceived the work. SS developed the codes and performed the computations. SS, FB, and MJ analyzed and interpreted the results. SS wrote the manuscript in consultation with FB and MJ.

REFERENCES

Antonietti, A., Casellato, C., D'Angelo, E., and Pedrocchi, A. (2017). Model-driven analysis of eyeblink classical conditioning reveals the underlying structure of cerebellar plasticity and neuronal activity. *IEEE Trans. Neural Networks Learn. Syst*. 28, 2748–2762. doi: 10.1109/TNNLS.2016.2598190

Antonietti, A., Casellato, C., Garrido, J. A., Luque, N. R., Naveros, F., Ros, E., et al. (2016). Spiking neural network with distributed plasticity reproduces cerebellar learning in eye blink conditioning paradigms. *IEEE Trans. Biomed. Eng.* 63, 210–219. doi: 10.1109/TBME.2015.2485301

Armano, S., Rossi, P., Taglietti, V., and D'Angelo, E. (2000). Long-term potentiation of intrinsic excitability at the mossy fiber–granule cell synapse of rat cerebellum. *J. Neurosci*. 20, 5208–5216. doi: 10.1523/JNEUROSCI.20-14-05208.2000

Badura, A., Clopath, C., Schonewille, M., and De Zeeuw, C. I. (2016). Modeled changes of cerebellar activity in mutant mice are predictive of their learning impairments. *Sci. Rep.* 6:36131. doi: 10.1038/srep36131

Badura, A., Schonewille, M., Voges, K., Galliano, E., Renier, N., Gao, Z., et al. (2013). Climbing fiber input shapes reciprocity of purkinje cell firing. *Neuron* 78, 700–713. doi: 10.1016/j.neuron.2013.03.018

Baudouin, S. J., Gaudias, J., Gerharz, S., Hatstatt, L., Zhou, K., Punnakkal, P., et al. (2012). Shared synaptic pathophysiology in syndromic and nonsyndromic rodent models of autism. *Science* 338, 128–132. doi: 10.1126/science.1224159

Belmeguenai, A., Botta, P., Weber, J. T., Carta, M., De Ruiter, M., De Zeeuw, C. I., et al. (2008). Alcohol impairs long-term depression at the cerebellar parallel fiber-Purkinje cell synapse. *J. Neurophysiol.* 100, 3167–3174. doi: 10.1152/jn.90384.2008

Borjkhani, M., Bahrami, F., and Janahmadi, M. (2018). Formation of opioid-induced memory and its prevention: a computational study. *Front. Comput. Neurosci.* 12:63. doi: 10.3389/fncom.2018.00063

Boyden, E. S., Katoh, A., and Raymond, J. L. (2004). Cerebellum-dependent learning: the role of multiple plasticity mechanisms. *Annu. Rev. Neurosci.* 27, 581–609. doi: 10.1146/annurev.neuro.27.070203.144238

Carrel, A. J. (2012). *The role of glutamatergic inferior olivary teaching signals in the acquisition of conditioned eyeblinks* (Graduate Theses and Dissertations). Iowa State University, Ames, IA, United States.

Chadderton, P., Margie, T. W., and Häusser, M. (2004). Integration of quanta in cerebellar granule cells during sensory processing. *Nature* 428:856. doi: 10.1038/nature02442

Cheron, G., Dan, B., and Márquez-Ruiz, J. (2013). Translational approach to behavioral learning: lessons from cerebellar plasticity. *Neural Plast.* 2013:853654. doi: 10.1155/2013/853654

Coesmans, M., Weber, J. T., De Zeeuw, C. I., and Hansel, C. (2004). Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control. *Neuron* 44, 691–700. doi: 10.1016/j.neuron.2004.10.031

D'Angelo, E., Antonietti, A., Casali, S., Casellato, C., Garrido, J. A., Luque, N. R., et al. (2016). Modeling the cerebellar microcircuit: new strategies for a long-standing issue. *Front. Cell. Neurosci.* 10:176. doi: 10.3389/fncel.2016.00176

D'Angelo, E., and De Zeeuw, C. I. (2009). Timing and plasticity in the cerebellum: focus on the granular layer. *Trends Neurosci.* 32, 30–40. doi: 10.1016/j.tins.2008.09.007

D'Angelo, E., Nieus, T., Bezzi, M., Arleo, A., and Coenen, O. J.-M. D. (2005). Modeling synaptic transmission and quantifying information transfer in the granular layer of the cerebellum. *Comput. Intell. Bioinspired Syst.* 3512, 107–114. doi: 10.1007/11494669_14

Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

De Zeeuw, C. I., and Ten Brinke, M. M. (2015). Motor learning and the cerebellum. *Cold Spring Harb. Perspect. Biol.* 7:a021683. doi: 10.1101/cshperspect.a021683

Evans, G. J. (2007). Synaptic signalling in cerebellar plasticity. *Biol. Cell* 99, 363–378. doi: 10.1042/BC20070010

Faulstich, M., van Alphen, A. M., Luo, C., du Lac, S., and De Zeeuw, C. I. (2006). Oculomotor plasticity during vestibular compensation does not depend on cerebellar LTD. *J. Neurophysiol.* 96, 1187–1195. doi: 10.1152/jn.00045.2006

Freeman, J. H. (2015). Cerebellar learning mechanisms. *Brain Res.* 1621, 260–269. doi: 10.1016/j.brainres.2014.09.062

Galliano, E., Gao, Z., Schonewille, M., Todorov, B., Simons, E., Pop, A. S., et al. (2013). Silencing the majority of cerebellar granule cells uncovers their essential role in motor learning and consolidation. *Cell Rep.* 3, 1239–1251. doi: 10.1016/j.celrep.2013.03.023

Gao, Z., Beugen, B. J., and Zeeuw, C. I. (2012). Distributed synergistic plasticity and cerebellar learning. *Nat. Rev. Neurosci.* 13, 619–635. doi: 10.1038/nrn3312

Garrido, J., Luque, N. R., D'Angelo, E., and Ros, E. (2013a). Distributed cerebellar plasticity implements adaptable gain control in a manipulation task: a closed-loop robotic simulation. *Front. Neural Circuits* 7:159. doi: 10.3389/fncir.2013.00159

Garrido, J., Ros, E., and D'Angelo, E. (2013b). Spike timing regulation on the millisecond scale by distributed synaptic plasticity at the

cerebellum input stage: a simulation study. *Front. Comput. Neurosci.* 7:64. doi: 10.3389/fncom.2013.00064

Geminiani, A., Casellato, C., Antonietti, A., D'Angelo, E., and Pedrocchi, A. (2018). A multiple-plasticity spiking neural network embedded in a closed-loop control system to model cerebellar pathologies. *Int. J. Neural Syst*. 28:1750017. doi: 10.1142/S0129065717500174

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics. From Single Neurons to Networks and Cognition*. Cambridge: Cambridge University Press.

Gosui, M., and Yamazaki, T. (2016). Real-world-time simulation of memory consolidation in a large-scale cerebellar model. *Front. Neuroanat*. 10:21. doi: 10.3389/fnana.2016.00021

Hammond, C., Goaillard, J. M., Debanne, D., and Gaiarsa, J. L. (2015). *Cellular and Molecular Neurophysiology*. London: Academic Press.

Hirano, T. (2013). Long-term depression and other synaptic plasticity in the cerebellum. *Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.* 89, 183–195. doi: 10.2183/pjab.89.183

Hirano, T. (2018). Regulation and interaction of multiple types of synaptic plasticity in a Purkinje neuron and their contribution to motor learning. *Cerebellum* 17, 756–765. doi: 10.1007/s12311-018-0963-0

Inoshita, T., and Hirano, T. (2018). Occurrence of long-term depression in the cerebellar flocculus during adaptation of optokinetic response. *eLife* 7:e36209. doi: 10.7554/eLife.36209

Ito, M. (2000). Mechanisms of motor learning in the cerebellum. *Brain Res*. 886, 237–245. doi: 10.1016/S0006-8993(00)03142-5

Ito, M. (2001). Cerebellar long-term depression: characterization, signal transduction, and functional roles. *Physiol. Rev*. 81, 1143–1195. doi: 10.1152/physrev.2001.81.3.1143

Ito, M. (2013). Error detection and representation in the olivo-cerebellar system. *Front. Neural Circuits* 7:1. doi: 10.3389/fncir.2013.00001

Ito, M., Yamaguchi, K., Nagao, S., and Yamazaki, T. (2014). Long-term depression as a model of cerebellar plasticity. *Prog. Brain Res*. 210, 1–30. doi: 10.1016/B978-0-444-63356-9.00001-7

Janahmadi, M., Goudarzi, I., Kaffashian, M. R., Behzadi, G., Fathollahi, Y., and Hajizadeh, S. (2009). Co-treatment with riluzole, a neuroprotective drug, ameliorates the 3-acetylpyridine-induced neurotoxicity in cerebellar Purkinje neurones of rats: behavioural and electrophysiological evidence. *Neurotoxicology* 30, 393–402. doi: 10.1016/j.neuro.2009.02.014

Jörntell, H., and Hansel, C. (2006). Synaptic memories upside down: bidirectional plasticity at cerebellar parallel fiber-Purkinje cell synapses. *Neuron* 52, 227–238. doi: 10.1016/j.neuron.2006.09.032

Kano, M., Aiba, A., and Tonegawa, S. (1994). Deficient cerebellar long-term depression in metabotropic glutamate receptor-1 (mGluR1) mutant mice. *Neurosci. Res. Suppl.* 19:S34. doi: 10.1016/0921-8696(94)92354-X

Katoh, A., Kitazawa, H., Itohara, S., and Nagao, S. (1998). Dynamic characteristics and adaptability of mouse vestibulo-ocular and optokinetic response eye movements and the role of the flocculo-olivary system revealed by chemical lesions. *Proc. Natl. Acad. Sci. U.S.A*. 95, 7705–7710. doi: 10.1073/pnas.95.13.7705

Kawato, M. (2009). Cerebellum: models. *Encyclopedia Neurosci*. 2, 757–767. doi: 10.1016/B978-008045046-9.01307-3

Kellett, D. O., Fukunaga, I., Chen-Kubota, E., Dean, P., and Yeo, C. H. (2010). Memory consolidation in the cerebellar cortex. *PLoS ONE* 5:e11737. doi: 10.1371/journal.pone.0011737

Koekkoek, S. K., Yamaguchi, K., Milojkovic, B. A., Dortland, B. R., Ruigrok, T. J., Maex, R., et al. (2005). Deletion of FMR1 in purkinje cells enhances parallel fiber LTD, enlarges spines, and attenuates cerebellar eyelid conditioning in fragile X syndrome. *Neuron* 47, 339–352. doi: 10.1016/j.neuron.2005.07.005

Kuwabara, Y., Ishizeki, M., Watamura, N., Toba, J., Yoshii, A., Inoue, T., et al. (2014). Impairments of long-term depression induction and motor coordination precede A? Accumulation in the cerebellum of APPswe/PS1dE9 double transgenic mice. *J. Neurochem.* 130, 432–443. doi: 10.1111/jnc.12728

Le, T. D., Shirai, Y., Okamoto, T., Tatsukawa, T., Nagao, S., Shimizu, T., et al. (2010). Lipid signaling in cytosolic phospholipase A2α-cyclooxygenase-2 cascade mediates cerebellar long-term depression and motor learning. *Proc. Natl. Acad. Sci. U.S.A*. 107, 3198–3203. doi: 10.1073/pnas.0915020107

Lee, K. H., Mathews, P. J., Reeves, A. M., Choe, K. Y., Jami, S. A., Serrano, R. E., et al. (2015). Circuit mechanisms underlying motor memory formation in the cerebellum. *Neuron* 86, 529–540. doi: 10.1016/j.neuron.2015.03.010

Lemon, R. N., and Edgley, S. A. (2010). Life without a cerebellum. *Brain* 133, 652–654. doi: 10.1093/brain/awq030

Lev-Ram, V., Wong, S. T., Storm, D. R., and Tsien, R. Y. (2002). A new form of cerebellar long-term potentiation is postsynaptic and depends on nitric oxide but not cAMP. *Proc. Natl. Acad. Sci. U.S.A*. 99, 8389–8393. doi: 10.1073/pnas.122206399

Luque, N. R., Garrido, J. A., Naveros, F., Carrillo, R. R., D'Angelo, E., and Ros, E. (2016). Distributed cerebellar motor learning: a spike-timing-dependent plasticity model. *Front. Comput. Neurosci.* 10:17. doi: 10.3389/fncom.2016.00017

Manto, M. U., and Jissendi, P. (2012). Cerebellum: links between development, developmental disorders and motor learning. *Front. Neuroanat*. 6:1. doi: 10.3389/fnana.2012.00001

Mapelli, L., Pagani, M., Garrido, J. A., and D'Angelo, E. (2015). Integrated plasticity at inhibitory and excitatory synapses in the cerebellar circuit. *Front. Cell. Neurosci*. 9:169. doi: 10.3389/fncel.2015.00169

Marzban, H., Del Bigio, M. R., Alizadeh, J., Ghavami, S., Zachariah, R. M., and Rastegar, M. (2015). Cellular commitment in the developing cerebellum. *Front. Cell. Neurosci*. 8:450. doi: 10.3389/fncel.2014.00450

Mauk, M. D., Medina, J. F., Nores, W. L., and Ohyama, T. (2000). Cerebellar function: coordination, learning or timing? *Curr. Biol*. 10, R522–R525. doi: 10.1016/S0960-9822(00)00584-4

Medina, J. F., and Mauk, M. D. (2000). Computer simulation of cerebellar information processing. *Nat. Neurosci.* 3, 1205–1211. doi: 10.1038/81486

Okamoto, T., Endo, S., Shirao, T., and Nagao, S. (2011). Role of cerebellar cortical protein synthesis in transfer of memory trace of cerebellum-dependent motor learning. *J. Neurosci*. 31, 8958–8966. doi: 10.1523/JNEUROSCI.1151-11.2011

Pijpers, A., Voogd, J., and Ruigrok, T. J. (2005). Topography of olivo-cortico-nuclear modules in the intermediate cerebellum of the rat. *J. Comp. Neurol*. 492, 193–213. doi: 10.1002/cne.20707

Piochon, C., Kloth, A. D., Grasselli, G., Titley, H. K., Nakayama, H., Hashimoto, K., et al. (2014). Cerebellar plasticity and motor learning deficits in a copy-number variation mouse model of autism. *Nat. Commun*. 5:5586. doi: 10.1038/ncomms6586

Pugh, J. R., and Raman, I. M. (2006). Potentiation of mossy fiber EPSCs in the cerebellar nuclei by NMDA receptor activation followed by postinhibitory rebound current. *Neuron* 51, 113–123. doi: 10.1016/j.neuron.2006.05.021

Schonewille, M., Belmeguenai, A., Koekkoek, S. K., Houtman, S. H., Boele, H. J., van Beugen, B. J., et al. (2010). Purkinje cell-specific knockout of the protein phosphatase PP2B impairs potentiation and cerebellar motor learning. *Neuron* 67, 618–628. doi: 10.1016/j.neuron.2010.07.009

Schonewille, M., Gao, Z., Boele, H. J., Vinueza Veloz, M. F., Amerika, W. E., Šimek, A. A., et al. (2011). Reevaluating the role of LTD in cerebellar motor learning. *Neuron* 70, 43–50. doi: 10.1016/j.neuron.2011.02.044

Seja, P., Schonewille, M., Spitzmaul, G., Badura, A., Klein, I., Rudhard, Y., et al. (2012). Raising cytosolic Cl^- in cerebellar granule cells affects their excitability and vestibulo-ocular learning. *EMBO J*. 31, 1217–1230. doi: 10.1038/emboj.2011.488

Sgritta, M., Locatelli, F., Soda, T., Prestori, F., and D'Angelo, E. U. (2017). Hebbian spike-timing dependent plasticity at the cerebellar input stage. *J. Neurosci*. 37, 2809–2823. doi: 10.1523/JNEUROSCI.2079-16.2016

Shim, H. G., Lee, Y.-S., and Kim, S. J. (2018). The emerging concept of intrinsic plasticity: activity-dependent modulation of intrinsic excitability in cerebellar Purkinje cells and motor learning. *Exp. Neurobiol*. 27, 139–154. doi: 10.5607/en.2018.27.3.139

Shutoh, F., Ohki, M., Kitazawa, H., Itohara, S., and Nagao, S. (2006). Memory trace of motor learning shifts transsynaptically from cerebellar cortex to nuclei for consolidation. *Neuroscience* 139, 767–777. doi: 10.1016/j.neuroscience.2005.12.035

Solouki, S., Bahrami, F., and Janahmadi, M. (2018). "Effects of irreversible olivary system lesion on the gain adaptation of optokinetic response eye movement: a model based study," in *25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)* (Qom: IEEE), 1–6. doi: 10.1109/ICBME.2018.8703572

Solouki, S., Bahrami, F., and Janahmadi, M. (2019). "In pursuit of activity dependent synaptic plasticity rules for cerebellar motor learning: a computational study," in *27th Iranian Conference on Electrical Engineering (ICEE)* (Yazd: IEEE), 1812–1818. doi: 10.1109/IranianCEE.2019.8786378

Solouki, S., and Pooyan, M. (2016). Arrangement and applying of movement patterns in the cerebellum based on semi-supervised learning. *Cerebellum* 15, 299–305. doi: 10.1007/s12311-015-0695-3

Stevens, C. F. (2003). Neurotransmitter release at central synapses. *Neuron* 40, 381–388. doi: 10.1016/S0896-6273(03)00643-3

Strata, P. (2009). David Marr's theory of cerebellar learning: 40 years later. *J. Physiol.* 587, 5519–5520. doi: 10.1113/jphysiol.2009.180307

Takeuchi, T., Ohtsuki, G., Yoshida, T., Fukaya, M., Wainai, T., Yamashita, M., et al. (2008). Enhancement of both long-term depression induction and optokinetic response adaptation in mice lacking delphilin. *PLoS ONE* 3:e2297. doi: 10.1371/journal.pone.0002297

Van Der Giessen, R. S., Koekkoek, S. K., van Dorp, S., De Gruijl, J. R., Cupido, A., Khosrovani, S., et al. (2008). Role of olivary electrical coupling in cerebellar motor learning. *Neuron* 58, 599–612. doi: 10.1016/j.neuron.2008.03.016

Wang, R., Tan, J., Guo, J., Zheng, Y., Han, Q., So, K.-F., et al. (2018). Aberrant development and synaptic transmission of cerebellar cortex in a VPA induced mouse autism model. *Front. Cell. Neurosci.* 12:500. doi: 10.3389/fncel.2018.00500

Welsh, J. P., Yamaguchi, H., Zeng, X.-H., Kojo, M., Nakada, Y., Takagi, A., et al. (2005). Normal motor learning during pharmacological prevention of Purkinje cell long-term depression. *Proc. Natl. Acad. Sci. U.S.A.* 102, 17166–17171. doi: 10.1073/pnas.0508191102

Wulff, P., Schonewille, M., Renzi, M., Viltono, L., Sassoè-Pognetto, M., Badura, A., et al. (2009). Synaptic inhibition of Purkinje cells mediates consolidation of vestibulo-cerebellar motor learning. *Nat. Neurosci.* 12, 1042–1049. doi: 10.1038/nn.2348

Yuzaki, M. (2013). Cerebellar LTD vs. motor learning—Lessons learned from studying GluD2. *Neural Networks* 47, 36–41. doi: 10.1016/j.neunet.2012.07.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Solouki, Bahrami and Janahmadi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 02 October 2019
doi: 10.3389/fncom.2019.00070

Tonality Tunes the Statistical Characteristics in Music: Computational Approaches on Statistical Learning

*Tatsuya Daikoku**

Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Statistical learning is a learning mechanism based on transition probability in sequences such as music and language. Recent computational and neurophysiological studies suggest that the statistical learning contributes to production, action, and musical creativity as well as prediction and perception. The present study investigated how statistical structure interacts with tonalities in music based on various-order statistical models. To verify this in all 24 major and minor keys, the transition probabilities of the sequences containing the highest pitches in Bach's Well-Tempered Clavier, which is a collection of two series (No. 1 and No. 2) of preludes and fugues in all of the 24 major and minor keys, were calculated based on *n*th-order Markov models. The transition probabilities of each sequence were compared among tonalities (major and minor), two series (No. 1 and No. 2), and music types (prelude and fugue). The differences in statistical characteristics between major and minor keys were detected in lower- but not higher-order models. The results also showed that statistical knowledge in music might be modulated by tonalities and composition periods. Furthermore, the principal component analysis detected the shared components of related keys, suggesting that the tonalities modulate statistical characteristics in music. The present study may suggest that there are at least two types of statistical knowledge in music that are interdependent on and independent of tonality, respectively.

Keywords: creativity, Markov model, n-gram, information theory, corpus, prediction, composition, implicit learning

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Tuo Zhang,
Northwestern Polytechnical University, China
Daya Shankar Gupta,
Camden County College, United States

***Correspondence:**
Tatsuya Daikoku
daikoku@cbs.mpg.de

Received: *20 April 2019*
Accepted: *19 September 2019*
Published: *02 October 2019*

Citation:
Daikoku T (2019) Tonality Tunes the Statistical Characteristics in Music: Computational Approaches on Statistical Learning. Front. Comput. Neurosci. 13:70. doi: 10.3389/fncom.2019.00070

INTRODUCTION

Prediction and Production in the Statistical Learning

The brain is innately equipped with statistical learning (SL) machineries that model external phenomena as a dynamical system that encode the probability distributions. The SL is thought as an implicit process in which the brain automatically calculate transitional-probability (TP) distribution of sequential information such as music and language (Saffran et al., 1996; Cleeremans et al., 1998). Furthermore, based on the internalized statistical model, it can predict a future state and optimize action for achieving a given goal (Monroy et al., 2017a,c) to resolve the uncertainty of information (Friston, 2010). The SL has also be thought to contribute to the encoding of the complexity in the information (Hasson, 2017), and to acquisition of musical and linguistic knowledge including tonality (Daikoku et al., 2016) and syntax (Daikoku et al., 2017a).

For example, an increasing volume of literature also demonstrates that SL and the knowledge associate with human's action (Zubicaray et al., 2013; Monroy et al., 2017a,b, 2018) and decision-making (Schwartenbeck et al., 2013; Friston et al., 2014, 2015; Pezzulo et al., 2015). For example, motor cortex activity contributes to SL of action words (Zubicaray et al., 2013). Furthermore, cerebellum and cerebral cortex partially share same network responsible for the interalized statistical model. That is, statistical knowledge formed in cerebral cortex may be sent to the cerebellum that is thought to play important roles in prediction of sequences (Lesage et al., 2012; Moberget et al., 2014), motor skill learning (Ito, 2008), habit learning (Friston et al., 2016), generalization or abstraction based on transitional probabilities (Shimizu et al., 2017), efficient performance in a learned context (Balsters et al., 2014). These findings may suggest that the internalized statistical model affects production of music (i.e., composition) (Daikoku, 2019a), the creativity (Wiggins, 2018), and individuality of artistic expression (Daikoku, 2018b) as well as the prediction and perception (Daikoku, 2019c). It is, however, unknown how the acquired statistical knowledge influences the production of music.

Statistical Learning Machinery in Musician

According to recent studies, musicians are better statistical learners than non-musicians (Francois and Schön, 2011; François et al., 2012; Hansen and Pearce, 2014; Przysinda et al., 2017; Elmer and Lutz, 2018). Furthermore, it is suggested that, through long-term musical training, musicians optimize the brain's probabilistic model of SL, and that the musically-optimized SL model allow the brain to precisely and efficiently predict tones during SL of another musical and auditory sequences (Francois and Schön, 2011; Kim et al., 2011; Hansen and Pearce, 2014; Przysinda et al., 2017). Recent computational studies also suggested that, from early to late periods in the composer's lifetime, the transitional probabilities of familiar phrase in each piece of music were gradually decreased (Daikoku, 2018d, 2019a). These findings were prominent in higher-, rather than lower-order SL models. These studies suggest that the higher-, rather than lower-, order statistical knowledge (Daikoku, 2018a) may be susceptible to long-term experience that modulates brain's SL model (Hansen and Pearce, 2014). Furthermore, computational studies on improvisation music suggested that lower-order SL models represented general characteristics shared among musicians, whereas higher-order SL models detected specific characteristics unique to each musician (Daikoku, 2018b). In this context, it can be hypothesized that statistical models in music, which may reflect the composer's statistical knowledge, interact with the music-specific structures of tonality. To our knowledge, however, few studies have examined how TP in music interacts with the tonalities. To understand the characteristics of music from interdisciplinary aspects that include informatics, musicology, and psychology, it is important to verify the interaction between tonality and statistical structure in music, especially regarding strategies of musical composition.

Computational Modeling

The computational model and simulation have been used to understand SL systems (e.g., Pearce and Wiggins, 2012; Rohrmeier and Rebuschat, 2012; Daikoku, 2018a, 2019b; Wiggins, 2018). Particularly, the prediction and production of SL is partially supported by chunking hypothesis that learning is based on extracting, storing, and combining small chunks. For example, information-theoretical models including *Markovian* processes have been applied to neurophysiological studies of SL in human brain as well as computational simulation (Pearce et al., 2010; Pearce and Wiggins, 2012; Daikoku et al., 2014, 2015, 2017b, 2018; Yumoto and Daikoku, 2016, 2018; Daikoku and Yumoto, 2017, 2019; Daikoku, 2018c). These neurophysiological experiments showed consistent evidence: neural activities for stimuli with high information content (i.e., low probability) are larger than those with low information content (i.e., high probability). This neural phenomenon is in agreement with a *Bayesian* hypothesis in theoretical neurobiology that the brain encodes probabilities (beliefs) about the causes of sensory data, and that these beliefs are updated in response to new sensory evidence based on Bayesian inference (Kersten et al., 2004; Knill and Pouget, 2004; Doya et al., 2007; Friston, 2010; O'Reilly et al., 2012; Parr and Friston, 2018; Parr et al., 2018). That is, information-theoretical computational models including *Markovian* processes can capture a variety of neurophysiological phenomena on prediction, chunk formation, action, and production in the framework of SL theory.

The Aim of the Present Study

This study aimed to examine how the statistical structure interacts with tonality. To verify the statistical relationships in all the keys of Western classical music (**Figure 1**), the TPs of the sequences containing the highest pitches in Bach's Well-Tempered Clavier, BWV 846–893, which is a collection of two series (No. 1 and No. 2) of preludes and fugues in all of the 24 major and minor keys (**Figure 1**), were calculated using six different orders of Markov or n-gram models (i.e., first- to sixth-order Markov chains). Johann Sebastian Bach (1685–1750) was a composer during the Baroque period, who contributed to the development of musical tonality and the Western classical music theory (Rohrmeier and Cross, 2008). His music is often used to verify the probabilities of musical sequences (Rohrmeier and Cross, 2008; Kim et al., 2011). Particularly, to understand the relationships between tonality and statistical structure in music, the Well-Tempered Clavier may be one of the best mediums because it is a collection of music containing all 24 of the major and minor keys by a single composer in Western classical music. Thus, the statistics in each piece of music with a key in the Well-Tempered Clavier could be, in part, regarded as an approximation of the statistics of the entire range of Western classical music in each key. Thus, to extract statistical knowledge dependent on keys and tonalities, the present study verified the statistical structure in each key and tonality. The TPs of each sequence were compared among tonalities (major and minor), two series (No. 1 and No. 2), and music types (prelude and fugue). It was

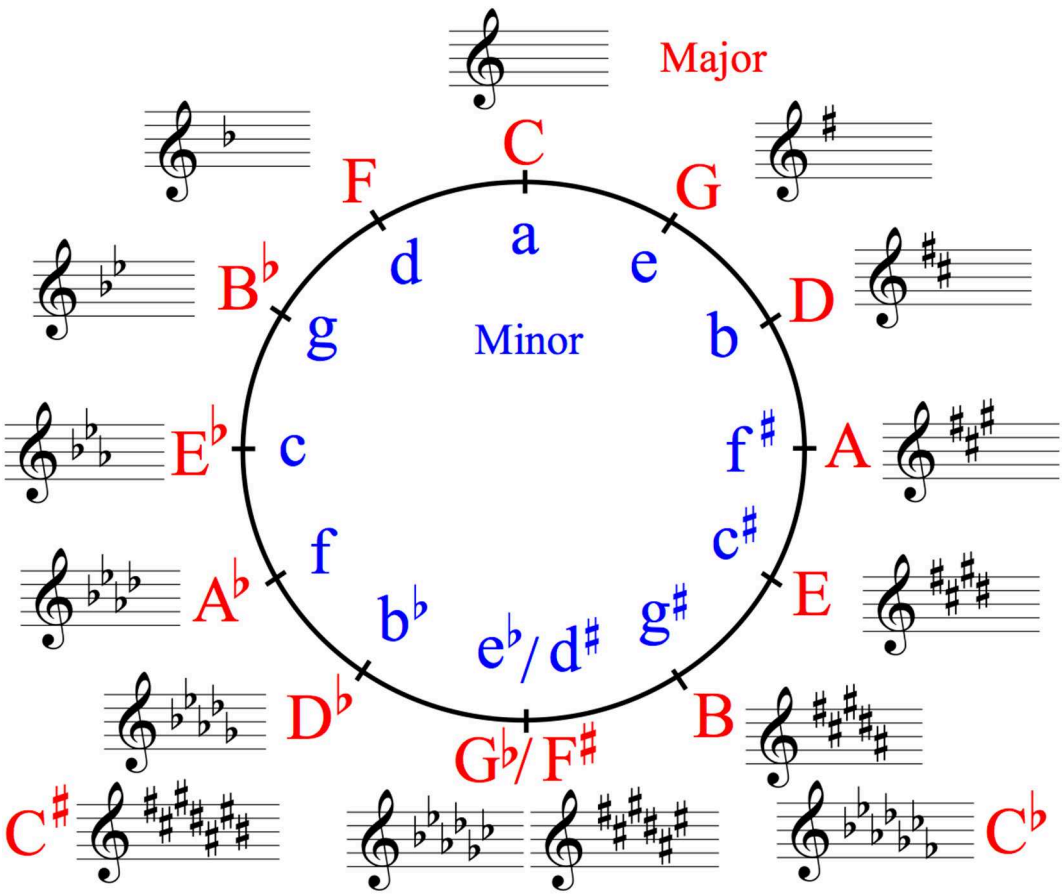


FIGURE 1 | Circle of fifths showing all 24 major and minor keys in Western classical music. A related key is one sharing many common tones with an original key, as opposed to a distant key. In music, such a key shares all, or all except one, pitches with a key with which it is being compared, and it is adjacent to it on the circle of fifths and its relative majors or minors. In a related key, a subdominant key has one more flat around the circle of fifths, and a relative key has the same key signature.

hypothesized that the statistical structure in music interacts with the tonality in music. If so, these findings suggest that music-specific knowledge of tonality modulates statistical knowledge in music.

METHODS

The Well-Tempered Clavier, BWV 846–893, which is a collection of two series (No. 1 and No. 2) of Preludes and Fugues in all 24 major and minor keys that was composed for solo keyboard by Johann Sebastian Bach, was used in the present study. Electronic scoring data of highest pitch were extracted from the Extensible Markup Language (XML) files. The highest pitches were chosen based on the following definitions (**Figure 2**): the grace notes were excluded, the pitches with slurs can be counted as one, and the highest pitches that can be played at a given point in time. According to SL theory, the brain automatically computes *n*th-order TPs of sequence. The transitional probability is a conditional probability of an event B given that the latest event A has occurred, written as P(B|A). The first- to six-order TPs of an event in SL were calculated from conditional probability (P) of an event e_{n+1}, given the preceding n events, based on the first- to six-order Markov models ($n = 1$–6):

$$P\left(e_{n+1}|e_n\right) = \frac{P(e_{n+1} \cap e_n)}{P(e_n)}$$

From the perspective of psychology, the formula can be interpreted as positing that the brain predicts a subsequent event e_{n+1} based on the preceding events e_n in a sequence (for more details, see Daikoku, 2018c). In other words, learners expect the events with higher TPs based on the latest n states (i.e., *n*th-order), whereas they are likely to be surprised by events with lower TPs. Then, all of the pitch transitions were numbered so that the first pitch was 0 in each sequential pattern, and an increase or decrease in a semitone was 1 and −1 based on the first pitch, respectively (**Figure 2**). This reveals interval patterns but not pitch pattern, and eliminates the effects of the change of key on sequential patterns. This procedure was employed because the interpretation of the change of key depends on musicians, and it is difficult to define it in an objective manner. Thus, the results in this study may represent statistics based on relative, rather than absolute pitches. To verify the difference in

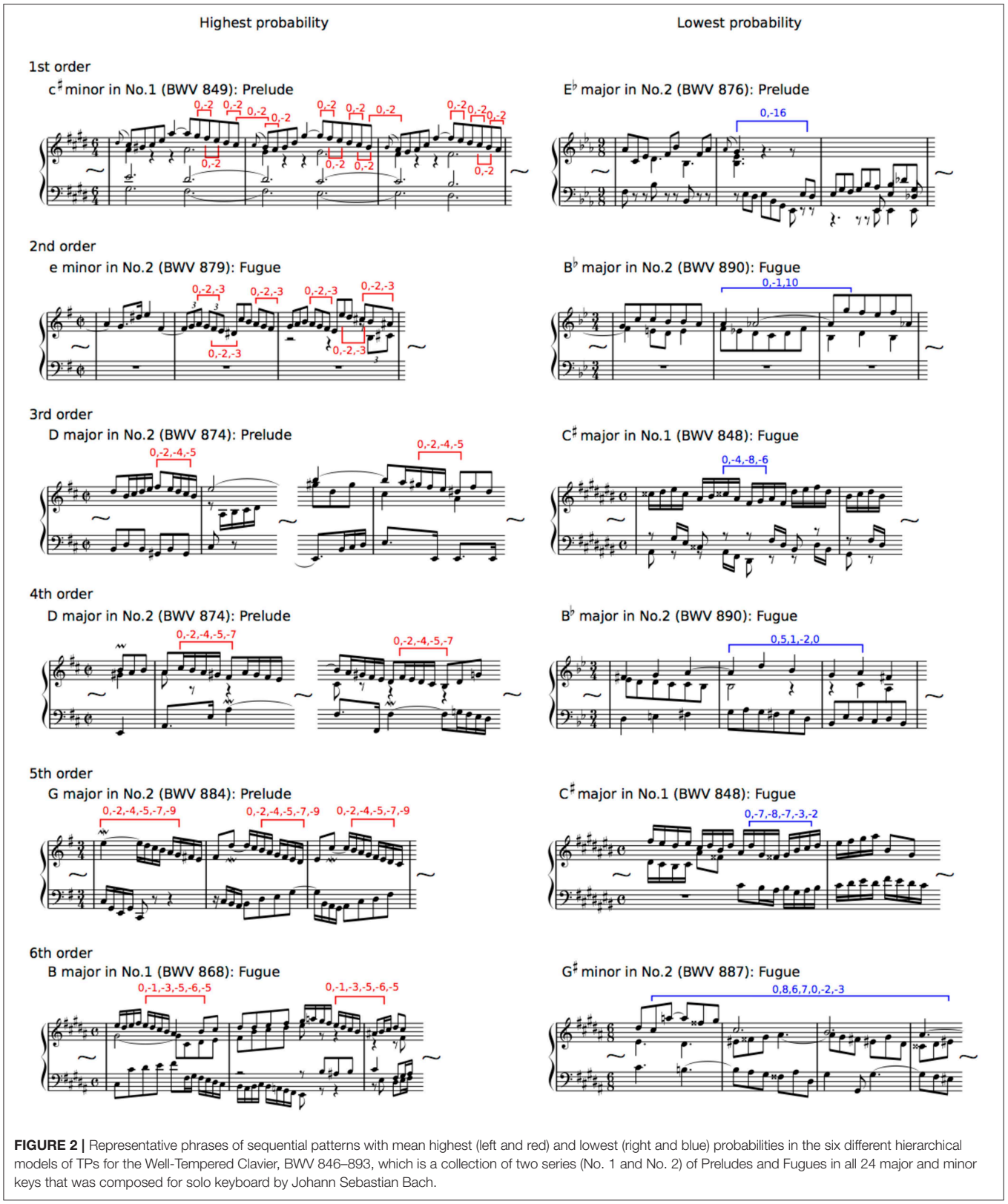


FIGURE 2 | Representative phrases of sequential patterns with mean highest (left and red) and lowest (right and blue) probabilities in the six different hierarchical models of TPs for the Well-Tempered Clavier, BWV 846–893, which is a collection of two series (No. 1 and No. 2) of Preludes and Fugues in all 24 major and minor keys that was composed for solo keyboard by Johann Sebastian Bach.

statistical structures between prelude and fugue, the sequential patterns that appear in all pieces of music that were divided between prelude and fugue were only used in the present study (1st: 4). In the second- to sixth-order Markov chains, sequential patterns that appear in all music could not be detected. The empirical logit transformation was applied to normalize the

TPs. The empirical logit transform allows data distribution to be normalized, and is used for a tolerence such that infinity is not returned when the argument is zero (0%) or one (100%). Thus, it is applicable when the TP values, which often show 0% and 100%, are analyzed. Then, we conducted repeated-measure analysis of variances (ANOVAs) based on a factor type (prelude vs. fugue), a factor tonality (major vs. minor), a factor number (No. 1 vs. No. 2), and a factor sequence (4 sequences) for the 1st-order Markov model. Bonferroni-corrected *post-hoc* tests were conducted for further analysis (Statistical significance levels: $p < 0.05$). It has been suggested that the TP distribution represents statistical characteristics in music (Daikoku, 2018b). Thus, using the nth-order TP distributions, the musical characteristic in each tonality was verified by correlation analysis. Furthermore, based on the result of correlation analysis, the TPs, in which there are a number of correlations of at least 0.3 (30), were analyzed by principal component analysis (PCA). The criteria of eigenvalue were set over 1. The first three components (i.e., the first to third highest cumulative contribution ratios) were adopted in the present study. The present study focus on the values of "*loadings*." The loading has generally been understood as the weights for each original variable when calculating the principal component. The representative phrases of sequential patterns with mean highest and lowest probabilities were decoded as musical scores (**Figure 2**). The criterion of the eigenvalue was set over 1 (Statistical significance levels: $p < 0.05$).

RESULTS

ANOVA

Higher-order of model represents exponentially larger numbers of sequential patterns: over forty in the first-order models, 600 in the second-order models, 3,500 in the third-order models, 9,000 in the fourth-order models, 15,000 in the fifth-order models, 20,000 in the sixth-order models. The results were shown in **Figure 3**. The main tonality effect showed that TPs of sequence that appear in all music in major key were lower than those in minor key [$F_{(1,11)} = 9.83$, $p = 0.009$, partial $\eta^2 = 0.47$; **Figure 3A**]. The main type effect showed that TPs of sequence that appear in all music in preludes were lower than those in fugues [$F_{(1,11)} = 140.74$, $p < 0.001$, partial $\eta^2 = 0.93$; **Figure 3B**]. The main sequence effect were significant [$F_{(2.16,23.76)} = 26.54$, $p < 0.001$, partial $\eta^2 = 0.71$; **Figure 3C**]. The TPs of [0, −2] was significantly higher compared with those of [0, −1], [0, 1], and [0, −3] (all: $p < 0.001$). The TPs of [0, −1] was higher compared with those of [0, 1] ($p = 0.005$) and [0, −3] ($p < 0.001$). The TPs of [0, 1] was higher compared with those of [0, −3] ($p < 0.001$). The tonality-number interactions were significant [$F_{(1,11)} = 7.57$, $p = 0.019$, partial$\eta^2 = 0.41$; **Figure 3D**]. In No. 1 of a collection of two series, the TPs in major key were significantly lower than those in minor key ($p = 0.001$). In minor key, the TPs in No. 1 were higher compared with those in No. 2 ($p = 0.044$). The tonality-sequence interactions were significant [$F_{(1.68,18.46)} = 5.35$, $p = 0.019$, partial $\eta^2 = 0.33$; **Figure 3E**]. In sequences of [0, −1], the TPs in major key were significantly lower than those in minor key ($p = 0.001$). In sequences of [0, 1], the TPs in major key were significantly lower than those in minor key ($p = 0.019$). In major key, the TPs of [0, −2] was higher compared with those of [0, −1], [0, 1], and [0, −3] (all: $p < 0.001$). The TPs of [0, −1] was higher compared with those of [0, −3] ($p < 0.001$). The TPs of [0, 1] was higher compared with those of [0, −3] ($p = 0.002$). In minor key, the TPs of [0, −2] was higher compared with those of [0,−1] (p = 0.001), [0, 1] ($p < 0.001$), and [0, −3] ($p < 0.001$). The TPs of [0, −1] was higher compared with those of [0, 1] ($p = 0.006$) and [0, −3] ($p < 0.001$). The TPs of [0, 1] was higher compared with those of [0, −3] ($p < 0.001$).

The type-sequence interactions were significant [$F_{(1.85,20.34)} = 7.64$, $p = 0.004$, partial $\eta^2 = 0.41$]. In sequences of [0, −2], the TPs in prelude were significantly lower than those in fugue ($p < 0.001$). In sequences of [0, −1], the TPs in prelude were significantly lower than those in fugue ($p = 0.012$). In sequences of [0, 1], the TPs in prelude were significantly lower than those in fugue ($p = 0.005$). In prelude, the TPs of [0, −2] was higher compared with those of [0, −1], [0, 1], and [0, −3] (all: $p < 0.001$). In fugue, the TPs of [0, −2] was higher compared with those of [0, −1], [0, 1], and [0, −3] (all: $p < 0.001$). The TPs of [0, −1] was higher compared with those of [0, 1] ($p = 0.006$) and [0, −3] ($p = 0.001$). The TPs of [0, 1] was higher compared with those of [0, −3] ($p = 0.010$). In fugue, the TPs of [0, −2] was higher compared with those of [0, −1], [0, 1], and [0, −3] (all: $p < 0.001$). The TPs of [0, −1] was higher compared with those of [0, 1] ($p = 0.047$) and [0, −3] ($p < 0.001$). The TPs of [0, 1] was higher compared with those of [0, −3] ($p < 0.001$).

Correlation Analysis

All the results of the correlation analysis are shown in **Supplementary Material**. In the first-order TPs, all the pieces of music are strongly ($0.7 \leqq |r| < 1.0$, $p < 0.01$; **Supplementary Material**, red) or moderately ($0.4 \leqq |r| < 0.7$, $p < 0.01$; **Supplementary Material**, green) related to each other (**Figure 4A**). In the second-order TPs, all the pieces of music are moderately ($0.4 \leqq |r| < 0.7$, $p < 0.01$; **Supplementary Material**, green) or weakly ($0.2 \leqq |r| < 0.4$, $p < 0.01$; **Supplementary Material**, yellow) related to each other (**Figure 4B**). In the third- and fourth-order TPs, some of the music is weakly ($0.2 \leqq |r| < 0.4$, $p < 0.01$; **Supplementary Material**, yellow) related to each other (**Figures 4C,D**). There are more weak correlations in the third-order than in the fourth-order TPs. In the fifth- and sixth-order TPs, no strong, moderate, or weak correlations were detected (**Figures 4E,F**).

Principal Component Analysis

Based on the results of correlation analysis, the first- and second-order TPs, in which there are a number of correlations of at least 0.3 (Tabachnick and Fidell, 2007), were analyzed by principal component analysis. In the first-order TP, the decision was made to specify two principal component solutions (eigenvalue >1; **Table 2A** and **Figure 5**). The two principal components accounted for 92.4% of the total variance. All of the music loaded higher than 0.58 on component 1. The "*loadings*" can be understood as the weights for each original variable when calculating the principal component. Thus, the result explains the general component of the Well-Tempered Clavier. The C major

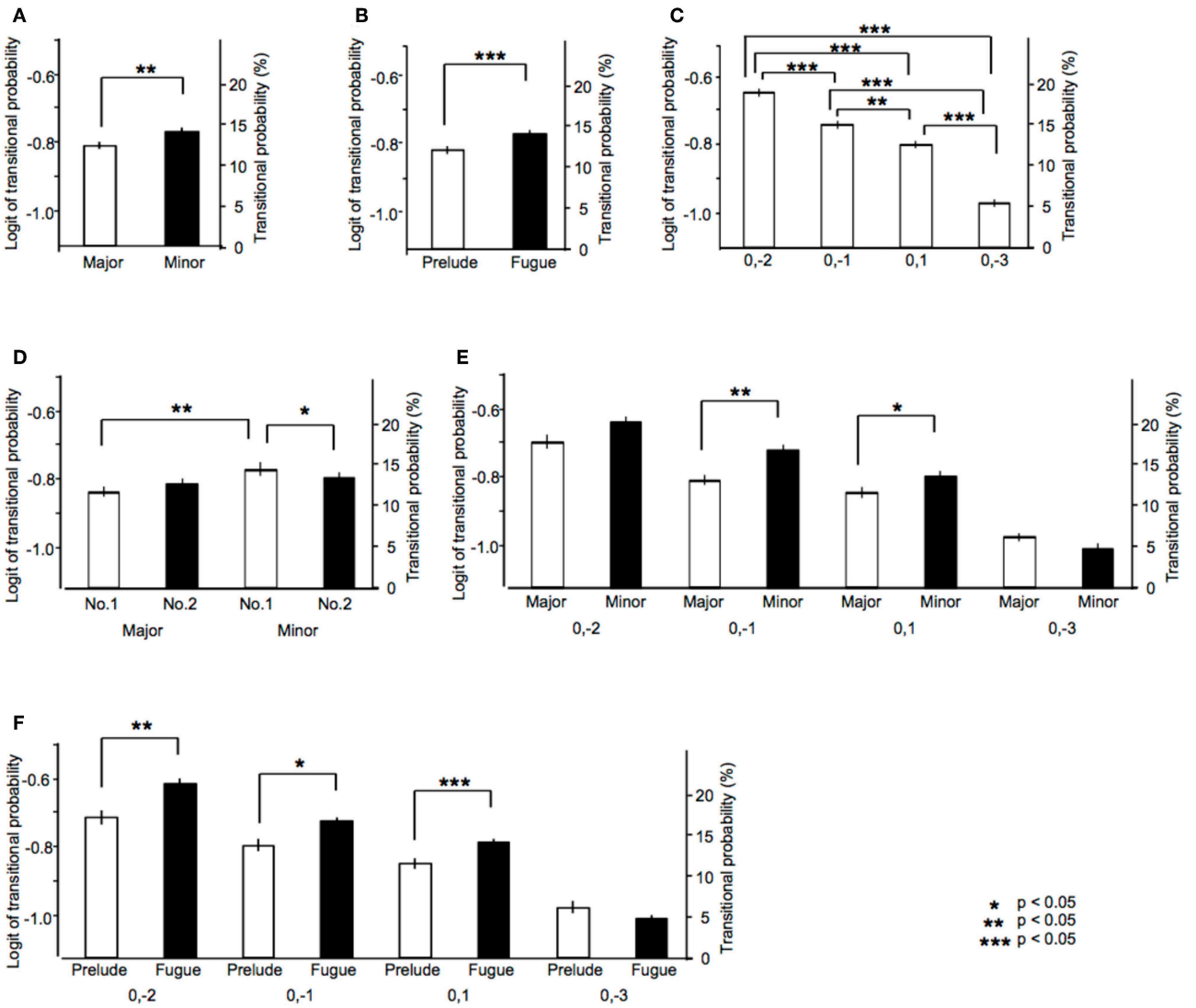


FIGURE 3 | The results of ANOVA in analysis 2. The main effects of **(A)** tonality, **(B)** type, and **(C)** sequence. The interactions of **(D)** tonality-number, **(E)** tonality-sequence, and **(F)** type-sequence.

and D minor in the first series (No. 1) of the Well-Tempered Clavier loaded higher than 0.45 on 2. This explains a component of related keys (i.e., the relative key of the subdominant key; **Table 1**) between C major and D minor. In the second-order TP, the decision was made to specify a three principal component solution (eigenvalue >1; **Table 2B** and **Figure 5**). The three principal components accounted for 83.2% of the total variance. All of the music loaded higher than 0.55 on 1,. This explains the general component of the Well-Tempered Clavier. On the other hand, compared to the other music, the C major and D minor in No. 1 of the Well-Tempered Clavier loaded <0.57 on component 1. The C minor in No. 1 and E♭ major in No. 2 of the Well-Tempered Clavier loaded at 0.41 or higher on component 2. This explains shared components of a related key (i.e., relative keys). The only D minor in No. 1 of the Well-Tempered Clavier loaded heavily (0.52) on component 3.

DISCUSSION

Psychological Aspects of TP in Musical Sequence

Based on the information theory (Shannon, 1948) covering multi-order Markov models and the cognitive models, a tone with a higher TP may be one that a composer is more likely to choose than those with lower TPs. Thus, the TP distributions sampled from music may represent the musical characteristics based on a composer's statistical knowledge underlying prediction. The present study aimed to examine how the statistical structure interacts with tonality in music. To verify it in all 24 major and minor keys (**Figure 1**), the TPs of the sequences containing the highest pitches in Well-Tempered Clavier were calculated based on Markov stochastic models. It was hypothesized that the statistical structure in

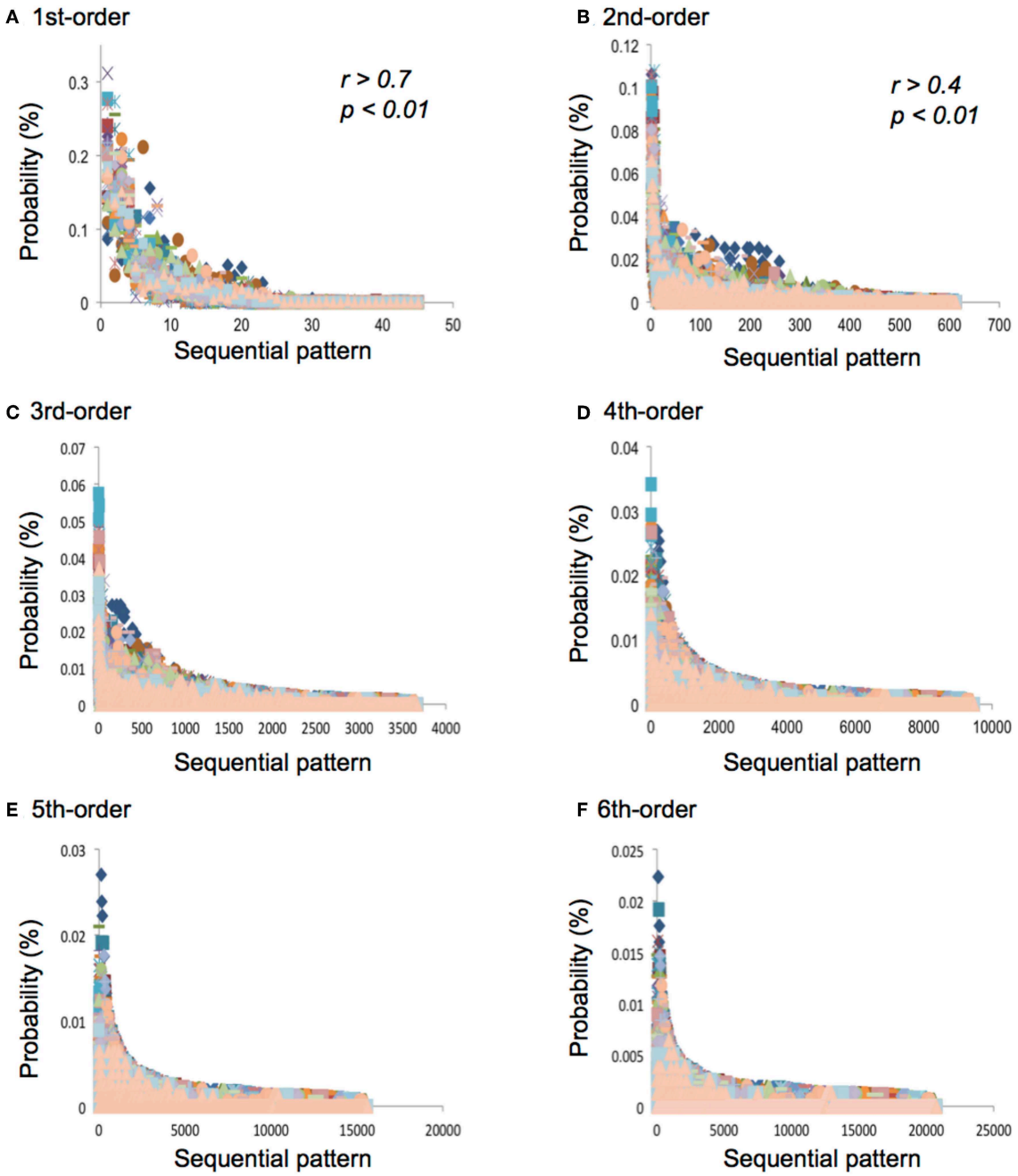


FIGURE 4 | (A) The first-, **(B)** second-, **(C)** third-, **(D)** fourth-, **(E)** fifth-, and **(F)** sixth-order TPs in each sequential pattern. The horizontal and vertical axes represent sequential patterns and the TPs, respectively. The sequential patterns were arranged in descending order in each hierarchy.

music interacts with tonality in music and that music-specific knowledge of tonality may modulates statistical knowledge in music.

The Relationships Between Tonality and Hierarchy of Stochastic Structure in Music

The present study adopted the sequences that appear in all pieces of music (i.e., universal sequences in the Well-Tempered Clavier). The TP differences between major and minor keys could be detected in lower-order (1st and 2nd in **Figure 3A**) but not in higher-order hierarchical models. This implies that these sequences may have specific semantics in each major and minor key. In the context of statistical learning, the tonality may modulate a lower- rather than a higher-order statistical knowledge of music. The TPs in the fugue were higher than those in the prelude (**Figure 3B**), and the difference

was prominent in sequences in which the interval was not more than a whole step (i.e., ±2), such as those found in musical scales (**Figure 3F**). It is well-known that the prelude less strictly follows the rules of Western classical music compared to the fugue. The findings in the present study may reflect the difference in statistical knowledge related to strategies for musical composition.

As a general tendency, the TPs of universal sequences were higher in minor than in major keys (**Figures 3A,E**). However, the difference became weaker in the series of No. 2 compared to that in No. 1 (**Figure 3D**). Statistical knowledge of universal sequences might be modulated from composition periods in No. 1 to No. 2. It would be interesting if the time-course variation of statistical structures may reflect the time-course variation of statistical knowledge. It is of note, however, that this study did not directly investigate the composer's statistical knowledge of music, as only the statistics of musical scores were analyzed. There may be other possible explanations for the findings of this study. For instance, it might have been Bach's intentional plan to compose music based on the statistical structure of music. Future studies should examine the effects of statistical knowledge on music compositions and neurological responses in parallel.

In the first- and second-order TPs, all of the pieces of music are related to each other (**Supplementary Material** and **Figure 4**). In the third- and fourth-order TPs, some of the music is related to each other, regardless of tonalities. There are more correlations in the third-order than fourth-order TPs. In the fifth- and sixth-order TPs, no remarkable correlations were detected. These results suggest that there are statistical characteristics that are shared among each piece of music at least in the first- and second-order hierarchical levels of statistical structure. In other words, there may be universal implicit knowledge of music

TABLE 1 | Related key in all 24 major and minor keys.

Major	Relative minor	Subdominant, dominant, and their relatives	Parallel minor
C	Am	F, G, Dm, Em	Cm
G	Em	C, D, Am, Bm	Gm
D	Bm	G, A, Em, F#m	Dm
A	F#m	D, E, Bm, C#m	Am
E	C#m	A, B, F#m, G#m	Em
B	G#m	E, F#, C#m, D#m	Bm
G♭	E♭m	C♭, D♭, A♭m, B♭m	F#m
D♭	B♭m	G♭, A♭, E♭m, Fm	C#m
A♭	Fm	D♭, E♭, B♭m, Cm	G#m
E♭	Cm	A♭, B♭, Fm, Gm	E♭m
B♭	Gm	E♭, F, Cm, Dm	B♭m
F	Dm	B♭, C, Gm, Am	Fm

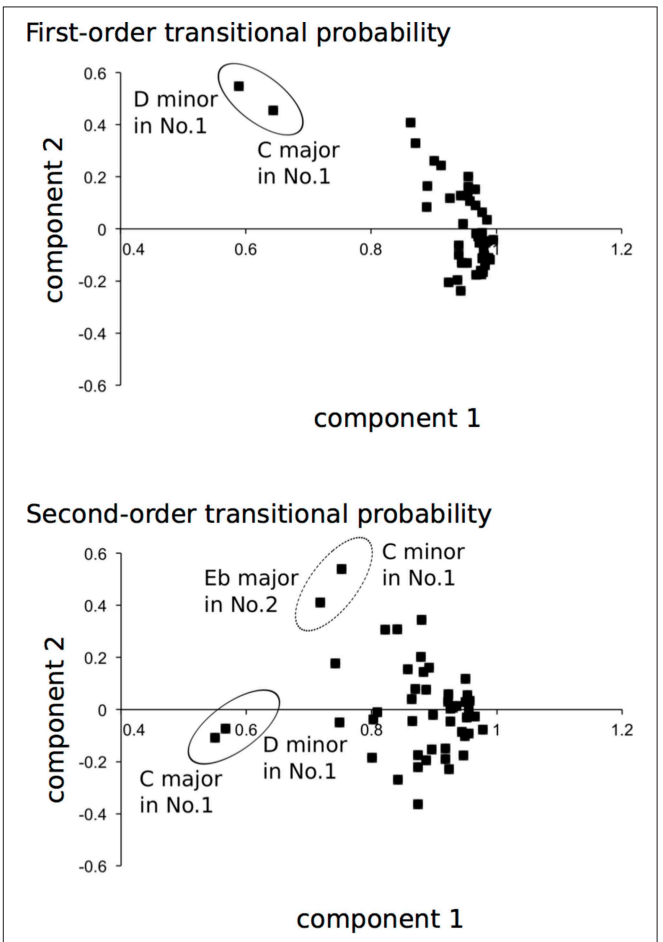


FIGURE 5 | Principal component analysis scatter plots. The dots represent each piece of music in the Well-Tempered Clavier, which is a collection of two series (No. 1 and No. 2) in all 24 major and minor keys that was composed for solo keyboard by Johann Sebastian Bach. The dots in each circle represent pieces of music with the component of each related key: between D minor and C major, Eb major and C minor, and C major and D minor.

TABLE 2A | The results of principal component analysis.

	1st-order transition probability			2nd-order transition probability		
	Total	Variance *	Cumulative *	Total	Variance *	Cumulative *
Component 1	42.814	89.196	89.196	37.073	77.236	77.236
Component 2	1.548	3.225	92.421	1.528	3.184	80.419
Component 3				1.331	2.773	83.193

The eigenvalue and percentages of variance and comulative variance.
**Percentage.*

TABLE 2B | The eigenvectors for the principal components.

		1st-order transitonal probability		2nd-order transition probability		
		Component 1	**Component 2**	**Component 1**	**Component 2**	**Component 3**
No. 1	C major	0.644	0.454	0.550	−0.109	−0.212
	C minor	0.968	−0.018	0.752	0.539	0.090
	Db major	0.912	0.243	0.809	−0.011	0.201
	Db minor	0.940	−0.063	0.918	−0.150	0.137
	D major	0.940	−0.100	0.864	0.040	−0.109
	D minor	0.589	0.547	0.567	−0.073	0.523
	Eb major	0.971	−0.030	0.947	−0.177	−0.017
	Eb minor	0.976	−0.175	0.952	−0.031	−0.200
	E major	0.946	−0.132	0.879	0.202	−0.289
	E minor	0.973	−0.053	0.926	−0.046	0.074
	F major	0.954	0.127	0.883	0.144	0.107
	F minor	0.990	−0.119	0.955	−0.093	−0.130
	Gb major	0.871	0.329	0.742	0.176	0.068
	Gb minor	0.980	−0.075	0.874	−0.222	−0.089
	G major	0.966	0.151	0.918	−0.190	0.173
	G minor	0.975	−0.160	0.953	0.055	−0.128
	Ab major	0.955	0.200	0.822	0.306	0.175
	Ab minor	0.953	−0.132	0.858	0.154	−0.173
	A major	0.943	0.128	0.874	−0.175	0.008
	A minor	0.988	−0.116	0.978	−0.077	−0.083
	Bb major	0.958	0.106	0.887	−0.195	0.051
	Bb minor	0.938	−0.197	0.926	0.003	−0.042
	B major	0.943	−0.238	0.922	0.029	−0.313
	B minor	0.968	−0.177	0.944	−0.086	−0.109
No. 2	C major	0.974	−0.051	0.880	0.344	−0.034
	C minor	0.977	−0.016	0.892	0.160	0.106
	Db major	0.901	0.261	0.865	−0.045	−0.026
	Db minor	0.983	−0.053	0.930	0.007	0.032
	D major	0.926	0.118	0.874	−0.364	0.171
	D minor	0.977	0.063	0.896	−0.154	0.063
	Eb major	0.955	0.162	0.718	0.411	0.342
	Eb minor	0.889	0.083	0.842	−0.270	0.251
	E major	0.987	−0.111	0.950	0.118	−0.011
	E minor	0.981	−0.103	0.924	−0.229	−0.167
	F major	0.979	−0.167	0.957	0.033	−0.169
	F minor	0.890	0.164	0.803	−0.038	0.214
	Gb major	0.967	0.090	0.887	0.076	0.026
	Gb minor	0.986	−0.047	0.923	0.059	0.046
	G major	0.863	0.408	0.749	−0.050	0.350
	G minor	0.947	0.020	0.870	0.079	−0.019
	Ab major	0.995	−0.042	0.965	−0.027	−0.050
	Ab minor	0.945	−0.130	0.898	−0.021	0.025
	A major	0.980	−0.110	0.841	0.307	−0.044
	A minor	0.924	−0.206	0.801	−0.186	−0.070
	Bb major	0.978	−0.113	0.950	0.028	−0.199
	Bb minor	0.982	−0.140	0.955	−0.001	−0.159
	B major	0.985	0.034	0.949	−0.102	0.052
	B minor	0.978	−0.114	0.936	0.013	−0.050

in the composer at the lower hierarchical levels, regardless of tonalities and pitch frequencies. The higher the hierarchical levels of TPs, the less the music was correlated with each other. From information theoretical viewpoint, the statistical models at lower hierarchical levels increases joint probability and mutual information, whereas statistical structures at higher hierarchical levels are less correlated, and interpreted as surprisal information (Gupta and Bahmer, 2019). The combined increase in mutual information at lower hierarchical level and surprisal information at higher hierarchical level would serve as the basis of specific knowledge about music (Gupta and Bahmer, 2019). These results also suggest that the higher the hierarchical level of statistical structure, the stronger the independence of characteristics in each piece of music. The specific characteristics in each piece of music may exist in higher hierarchical levels of statistical structure. This may imply that greater creativity is attributed at higher hierarchical level (Daikoku, 2018b). Thus, it could be assumed that the general statistical structure that is shared among many pieces of music is formed by low-hierarchical implicit knowledge, whereas the specific structure that is independent of each piece of music is formed by high-hierarchical implicit knowledge (Gupta and Bahmer, 2019).

J.S. Bach's Music for Study on Implicit and Explicit Knowledge

Johann Sebastian Bach (1685–1750), a German composer and musician of the Baroque period, is considered to have contributed to the development of musical tonality and has been central to Western classical music theory until the present (Rohrmeier and Cross, 2008). His music is often used to investigate the probabilities of musical sequences. Furthermore, to investigate the relationships between tonality and statistical structure in music, the Well-Tempered Clavier is considered an excellent medium because it is a collection of music containing all the keys of Western classical music (i.e., 24 major and minor keys). Thus, the statistical characteristics of each piece of music with a key in the Well-Tempered Clavier could be, in part, regarded as approximations of the statistical characteristics of the entire range of Western classical music in each key. In other words, the findings in the present study may reflect the implicit knowledge in each musical key in humans who explicitly learn the music-specific knowledge based on Western classical music and who intentionally follow these frameworks when composing music. Furthermore, the present study may suggest that there are at least two types of implicit knowledge that are dependent on and independent of tonality, respectively. This study, however, did not directly demonstrate that the implicit musical knowledge is reflected in music, as only the statistics of musical scores were analyzed. Future studies should investigate, in parallel, how implicit learning in music is reflected in the neurological response and how the learned knowledge is expressed when composing music.

The representative phrases of sequential patterns with mean highest and lowest probabilities were decoded as musical scores in **Figure 2**, based on each hierarchical level of first- (highest: $P[-2|0]$, lowest: $P[-16|0]$), second- (highest: $P[-3|0, -2]$, lowest: $P[10|0, -1]$), third- (highest: $P[-5|0, -2, -4]$, lowest: $P[-6|0, -4, -8]$), fourth- (highest: $P[-7|0, -2, -4, -5]$, lowest: $P[0|0, 5, 1, -2]$), fifth- (highest: $P[-9|0, -2, -4, -5, -7]$, lowest: $P[-2|0, -7, -8, -7, -3]$), and sixth- (highest: $P[-3|0, -1, -3, -5, -6, -5]$, lowest: $P[-3|0, 8, 6, 7, 0, -2]$). The sequential patterns with the highest sequential patterns are familiar ones in Western classical music, suggesting that implicit statistical knowledge and explicit music-specific knowledge interact, in part, with each other. The principal component analysis detected the shared components of related keys (**Figure 5**). This suggests that tonalities modulate implicit knowledge in music. However, these findings are not detected in all the types of related keys (**Supplementary Material**). Future studies will be needed to clarify the relationships between statistical structure and tonalities in music. In the present study, all of the pitch transitions were numbered to understand how the pitches, but not the notes, were transitioned to from the first pitch. This was performed to eliminate the effects of the change of key on sequential patterns. Thus, the results may represent statistics based on relative pitches rather than absolute pitches. Nonetheless, the present study suggests that explicit knowledge on tonality could, in part, modulate implicit knowledge in music.

CONCLUSION

The present study indicated that, in the lower hierarchical levels of statistical structure (first and second orders), all the pieces of music are related to each other. However, the higher the hierarchical levels of TPs, the less the music was correlated with each other, regardless of tonality. These findings suggest that the general statistical structure that is shared among many pieces of music is formed by low-hierarchical implicit knowledge, whereas the specific structure that is independent of each piece of music is formed by high-hierarchical implicit knowledge. This may imply that greater creativity is attributed at higher hierarchical level. On the other hand, the principal component analysis detected the shared components of related keys, suggesting that tonalities modulate implicit knowledge in music. The implicit statistical knowledge and explicit music-specific knowledge could, in part, interact with each other. It is suggested that there are at least two types of implicit knowledge that are dependent on and independent of tonality, respectively. The present study sheds new light on novel methodologies that can be employed to evaluate the implicit knowledge of a composer using musical scores in interdisciplinary studies that include psychology, informatics, and musicology.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

The methodology of the present study was considered by the authors. The author analyzed all of the

data and prepared the figures, and wrote the manuscript text.

FUNDING

The present study was supported by Suntory Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00070/full#supplementary-material

Supplementary Table 1 | The results of correlation analysis. The red, green, yellow cells represent strong ($0.7 \leqq |r| < 1.0$), moderate ($0.4 \leqq |r| < 0.7$), and weak ($0.2 \leqq |r| < 0.4$) correlations, respectively. M and m indicate major and minor keys, respectively.

REFERENCES

Balsters, J. H., Laird, A. R., Fox, P. T., and Eickhoff, S. B. (2014). Bridging the gap between functional and anatomical features of cortico-cerebellar circuits using meta-analytic connectivity modeling. *Human Brain Mapping* 35, 3152–3169. doi: 10.1002/hbm.22392

Cleeremans, A., Destrebecqz, A., and Boyer, M. (1998). Implicit learning: news from the front. *Trends Cogn. Sci.* 2, 406–416. doi: 10.1016/S1364-6613(98)01232-7

Daikoku, T. (2018a). Entropy, uncertainty, and the depth of implicit knowledge on musical creativity: computational study of improvisation in melody and rhythm. *Front. Comput. Neurosci.* 12:97. doi: 10.3389/fncom.2018.00097

Daikoku, T. (2018b). Musical creativity and depth of implicit knowledge: spectral and temporal individualities in improvisation. *Front. Comput. Neurosci.* 12, 1–27. doi: 10.3389/fncom.2018.00089

Daikoku, T. (2018c). Neurophysiological markers of statistical learning in music and language: hierarchy, entropy, and uncertainty. *Brain Sci.* 8:E114. doi: 10.3390/brainsci8060114

Daikoku, T. (2018d). Time-course variation of statistics embedded in music: corpus study on implicit learning and knowledge. *PLoS ONE* 13:e0196493. doi: 10.1371/journal.pone.0196493

Daikoku, T. (2019a). Implicit knowledge and the uncertainty on musical creativity fluctuate over a composer's lifetime. *Front. Comput. Neurosci.* 13:27. doi: 10.3389/fncom.2019.00027

Daikoku, T. (2019b). Computational models and neural bases of statistical learning in music and language. *Phys. Life Rev*. doi: 10.1016/j.plrev.2019.09.001. [Epub ahead of print].

Daikoku, T. (2019c). Implicit learning in the developing brain: an exploration of ERP indices for developmental disorders. *Clin. Neurophysiol.* doi: 10.1016/j.clinph.2019.09.001. [Epub ahead of print].

Daikoku, T., Takahashi, Y., Futagami, H., Tarumoto, M., and Yasuda, H. (2017b). Physical fitness modulates incidental but not intentional statistical learning of simultaneous auditory sequences during concurrent physical exercise. *Neurol. Res.* 30, 107–116. doi: 10.1080/01616412.2016.1273571

Daikoku, T., Takahashi, Y., Tarumoto, M., and Yasuda, H. (2018). Auditory statistical learning during concurrent physical exercise and the tolerance for pitch, tempo, and rhythm changes. *Motor Control* 22, 233–244. doi: 10.1123/mc.2017-0006

Daikoku, T., Yatomi, Y., and Yumoto, M. (2014). Implicit and explicit statistical learning of tone sequences across spectral shifts. *Neuropsychologia* 63, 194–204. doi: 10.1016/j.neuropsychologia.2014.08.028

Daikoku, T., Yatomi, Y., and Yumoto, M. (2015). Statistical learning of music- and language-like sequences and tolerance for spectral shifts. *Neurobiol. Learn. Mem.* 118, 8–19. doi: 10.1016/j.nlm.2014.11.001

Daikoku, T., Yatomi, Y., and Yumoto, M. (2016). Pitch-class distribution modulates the statistical learning of atonal chord sequences. *Brain Cogn.* 108, 1–10. doi: 10.1016/j.bandc.2016.06.008

Daikoku, T., Yatomi, Y., and Yumoto, M. (2017a). Statistical learning of an auditory sequence and reorganization of acquired knowledge: a time course of word segmentation and ordering. *Neuropsychologia* 95, 1–10. doi: 10.1016/j.neuropsychologia.2016.12.006

Daikoku, T., and Yumoto, M. (2017). Single, but not dual, attention facilitates statistical learning of two concurrent auditory sequences. *Sci. Rep.* 7:10108. doi: 10.1038/s41598-017-10476-x

Daikoku, T., and Yumoto, M. (2019). Concurrent statistical learning of ignored and attended sound sequences: an MEG study. *Front. Hum. Neurosci.* 13:102. doi: 10.3389/fnhum.2019.00102

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Oxford: MIT Press.

Elmer, S., and Lutz, J. (2018). Relationships between music training, speech processing, and word learning: a network perspective. *Ann. N. Y. Acad. Sci.* 1423, 10–18. doi: 10.1111/nyas.13581

François, C., Chobert, J., Besson, M., and Schön, D. (2012). Music training for the development of speech segmentation. *Cereb. Cortex* 23, 2038–2043. doi: 10.1093/cercor/bhs180

Francois, C., and Schön, D. (2011). Musical expertise boosts implicit learning of both musical and linguistic structures. *Cereb. Cortex* 21, 2357–2365. doi: 10.1093/cercor/bhr022

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., Doherty, J. O., and Pezzulo, G. (2016). Neuroscience and Biobehavioral Reviews Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–224. doi: 10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369. doi: 10.1098/rstb.2013.0481

Gupta, S. D., and Bahmer, A. (2019). Increase in mutual information during interaction with the environment contributes to perception. *Entropy* 21:365. doi: 10.3390/e21040365

Hansen, N. C., and Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Front. Psychol.* 5, 1–17. doi: 10.3389/fpsyg.2014.01052

Hasson, U. (2017). The neurobiology of uncertainty: implications for statistical learning. *Phil. Trans. R. Soc. B* 372:20160048. doi: 10.1098/rstb.2016.0048

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nat. Rev. Neurosci.* 9, 304–313. doi: 10.1038/nrn2332

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Ann. Rev. Psychol.* 55, 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Kim, S. G., Kim, J. S., and Chung, C. K. (2011). The effect of conditional probability of chord progression on brain response: an MEG study. *PLoS ONE* 6:17337. doi: 10.1371/journal.pone.0017337

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Lesage, E., Morgan, B. E., Olson, A. C., Meyer, A. S., and Miall, R. C. (2012). Cerebellar rTMS disrupts predictive language processing. *Curr. Biol.* 22, R794–R795. doi: 10.1016/j.cub.2012.07.006

Moberget, T., Gullesen, E. H., Andersson, S., Ivry, R. B., and Endestad, T. (2014). Generalized role for the cerebellum in encoding internal models: evidence from semantic processing. *J. Neurosci.* 34, 2871–2878. doi: 10.1523/JNEUROSCI.2264-13.2014

Monroy, C., Meyer, M., Gerson, S., and Hunnius, S. (2017c). Statistical learning in social action contexts. *PLoS ONE* 12, 1–20. doi: 10.1371/journal.pone.0177261

Monroy, C. D., Gerson, S. A., Domínguez-Martínez, E., Kaduk, K., Hunnius, S., and Reid, V. (2017a). Sensitivity to structure in action sequences: an infant event-related potential study. *Neuropsychologia* 126, 92–101. doi: 10.1016/j.neuropsychologia.2017.05.007

Monroy, C. D., Gerson, S. A., and Hunnius, S. (2018). Translating visual information into action predictions: statistical learning in action and nonaction contexts. *Memory Cogn.* 46, 600–613. doi: 10.3758/s13421-018-0788-6

Monroy, C. D., Meyer, M., Schröer, L., Gerson, S. A., and Hunnius, S. (2017b). The infant motor system predicts actions based on visual statistical learning. *Neuroimage* 185, 947–954. doi: 10.1016/j.neuroimage.2017.12.016

O'Reilly, J. X., Jbabdi, S., and Behrens, T. E. J. (2012). How can a Bayesian approach inform neuroscience? *Eur. J. Neurosci.* 35, 1169–1179. doi: 10.1111/j.1460-9568.2012.08010.x

Parr, T., and Friston, K. J. (2018). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Parr, T., Rees, G., and Friston, K. J. (2018). Computational neuropsychology and Bayesian inference. *Front. Human Neurosci.* 12, 1–14. doi: 10.3389/fnhum.2018.00061

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., and Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage* 50, 302–313. doi: 10.1016/j.neuroimage.2009.12.019

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Topics Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress Neurobiol.* 134, 17–35. doi: 10.1016/j.pneurobio.2015.09.001

Przysinda, E., Zeng, T., Maves, K., Arkin, C., and Loui, P. (2017). Jazz musicians reveal role of expectancy in human creativity. *Brain Cogn.* 119, 45–53. doi: 10.1016/j.bandc.2017.09.008

Rohrmeier, M., and Cross, I. (2008). "Statistical properties of tonal harmony in Bach's Chorales," in *Proceedings of 10th International Conference on Music Perception and Cognition*. Retrieved from: http://icmpc10.psych.let.hokudai.ac.jp/%5Cnhttp://www.mus.cam.ac.uk/files/2009/09/bachharmony.pdf

Rohrmeier, M., and Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics Cogn. Sci.* 4, 525–553. doi: 10.1111/j.1756-8765.2012.01223.x

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4, 1–5. doi: 10.3389/fpsyg.2013.00710

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical J.* 27, 623–656.

Shimizu, R. E., Wu, A. D., Samra, J. K., and Knowlton, B. J. (2017). The impact of cerebellar transcranial direct current stimulation (Tdcs) on learning fine-motor sequences. *Philos. Transac. R. Soc. B Biol. Sci.* 372:20160050. doi: 10.1098/rstb.2016.0050

Tabachnick, B. G., and Fidell, L. S. (2007). *Using Multivariate Statistics, 5th edn*. New York, NY: Allyn & Bacon.

Wiggins, G. A. (2018). Creativity, information, and consciousness: the information dynamics of thinking. *Phys. Life Rev.* 1, 1–39. doi: 10.1016/j.plrev.2018.05.001

Yumoto, M., and Daikoku, T. (2016). "Basic function," in *Clinical Applications of Magnetoencephalography*, eds S. Tobimatsu and R. Kakigi (New York, NY: Springer Science + Business Media), 97–112.

Yumoto, M., and Daikoku, T. (2018). Neurophysiological studies on auditory statistical learning. *Jpn. J. Cogn. Neurosci.* 20, 38–43. doi: 10.11253/ninchishinkeikagaku.20.38

Zubicaray, G., Arciuli, J., and Mcmahon, K. (2013). Putting an "end" to the motor cortex representations of action words. *J. Cogn. Neurosci.* 25, 1957–1974. doi: 10.1162/jocn_a_00437

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Daikoku. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 04 October 2019
doi: 10.3389/fnins.2019.01053

Quantifying Neurodegenerative Progression With DeepSymNet, an End-to-End Data-Driven Approach

Danilo Pena[1,2], Arko Barman[1,2], Jessika Suescun[3], Xiaoqian Jiang[1], Mya C. Schiess[3], Luca Giancardo[1,4] and the Alzheimer's Disease Neuroimaging Initiative[†]*

[1] School of Biomedical Informatics, University of Texas Health Science Center at Houston (UTHealth), Houston, TX, United States, [2] Center for Precision Health, UTHealth, Houston, TX, United States, [3] Department of Neurology, McGovern Medical School, UTHealth, Houston, TX, United States, [4] Center for Precision Health, UTHealth Diagnostic and Interventional Imaging, McGovern Medical School, UTHealth Institute for Stroke and Cerebrovascular Diseases, UTHealth, Houston, TX, United States

Alzheimer's disease (AD) is the most common neurodegenerative disorder worldwide and is one of the leading sources of morbidity and mortality in the aging population. There is a long preclinical period followed by mild cognitive impairment (MCI). Clinical diagnosis and the rate of decline is variable. Progression monitoring remains a challenge in AD, and it is imperative to create better tools to quantify this progression. Brain magnetic resonance imaging (MRI) is commonly used for patient assessment. However, current approaches for analysis require strong a priori assumptions about regions of interest used and complex preprocessing pipelines including computationally expensive non-linear registrations and iterative surface deformations. These preprocessing steps are composed of many stacked processing layers. Any error or bias in an upstream layer will be propagated throughout the pipeline. Failures or biases in the non-linear subject registration and the subjective choice of atlases of specific regions are common in medical neuroimaging analysis and may hinder the translation of many approaches to the clinical practice. Here we propose a data-driven method based on an extension of a deep learning architecture, DeepSymNet, that identifies longitudinal changes without relying on prior brain regions of interest, an atlas, or non-linear registration steps. Our approach is trained end-to-end and learns how a patient's brain structure dynamically changes between two-time points directly from the raw voxels. We compare our approach with Freesurfer longitudinal pipelines and voxel-based methods using the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Our model can identify AD progression with comparable results to existing Freesurfer longitudinal pipelines without the need of predefined regions of interest, non-rigid registration algorithms, or iterative surface deformation at a fraction of the processing time. When compared to other voxel-based methods which share some of the same benefits, our model showed a statistically significant performance improvement. Additionally, we show that our model can differentiate between healthy subjects and patients with MCI. The model's decision was investigated using the epsilon layer-wise propagation algorithm. We found that the

OPEN ACCESS

Edited by:
Mackenzie W. Mathis, Harvard University, United States

Reviewed by:
Blake A. Richards, Montreal Neurological Institute, Mcgill University, Canada
Bing Zhang, Nanjing Drum Tower Hospital, China

***Correspondence:**
Luca Giancardo
luca.giancardo@uth.tmc.edu

[†] *Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 July 2019*
Accepted: *19 September 2019*
Published: *04 October 2019*

Citation:
Pena D, Barman A, Suescun J, Jiang X, Schiess MC, Giancardo L and the Alzheimer's Disease Neuroimaging Initiative (2019) Quantifying Neurodegenerative Progression With DeepSymNet, an End-to-End Data-Driven Approach. Front. Neurosci. 13:1053. doi: 10.3389/fnins.2019.01053

predictions were driven by the pallidum, putamen, and the superior temporal gyrus. Our novel longitudinal based, deep learning approach has the potential to diagnose patients earlier and enable new computational tools to monitor neurodegeneration in clinical practice.

Keywords: Alzheimer's disease, deep learning, magnetic resonance imaging, progression, biomarkers, longitudinal, ADNI

INTRODUCTION

Alzheimer's disease (AD) is the leading cause of dementia globally (50–75%) and is distinguished by a progressive cognitive decline (Lane et al., 2018). Currently, 5.8 million Americans suffer AD, and by 2050, this number will rise to 14 million (Alzheimer's Association, 2016). Criteria for the diagnosis of probable AD is based on subjective clinical assessments (Pfeffer et al., 1982; Marshall et al., 2015). There are multiple treatments available that can ameliorate some of the symptoms, but none of these drugs alter the course of the disease, and inevitably the dementia progresses in all patients. Neuroprotective and other disease-modifying therapies are under active development, however, to demonstrate their efficacy, sensitive, and reproducible metrics to measure disease progression are urgently needed, particularly at the earliest stage of the disease when therapies are more likely to slow the neurodegenerative progression (Aisen et al., 2017).

MRI based biomarkers for AD have been widely studied. Multiple groups have used imaging data to understand how regional brain atrophy, connectivity, or physical proximity can serve as biomarkers for dementia (Lillemark et al., 2014) or how these can be used to develop deep learning networks for AD classification (Litjens et al., 2017). These techniques are being developed to improve the accuracy of and to provide a quantitative, data-driven approach for AD disease diagnosis (Weiner et al., 2017).

In the past decade, researchers explored many avenues of the AD classification problem using machine learning. Over the years, there have been extensive reviews summarizing the state of the art of these methods (Rathore et al., 2017; Pellegrini et al., 2018). Recently, some approaches include one imaging modality (typically MRI) (Long et al., 2017), multiple imaging modalities data (Zhang et al., 2011, 2012), brain connectivity (de Vos et al., 2018), and genetic data (Peng et al., 2016). The most commonly used machine learning models are support vector machines, though there is definitely diversity in techniques (Pellegrini et al., 2018). Many of these papers focus on AD vs. CN classification, but studies are also looking at other classification tasks such as CN vs. MCI (Samper-González et al., 2018). All these methods are used for diagnosing a patient based on a model that was trained on cross-sectional data.

There is a growing interest in using machine learning to understand disease progression, and this is made possible by the available datasets for neurodegenerative disorders (Marcus et al., 2010). Researchers have used this longitudinal data to create brain development trajectories used to predict the risk of developing AD (Lawrence et al., 2017), to develop clinical symptom trajectories (Bhagwat et al., 2018) to extract essential brain features in MCI classification (Huang et al., 2017; Sun et al., 2017), and to investigate different stages of AD progression from a multi-modal imaging standpoint (Gray et al., 2012; Rodrigues et al., 2014; Nozadi et al., 2018). Similar work in Parkinson's disease used longitudinal connectome data as a marker for neurodegenerative progression (Peña-Nogales et al., 2018). In areas such as genetics, longitudinal studies have proved beneficial by revealing more single nucleotide polymorphism phenotype associations than cross-sectional studies in AD research (Xu et al., 2014). Despite the progress, existing methods use a priori hypotheses and feature engineering in their neuroimaging processing pipeline. A typical example would be the computation of the volumetric changes on a predefined number of brain areas which are used as an input to a machine learning or statistical model. This approach has two main limitations: (1), it is bound to the a-priori selection of specific brain areas, making it impractical to model disease progressions that are not fully understood; (2), any error in the estimation of the brain areas metrics would negatively influence the machine learning or statistical model (i.e., garbage in, garbage out).

To overcome the limitation of engineered features, recent studies have used the concept of feature extraction through deep learning techniques. This allows researchers to automatically extract image representations from the raw voxels specific to the outcome needed. Studies have used deep learning to tune convolutional neural networks (CNNs) on MRI images (Backstrom et al., 2018) and to process multi-modal information including genetic and neuropsychological data (Spasov et al., 2019). Further, others have used deep learning to complete other related tasks like segmentation and brain parcellation (Li et al., 2017; Gibson et al., 2018). ADNI-based machine and deep learning reviews are being written as a result (Weiner et al., 2017). However, we are not aware of end-to-end feature learning approaches to measure longitudinal changes that does not require pre-defined brain areas or region of interests for training.

In this work, we propose to use a DeepSymNet-based model, a novel end-to-end deep learning architecture, to identify longitudinal neurodegenerative progression between structural MRI images with minimal preprocessing at two-time points. We adapt the DeepSymNet architecture presented by Barman et al. (2019) to identify structural brain differences by learning time-sensitive representation on a subject-level. The imaging preprocessing pipeline required by the architecture does not use a priori brain regions or non-rigid registration algorithms making the process more robust by having fewer steps throughout the pipeline and more efficient in terms of time required to

generate hypotheses and computational time than common longitudinal processing pipelines such as Freesurfer. This work has four main contributions: (1) a new neuroimaging pipeline to measure neurodegenerative progression that does not require pre-defined brain areas or region of interests for training, (2) comparable classification performance when compared with existing Freesurfer and voxel-based longitudinal pipelines for AD-relevant progression, (3) higher computational efficiency and generalizability to external dataset of mild cognitive impairment (MCI) subjects, and (4) analysis of the brain areas that drive the model's decision. This manuscript's code can be found at https://gitlab.com/lgianca/longitudinal-deepsymnet.

MATERIALS AND METHODS

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) in November 2018. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The ADNI data were downloaded in November 2018 (https://ida.loni.usc.edu/login.jsp). The data were then processed according to the Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016). This structure allows researchers to organize their neuroimaging-related data in a concise way that allows groups to access a growing number of computational tools and pipelines compatible with the BIDS format.

In this study, we included AD, MCI, and CN subjects from ADNI who had at least two T1-weighted brain images at least 6 months apart from ADNI1, ADNIGO, and ADNI2, and this resulted in 971 patients. When patients who had more than two imaging time points, we chose the first and last sessions. The two sessions are referred to as "session 1" and "session 2" throughout the paper. Further, all the patients chosen stayed within their disease phenotype at and between the two imaging sessions and successfully passed the image preprocessing pipeline. As shown in **Table 1**, after the above criteria were applied, we had a total of 632 subjects between the three groups. A Kruskal-Wallis test was performed on the cohort demographics. The statistical test succeeded to reject the null hypothesis that the samples originate from the same distribution for the sex and time between sessions variables. In our comparative analysis described later, we corrected for these potential confounders.

We will divide the methods and techniques we used into four main topics: (1) overviews of the Freesurfer-based, voxel-based, and DeepSymNet pipelines, (2) experimental design, (3) computational time analysis, (4) progression generalizability tests, and (5) confounding variable adjustment.

In the three sections, we provide an overview of the Freesurfer longitudinal pipeline, a description of the feature sets used for machine learning, and their accompanying experiments. Second, we describe the DeepSymNet-based pipeline, the experiments conducted, and the relevant brain area analysis. Third, we discuss the voxel-based machine learning method. Next, we describe comparison experiments for the above pipelines where we evaluated the computational time requirements. We then apply the models on an external MCI cohort to test the generalizability of the different pipelines and to identify progression patterns on this cohort at risk of developing AD. Finally, we describe how we evaluate the effect of potential confounders in the models.

Data Preprocessing and Feature Set Creation Using Freesurfer Pipelines

As seen in **Figure 1** below, Freesurfer extracts the brain region volumes using multiple predefined (and time intensive) steps such as within-subject template creation, atlas registration, non-linear transformations, surface inflating. Here, we aimed to compare the model performance using two different Freesurfer version's atlas-based, longitudinal pipelines. The pipeline features will be referred to as the first feature set (FS 1) and the second feature set (FS 2), respectively. For FS 1, the "University of California San Francisco's Longitudinal Freesurfer (5.1) All Available Base Image [ADNIGO, 2]" file downloaded from the ADNI website. This independent and external source was critical to ensure that our results were comparable with already published research. The FS 2 set was taken from the output of the Freesurfer (6.0) longitudinal pipeline. A full description of the Freesurfer longitudinal pipeline is beyond the scope of this paper, and we would like to refer readers to the original publications for more information (Reuter et al., 2010, 2012; Iglesias et al., 2016). These steps are typically what researchers use in neuroimaging pipelines.

The patients between the two datasets (FS 1 and FS 2) were matched using the patient's RID (specific to ADNI's protocol) and by the closest visit date. A total of 93 subjects matched between the two datasets. For each patient, the structural MRI variables output from the pipelines were used as a feature vector. The FS 2 regional volume data came from the subcortical segmentation

TABLE 1 | Demographics and time between imaging sessions of the AD, CN, and MCI patients used in this study (one standard deviation–s.d.).

	AD	CN	MCI	*p*-value
Number of patients	212	270	150	
First session age, years [mean (s.d.)]	74.8 (7.7)	74.7 (5.9)	75.3 (7.3)	0.381
Time between sessions, years [mean (s.d.)]	1.6 (0.6)	4.4 (2.6)	3.4 (2.8)	<0.001
Sex [male, n (%)]	108 (50.9%)	138 (51.1%)	99 (66%)	<0.01
Race/Ethnicity [n (%)] -	0 (0%)	0 (0%)	0 (0%)	0.403
American Indian/Alaskan Native,	4 (1.9%)	6 (2.2%)	3 (2.0%)	
Asian	0 (0%)	0 (0%)	0 (0%)	
Native Hawaiian/Other Pacific Islander	12 (5.6%)	20 (7.4%)	8 (5.3%)	
Black/African American	193 (91.1%)	242 (89.6%)	139 (92.7%)	
White, More than one race	3 (1.4%)	2 (0.8%)	0 (0%)	
Unknown				

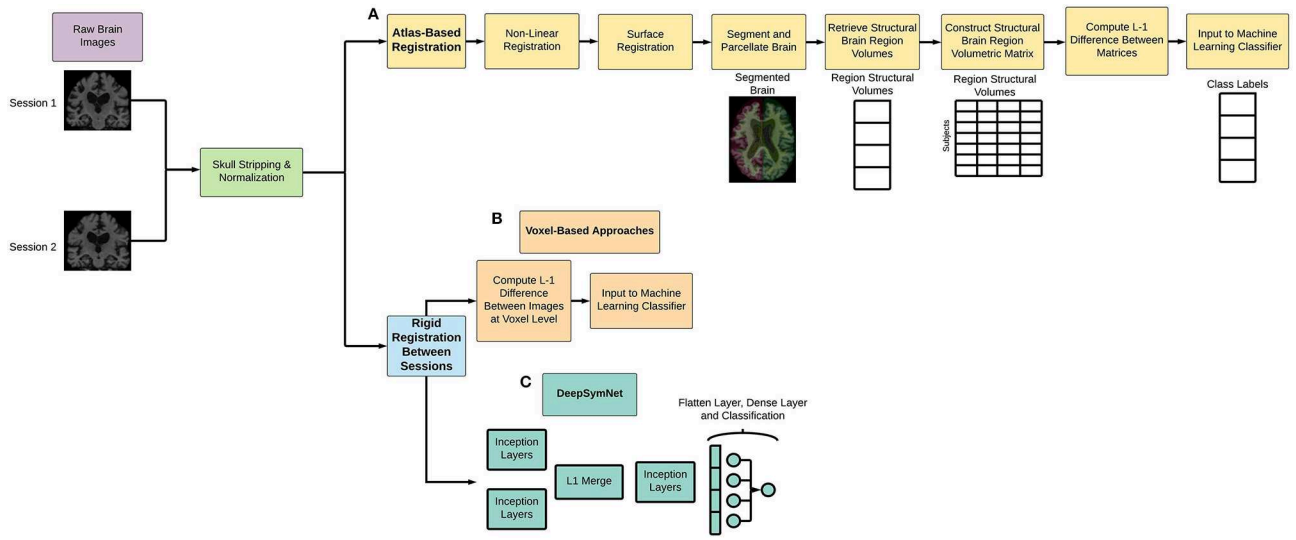


FIGURE 1 | Overview of **(A)** atlas-based and non-linear registration imaging preprocessing pipeline, **(B)** voxel-based methods, and **(C)** DeepSymNet-based pipeline implementation. Note that all processes involve rigid registration to align the patient's brains longitudinally.

file (aseg.stats) and white matter parcellation files (wmparc.stats) uniquely using the region sizes. The FS 1 pipeline took data from both the header and body of the segmentation files in addition to additional statistics data from the parcellation files (e.g., thickness average and standard deviation, and surface area) which resulted in a larger feature set. For both pipelines, the same cortical and subcortical regions were investigated. Finally, the FS 2 pipeline was compared to the DeepSymNet pipeline using both the limited cohort of 93 patients and the full cohort.

DeepSymNet-Based Pipeline Overview

Data Preprocessing

For the DeepSymNet preprocessing implementation, a simple longitudinal pipeline that used skull stripping, normalization, and a patient-specific alignment (**Figure 1**). This pipeline has several advantages for longitudinal studies. First, it uses rigid registration, which decreases computational and time costs associated with imaging pipelines. Second, there is no dependence on predefined brain regions, which enables us to take a more data-driven approach to understand progression. Third, the pipeline finds the best space to register the individual patient to which reduces the overall noise and bias that occurs when comparing samples between each other. The output of the MRI images was $182 \times 218 \times 182$ with a resolution of $1 \times 1 \times 1$ mm^3. Empty voxels outside of the brain were deleted from the image and not used in the subsequent analyses.

DeepSymNet Architecture Overview

The Deep Symmetry-Sensitive Convolutional Neural Network (DeepSymNet) architecture (**Figure 2**) used in this study was inspired from a model designed to identify spatial symmetries in brain angiograms (Barman et al., 2019). We applied this architecture to identify changes through time as opposed to spatial symmetries. This enables the architecture to directly learn a representation sensitive to intra-patients changes, rather than model the complex inter-patient heterogeneity and measure it over time. The model receives as input two brains at different time points, an initial step that learns a common representation between the two time points by 3-dimensional (3-D) Inception modules with weight sharing, which is then followed by a merge layer where the output of the filters is subtracted from one another, then, another set of 3-D Inception modules learn a representation sensitive to change. Finally, a max pooling and fully connected layer estimate the likelihood of a disease-relevant progression.

As AD is characterized by structural brain degeneration, a model like the one described has the potential of identifying the structural differences or progression patterns between the two MRI acquisition regardless of the brain appearance at during the first imaging session.

DeepSymNet Architecture Detailed

Here, we summarize the different components of DeepSymNet and walk the readers through our specific design.

Shared weights

Each of the two images is fed into identical neural networks before they are merged together. Additionally, these identical neural networks share weights, allowing them to learn the same patterns in the two images. This architecture allows the network to learn complex differences between the two imaging timepoints and encode information specific to that visit. Further, this part of the network can learn asymmetric patterns as neurodegeneration may differ between right and left hemispheres.

Inception module

The Inception modules are composed of multiresolution 3-D convolutional filters that learn to represent the T1 images at

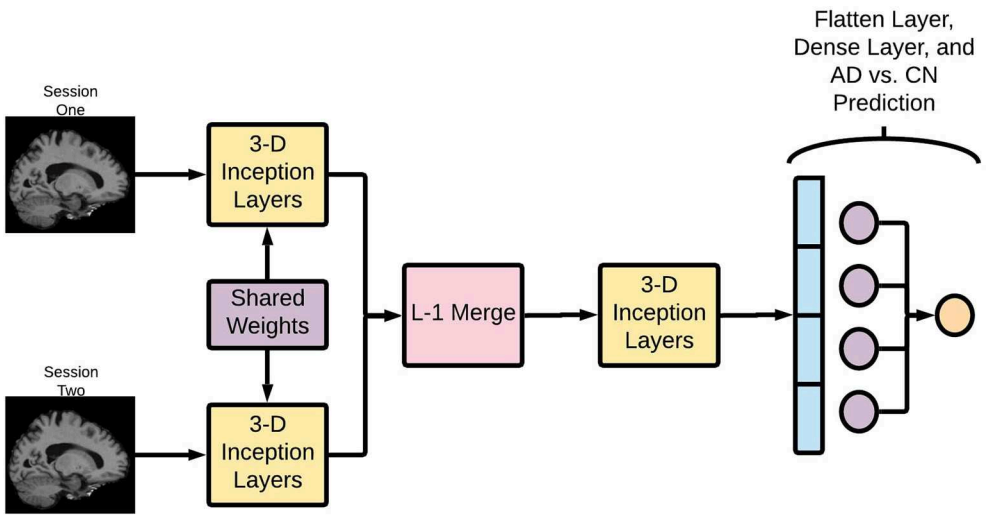


FIGURE 2 | Deep Symmetry-Sensitive Convolutional Neural Network (DeepSymNet) architecture overview. Longitudinal images go through a Siamese network composed of 3-D Inception modules that share weights. These outputs are passed through a L-1 merge layer that computes differences between these two sessions. Then, this is passed through a final 3-D Inception layer to learn from these differences. Lastly, these outputs are flattened and put through a dense layer for the final AD-progression prediction.

different time points according to a loss function. It should be noted that these Inception modules are the 3-D extension to what is presented by Szegedy et al. (2015), and the modules are not the same full Inception network architecture. The Inception module concatenates several parallel convolutional layers. It can be thought of as a mini network within a larger network. Since MRI images are 3-D images, 3-D convolutional filters are used within the Inception modules. For this application, the Inception module consisted of $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$ convolutions, followed by concatenation and a max pooling layer. These different convolution sizes enable the network to learn features at different scales, which could allow for more complex pattern recognition. Sixty-four filters have been used in the Inception modules and the network uses rectified linear unit (ReLU) activations.

Merge layer

An L-1 difference was then used to combine the learned convolutional features from both sessions. This encodes critical information about the structural brain changes between the timepoints, and this information is passed through another Inception layer.

Fully connected layers

After the merge and Inception layers, the difference between the two images is transformed into a feature vector. This vector is then passed through a fully connected layer, which creates a linear combination of the filter outputs from the penultimate layer of the network. A SoftMax operation is used as the activation function in the last layer for AD-progression prediction.

The DeepSymNet network was not designed to be fully translational invariant like the classical Inception Network (Szegedy et al., 2015). Rather, the algorithm is designed to be insensitive to the inevitable small registration inaccuracies between the two timepoints that may be present due to the rigid registration step. This is achieved through the aforementioned max pooling layers. Overall, this network is relatively shallow compared to some of the deep networks commonly reported in literature. The multiresolution filters within the inception modules in addition to the L-1 merge layer allow for complex image representation through space and time.

Implementation

The experiments were completed with Python (3.6.8). The DeepSymNet was implemented with Keras (2.2.4) with TensorFlow (1.12.0) as the backend. From a hardware perspective, we used Nvidia's Tesla V100 graphics cards with 32 GB RAM. The training times for each epoch varied depending on the number of model parameters, and these training lengths could range from 9 to 400 s. Each fold had 150 epochs, and a batch size of four was used. Early stopping based on a lack of improvement in validation loss for 30 epochs was also employed to reduce unnecessary computational cost. Binary cross-entropy was used as the loss function, and the Adam optimizer was used (Kingma and Ba, 2015).

Brain Region Relevance Analysis

In order to understand which voxels contributed to the model's predictions, the epsilon layer-wise relevance propagation (ϵ-LRP) method was used to analyze the contributions individual voxels on the final prediction (Bach et al., 2015). In summary, the ϵ-LRP method decomposes the output of deep learning architectures, on a sample level, into relevance scores in a backward fashion layer-by-layer. These scores can be projected on the pixel or voxel level of the input, which in our case is the

full brain MRI, and we used a heatmap to visualize the relative magnitude of the scores. The ϵ-LRP implementation of the open-source package DeepExplain (https://github.com/marcoancona/DeepExplain) was used (Ancona and Gross, 2017).

To create this visualization that demonstrated which parts of the brain were important for DeepSymNet's predictions on AD and CN subjects, we started with the chosen model after the hyperparameter tuning test discussed in the next section. We computed the subject-level relevance with the ϵ-LRP method for both imaging sessions, and we ensured that the voxel magnitudes used were from the test set subjects. For each n-th patient, we added the absolute value of all the patients' voxel magnitude, m, for both sessions together into a single brain volume, M_N.

$$M_N = \sum_{i\ =\ n}^{N} |m_n|$$

Where m is the voxel in the brain volume after registering to a common space and N represents the total number of patients.

Next, we wanted to determine the relative importance of each brain region using the magnitude of a voxel's relevance. We used the Harvard-Oxford cortical and subcortical maps to segment the brain into interpretable regions (Caviness et al., 1996). We used these respective maps to segment the voxels into different brain regions, R. We took the sum of all the voxel's magnitudes m_v^r, within each region, r, and then normalized this summation by the volume, V_r, of the respective regions. This would result in the normalized volume magnitude, M_r.

$$M_r = \frac{\sum_{v\ =1}^{V_r} m_v^r}{V_r}$$
$$\forall r = 1, 2, \ldots R$$
$$\forall v = 1, 2, \ldots V$$

This allowed us to see and interpret the relative progression-related importance of these different regions for the model's decision.

Voxel-Based Machine Learning

In addition to the models discussed above, we experimented with a set of general-purpose machine learning models receiving the same input as the DeepSymNet architecture (**Figure 1B**). To construct the feature vector for these models, the L-1 difference between the two images was calculated, and the resulting 3-D array was flattened into one dimension and given as input to the model. Linear support vector machine and random forest classifiers were chosen as preliminary models. However, in order to avoid any model bias, we also included a strategy that incorporated robust ensemble model construction through meta-learning and Bayesian optimization (Feurer et al., 2015). This strategy was completed using the AutoML library (https://automl.github.io/auto-sklearn/master/). The architecture here is essentially optimized over the parameter space to find the "optimal" solution in order to mitigate human negligence, which offers a good baseline for our comparison.

Experimental Design for Machine Learning Models

For the FS 1 and FS 2 sets, longitudinal based classification tasks were performed to assess the reliability between the two different pipelines. These tests used an L1-regularized logistic regression as the classifier, and the features (e.g., regional brain volumes) were scaled with respect to their inter-quartile range

$$\frac{x_r - Q_1\ (x_r)}{Q_3\ (x_r) - Q_1\ (x_r)}, \forall x_r, r_{FS} = 1, 2 \ldots R_{FS}$$

where the values, x_r, of brain region, r_{FS}, are transformed using the quartiles Q.

A 10-fold cross-validation was used where the split between training and test sets were 90% and 10%, respectively. For the longitudinal classification, each patient had two feature vectors (first and last sessions), and the L-1 difference between these vectors was used as the final feature vector. The cross-validation process with random splits was conducted 100 times, and the average classification probability was taken as the average of all of these test set trials. This previously validated method offers more reliable performance on relatively small datasets as it increases the number of cross-validations without decreasing the size of the test set (Pedregosa et al., 2011; Varoquaux et al., 2017). Sensitivity and specificity metrics were calculated by choosing a cutoff point that was the minimum distance from the upper left corner in the Area Under the Receiver Operating Curve (AUC ROC). Finally, to ensure the validity of this process, tests were also conducted solely using either session 1 or session 2 data. The results from these tests were comparable to existing literature that used machine learning approaches to distinguish AD from CN with structural brain regional volumes and the ADNI dataset. Discussion of these results are beyond the scope of this paper.

For the DeepSymNet pipeline, a single 10-fold cross-validation was used due to computational constraints. Training a cross-validation approach with random splits was not feasible. Each fold contained a train and validation set, and once the fold was completed, the metrics were evaluated on a separate test set. The train, validation, and test set's proportions were 80% and 10%, and 10%, respectively. Learning rate and regularization optimization tests were conducted manually. The voxel-based approaches followed the same 10-fold cross validation and data split schematic as the DeepSymNet. Within each fold, parameters were tuned using the validation set, and the probabilities from the test sets were taken into account.

The metrics from the models trained on FS 1, FS 2, and the voxel-based methods served as the baseline metrics to compare against the DeepSymNet architecture. We chose these Freesurfer-based models for two reasons: (1) Freesurfer is arguably the most used and tested image preprocessing approach to create a representation from T1-weighted volumes. Additionally, Freesurfer has a tested and well-recognized longitudinal pipeline used by multiple research groups around the globe. (2) To ensure that we were not biased to a specific FS version's representation, we employed two sets of features.

The metrics used to evaluate the models are the AUC ROC score, balanced accuracy, sensitivity, and specificity. Further

testing required the ROC curves to be compared, and these were tested for significance using the DeLong's test (DeLong et al., 1988). The ROC curves' confidence intervals were calculated using the model's predictions with a Monte Carlo resampling method with 1,000 iterations, with 80% of the data per iteration.

Computational Time Analysis

Ultimately, we are interested in developing the underlying algorithm enabling a measuring tool for clinicians or researchers who wants to quantify an AD-relevant progression from T1-weighted brain images. Therefore, we tested the time needed to go from the raw brain images to a prediction in order to assess the feasibility of using a similar application in a clinical setting or large clinical trials. We assumed that that all ML models were already trained, as it is standard for ML applications deployment. For these tests, we used a machine with 8 CPU cores and 1 GPU. We ran 450 iterations of each pipeline and computed the mean and standard deviation of the time needed to complete. All parallelization speed-up for the Longitudinal Freesurfer-based pipeline were enabled.

Generalizability: Detection of "AD-like" Progression Pattern on MCI Cohort

After the experiments, we wanted to see if the DeepSymNet model could apply the learned progression pattern on an external set of high-risk patients. The final DeepSymNet model was applied to a cohort of MCI patients from the ADNI protocol. Each of the 10-folds' respective best DeepSymNet model was applied to each MCI patient progression so that each patient had 10 prediction probabilities. These probabilities were averaged together for the final probability measurement. These prediction probabilities were then used to construct AUC ROC curves of MCI vs. CN. Note that the control group probabilities were taken from the test set from the 10-fold cross-validation process explained above for the MCI vs. CN AUC ROC curve. In addition, these MCI results from the DeepSymNet were compared against the same method from the logistic regression that used atlas-based registration pipeline outlined above.

Confounding Variable Adjustment

Finally, we adjusted the DeepSymNet's probability output for confounding variables through a logistic regression method. The confounders used were the time between the two imaging sessions, gender, and the baseline age at the first imaging session.

TABLE 2 | High-level comparison between the two feature sets from different imaging pipelines used to test the robustness of the new method.

	Freesurfer version	Number of samples used for ML pipeline (AD/CN)	Number of features
Feature set 1	5.1	93 (60/39)	340
Feature set 2	6.0	93 (60/39)	117

$$
\begin{aligned}
logit\, E(Y) &= \beta_0 + \beta_1{}^* X_1 + \beta_2{}^* X_2 + \beta_3{}^* X_3 + \beta_4{}^* X_4 \text{ where} \\
Y &= \text{Subject Group } (e.g.\ AD, CN, MCI) \\
X_1 &= \text{Time between sessions} \\
X_2 &= \text{Gender} \\
X_3 &= \text{Age} \\
X_4 &= \text{DeepSymNet Probability}
\end{aligned}
$$

The feature coefficients from the logistic regression model for these variables along with their 95% confidence interval and *p*-value were reported.

RESULTS

In this study, we aimed to examine: (1) various models' performance on learning AD-related progression patterns, (2) image resolution and network architecture hyperparameter tests on DeepSymNet performance, (3) the evaluation of the models on an external set of MCI patients, and (4) the influence of the selected confounding variables. The purpose of the first aim was to investigate the advantages of using an end-to-end data-driven approach to understand AD progression. The second aim allows readers to understand how the model behaved with hyperparameter tuning. The third aim validated the learned AD progression-specific pattern from DeepSymNet, and the fourth aimed ensured that these output probabilities were statistically significant.

Longitudinal Pipeline and Model Comparison

As seen in **Table 2**, the FS 1 and FS 2 pipelines had a differing number of structural volumetric features due to the reasons described in the Methods section. As seen in **Figure 3** and **Table 3** below, the DeepSymNet had the highest AUC ROC of all the methods. Of the machine learning voxel-based methods, the random forests approach performed the best. There were statistical differences in the performance between the random forest voxel-based method and the DeepSymNet. However, there was not a significant difference between the DeepSymNet and the model using Feature Set 2. This suggests that a DeepSymNet architecture learns a high-level representation of longitudinal changes that is, at least, as informative as the changes in brain regional volumes and outperforms the voxel-based general-purpose machine learning approaches tested. Finally, as seen in **Table 4**, the preprocessing time was much faster in the DeepSymNet and voxel-based methods vs. the traditional neuroimaging pipelines used in research.

DeepSymNet Hyperparameter Experiments

A non-exhaustive manual hyperparameter search for the optimal image resolution and configuration of Inception modules was conducted (**Table 5**). These tests were not completed with a grid-search method in order to conserve time and computational power. The resolution changes were varied between 10, 20, 25, 30, and 35 percent and were applied to the original image isotropically. Afterwards, the number of Inception modules

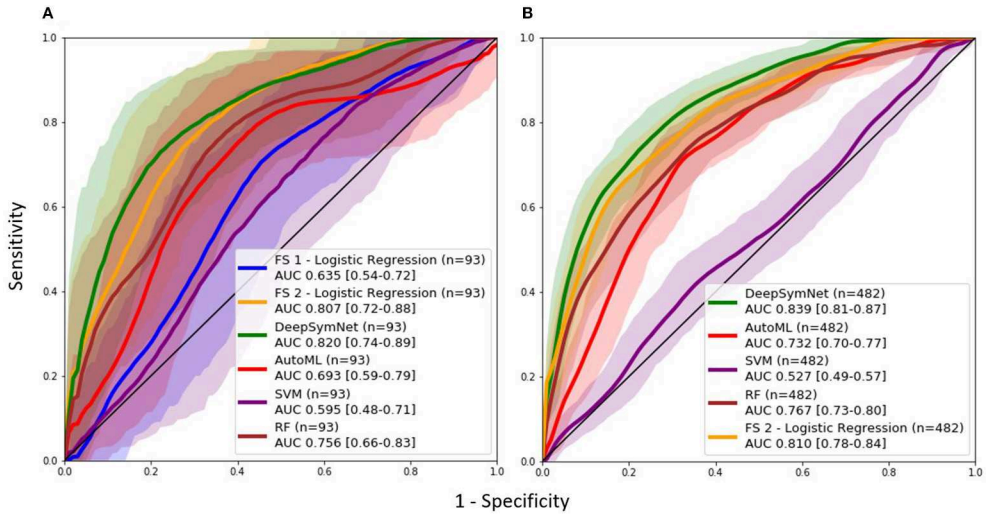


FIGURE 3 | AUC ROC curves comparing performance using a **(A)** data subset allowing for a comparison with samples found in Feature Set 1 (externally computed by UCSF) and the **(B)** full dataset.

TABLE 3 | Longitudinal models' metric evaluation for AUC ROC, sensitivity, specificity, and balanced accuracy on the full dataset.

Model	AUC ROC	Sensitivity	Specificity	Balanced accuracy	
FS 2 ($n = 482$)	81.0	75.9	71.9	72.3	N.S.
DeepSymNet ($n = 482$)	83.9	79.2	72.9	75.7	*** N.S.
Voxel-based random forest ($n = 482$)	76.7	68.4	72.6	64.2	

p-values were computed from the DeLong test for correlated ROC curves to reject the null hypothesis that there is no statistical difference between the AUCs.
N.S., not significant.
$^{***}p < 0.0001$.

TABLE 4 | Time performance for voxel-based pipeline (including DeepSymNet) and Longitudinal Freesurfer-based pipelines to generate an AD-relevant progression metric at inference time.

Pipeline	Computation time	
	Mean	Standard deviation
DeepSymNet and other voxel-based ML Methods	6.6 min	1.2 min
Longitudinal Freesurfer-based	17.06 h	2.7 h

TABLE 5 | Detailed view of DeepSymNet model tuning experimental AUC ROC results.

Experiment	Image resolution	Module configuration (Before/After)	AUC ROC
Input image resolution	[10, 20, **25**, 30, 35%]	1/1	83.9
Inception module configuration	25%	[**1/1**, 1/2, 2/1, 2/2, 3/3]	83.9

Input image resolution and Inception module configuration were examined. The tests were completed in a stepwise fashion as outlined in the table. Bolded values indicate the model chosen for further analysis.

before and after the merge layer were varied to find the optimal model.

Though the model with the 35% image resolution had a higher AUC ROC, the model that used 25% image resolution was chosen as the final model for two reasons. This model had fewer parameters (8.6 M vs. 21.2 M), and there was no statistical difference between the two curves. Several tests were conducted where the number of Inception modules before and after the merge layer was changed. As seen in **Figure 4**, the model that performed the best in this test contained 1 and 1 Inception modules before and after the merge layer, respectively.

DeepSymNet Output

The distribution of classification probabilities AD-like progression estimated by the DeepSymNet architecture was visualized in **Figure 5** below. As expected, the MCI subjects' probabilities were in between the CN and AD groups. This indicates that the MCI group have structural progression patterns that are similar to AD. Additionally, the MCI cohort qualitatively had a smaller interquartile range compared to the other two classes.

DeepSymNet Brain Region Relevance Analysis

After the model evaluation was completed, we wanted to understand which regions of the brain were relevant for the model's decision. The brain regions of interest are visualized both globally and based on location (subcortical

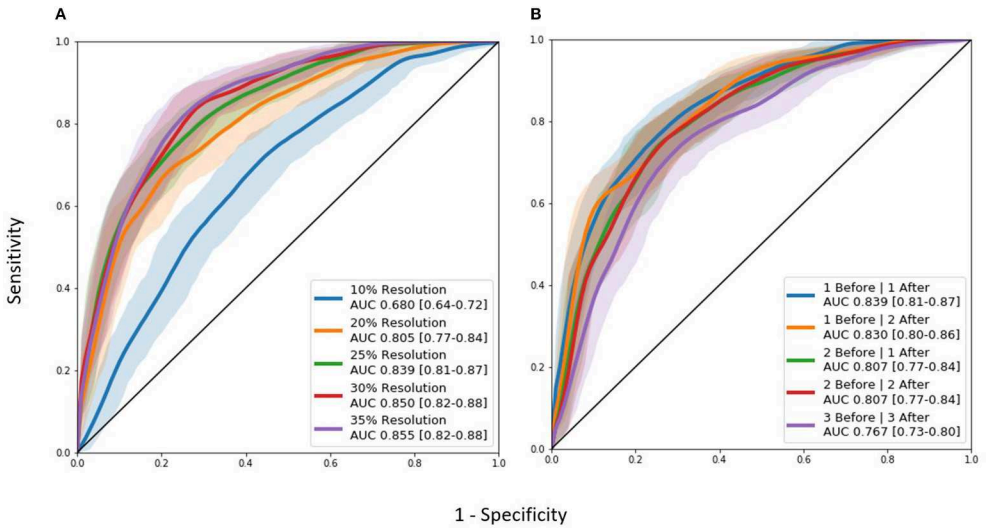


FIGURE 4 | AUC ROC curves for hyperparameter tuning experiments: **(A)** Image resolution experiments. **(B)** Differing Inception module architecture experiments. "Before" and "after" labels represent the number of 3D Inception modules before and after the L-1 layer where the two timepoints are combined together.

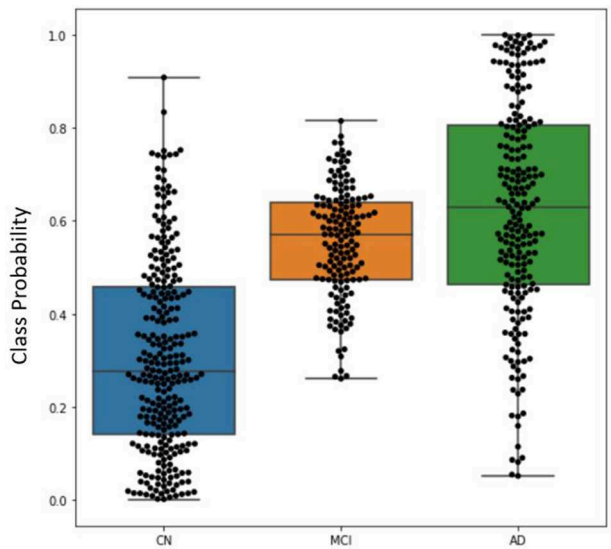


FIGURE 5 | Boxplot visualizing the distribution of DeepSymNet's output probabilities for the three classes, where each dot represents the progression probability for one subject. The class probability can be used as an indicator of AD-like longitudinal progression over two timepoints.

vs. cortical) in **Figure 6**. The palladium, white matter, and putamen subcortical regions had the highest overall relevance magnitude across all the subjects and sessions. Of the cortical regions, the superior temporal gyrus cortex anterior division had the highest magnitude. The top five activated regions from the subcortical and cortical areas are summarized in **Figure 7**. All of the cortical and subcortical regions and their relevance magnitude can be found in the **Supplementary Tables**.

Models Evaluated on an External Set of MCI Patients

Further, the final DeepSymNet model was assessed on an external set of MCI patients that DeepSymNet was never trained on. As seen in **Figure 8**, the AUC ROC score for identifying an AD-like progression in the MCI cohort for DeepSymNet was significantly higher than the machine learning models trained on brain regional volumes, i.e., Freesurfer-based, or voxels-based. Additionally, the AUC ROC performance is comparable to the AD vs. CN progression prediction task. This indicates DeepSymNet's ability to generalize an AD-progression specific pattern that is applicable to high-risk patients and that could not be achieved by MRI-based regional volume measures.

Confounding Variable Adjustment

Finally, the DeepSymNet's output probabilities for the AD vs. CN and MCI vs. CN prediction tasks were adjusted using logistic regression with confounders. In both tasks, the output probabilities taken from DeepSymNet were statistically significant ($p < 0.0001$) and had the highest coefficients relative to the potential confounding variables (**Table 6**).

DISCUSSION

In this study, our novel deep learning architecture, DeepSymNet, learned from temporal differences on the individual level to quantify AD progression. The DeepSymNet architecture combines the benefit of distance-based objective functions (which typically require smaller datasets) with prediction error-based objective functions (which lead to higher classification performance). The regions of the brain that drove the model's decision were visually analyzed using epsilon layer-wise relevance propagation methods. In addition, the

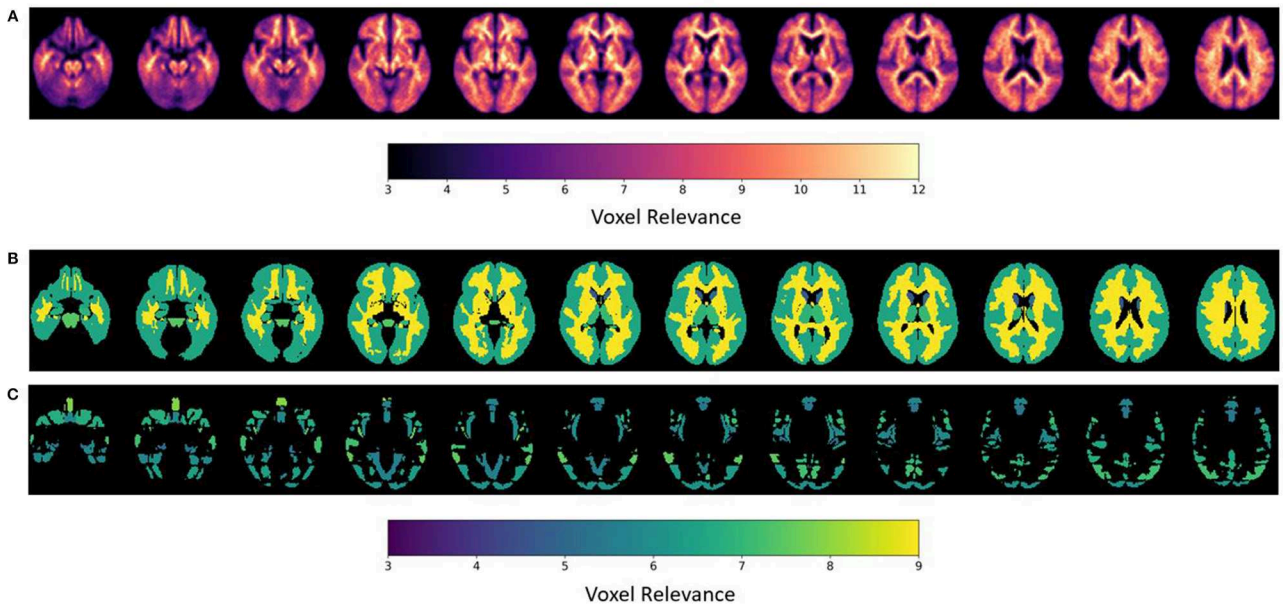


FIGURE 6 | ε-LRP relevance maps indicating the contribution of each voxel and brain region to the AD-progression classification at group level. **(A)** Voxel-based relevance map; **(B,C)** normalized relevance maps mapped to brain regions; **(B)** sub-cortical **(C)** cortical. Relevance scales help denote the degree of importance for the model's decision.

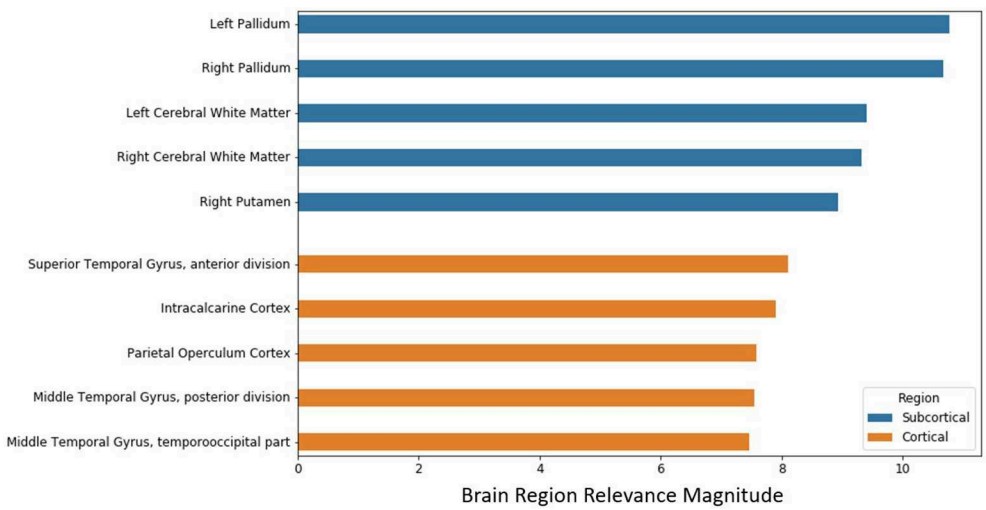


FIGURE 7 | Brain regional relevance magnitudes of top five subcortical and cortical region, sorted in descending order. These values were computed by summation of all the activations within the respective region and normalization by regional volume.

DeepSymNet pipeline did not use typical image preprocessing steps, predefined brain regions, or non-rigid registration algorithms. These commonly used steps can be a significant source of downstream bias and computational cost. The robustness of our pipeline was benchmarked against pipelines that used atlas-based methods and baseline voxel-based machine learning models. The DeepSymNet architecture and imaging pipeline is disease-agnostic and could be used for other problems that utilize brain imaging for measuring disease progression.

AD is an ideal case study for this work as there is no current established framework to numerically quantify AD neurodegenerative progression. For clinicians to properly test new disease-modifying drugs, there is a need to develop tools that quantify AD degeneration with commonly used brain imaging scans such as MRI. Many studies have found that AD-relevant changes are visible on the T1-weighted images, and all the AD population's progression will at some point appear. This idea was supplemented with longitudinal data where our model learned from the differences between the two imaging time points.

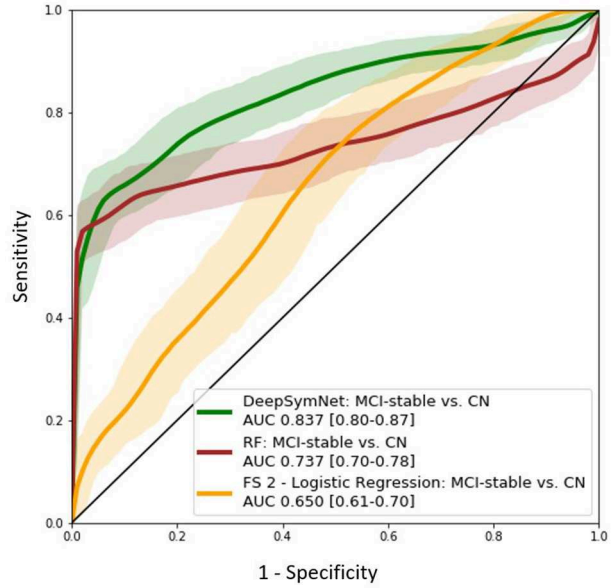


FIGURE 8 | AUC ROC curves for DeepSymNet, Random forests, and logistic regression (using Feature Set 2) for the identification of an AD-like progression in the MCI ($n = 150$) and CN ($n = 270$) groups. The models were not retrained on the MCI group, as the MCI patients were used only at inference time.

There are several advantages to learning from time differences. First, we effectively reduce the risk of outside confounders affecting the experimental results as the individual patient's data is registered to a common space within their own specific longitudinal data. The L-1 difference used between the same subject will also reduce the magnitude of the cross-sectional patient-specific features and will instead magnify the structural differences over time. Lastly, the preprocessing method that considers individual brain morphology is in line with providing individualized precision medicine.

In order to compare the robustness of the novel longitudinal pipeline, we compared models that used either pre-defined brain region volumes or voxels as to its input against DeepSymNet. DeepSymNet could represent AD-progression comparable to region-based methods and better than voxel-based methods as indicated from the superior performance over voxel-based methods. We believe this improved performance was due to the model learning both representations of the brain and the differences between the two timepoints. Further, our deep learning model was entirely data-driven and had fewer preprocessing steps.

Two experimental hyperparameter tuning tests were conducted to improve the model performance: image resolution and Inception module architecture. Due to time and computational constraints, the authors explored through a narrow search space for tuning the model. The DeepSymNet model chosen for further analysis used images with 25% resolution and had one Inception module before and after the merge layer. This model had an AUC ROC of 0.84 (0.81–0.87).

Once tuned, DeepSymNet pipeline was then applied to an external set of MCI patients. From **Figure 8**, we saw an improved AUC ROC performance MCI vs. CN progression identification (0.84). This indicated that the model was able to generalize AD-specific progression patterns that are also seen in prodromal MCI patients. The DeepSymNet pipeline also achieved a performance similar to the AD vs. CN prediction task. A non-perfect classification was expected as the MCI cohort is a heterogeneous group where not all subjects will develop AD. Additionally, both classification task probabilities were statistically significant after adjusting for confounders. Finally, as seen in **Figure 5**, there was a clear trend where the MCI probabilities were between the CN and AD probabilities, which may indicate the degree of neurodegenerative progression.

Once the model experiments were completed, the top activated brain regions were analyzed. Previous AD studies corroborate our subcortical and cortical regional findings. Researchers found significant white matter reductions in AD patients throughout the brain, particularly in the temporal lobe (Guo et al., 2010) and elevated mean diffusivity in precuneus and entorhinal white matter microstructures (Kantarci et al., 2017). The pallidum region was found to have a significant difference in beta-amyloid burden between early and late-onset AD (Youn et al., 2017) and differences in RNA binding protein TDP-43 deposits (Josephs et al., 2016). Finally, researchers using different imaging modalities to discriminate AD patients found the pallidum and putamen to be consistently important (Rondina et al., 2018). Various frontal regions were relevant for the model decision; this might represent an advanced disease stage in the selected population.

Further, we looked at evidence surrounding our cortical region findings. An AD disease progression timeline analysis found that the superior temporal gyrus anterior division was among the top biomarkers to first become abnormal (Venkatraghavan et al., 2019). Functional connectivity analysis found decreased connectivity in the superior temporal gyrus in dementia patients including AD (Hafkemeijer et al., 2015; Schwab et al., 2018). The middle temporal gyrus has been shown to atrophy significantly in both MCI and AD patients when compared to controls in longitudinal studies (Ghazi et al., 2019) and research that combined multi-modal data types (Convit et al., 2000; Korolev et al., 2016). Finally, Guo et al. found significant gray matter volume reductions in the superior and middle temporal gyrus (2010). These subcortical and cortical structural changes might provide insight into pathophysiology process of AD and potentially serve as biomarkers for identifying those who are at risk of developing AD.

There were several limitations in this study to note. As stated in the Methods section, the first and last imaging sessions were considered for this study. Many patients that had more than two visits with MRI imaging, which indicates that there is much more data available that could be incorporated into the model. Further, the hyperparameter tuning tests were all conducted manually, in a semi-structured way, and non-exhaustively. Machine learning methods like grid search or random grid search are used to find optimal network parameters. However, due to the computational costs of these methods at training time, they were not employed. Additionally, the time between the two imaging sessions for each patient was not controlled during the sample selection,

TABLE 6 | Summary of the logistic regression coefficients with associated confidence interval and p-values for DeepSymNet output probability and confounding variables in both classification tasks.

Classification task	Time between sessions	Gender	Baseline age	DeepSymNet output probability
Progression of AD vs. CN	1.25 (0.95 to 1.54)***	0.03 (−0.50 to 0.56)	−0.07 (−0.08 to −0.05)***	4.22 (3.01 to 5.44)***
Progression of MCI vs. CN	−0.2 (−0.11 to 0.08)	−0.41 (−0.89 to −0.06)	−0.04 (−0.05 to −0.03)***	6.41 (5.01 to 7.81)***

*** $p < 0.0001$.

which could present itself as a confounder. The time between sessions was corrected for and controlled in the statistical analysis (**Table 6**). Finally, a fully external validation of these results in other AD-specific imaging datasets was not conducted.

Future studies could expand the generalizability of AD classification by using other open-source datasets such as Open Access Series of Imaging Studies (https://www.oasis-brains.org/). Also, studies could look at improving the model performance by taking a Bayesian approach for hyperparameter tuning. Other work could make use of our pipeline as a progression phenotype to assess relationship with other data sources such as cerebrospinal/blood biomarkers or genetic data. Finally, newer models such as advances in recurrent neural networks could also incorporate more time points which may provide a richer representation of a patient's progression over time.

CONCLUSION

In summary, we implemented a novel pipeline based on a DeepSymNet architecture, that was able to detect an AD progression pattern by learning from structural differences of inter-subject MRI scans at two-time points. The paper's image preprocessing pipeline did not use predefined brain regions or non-rigid registration, which significantly reduced the opportunity for intermediate bias. In addition, the DeepSymNet pipeline was benchmarked against models that used standard imaging pipelines. From the brain region relevance analysis using the ε-LRP algorithm, the pallidum, putamen, and the superior temporal gyrus regions were critical in the model's final decision. Further, the model learned an AD progression pattern that was generalizable on an independent, external set of MCI patients. This architecture has the potential to be applied to multiple other applications where longitudinal changes need to be detected and measured. Finally, our pipeline can be used to improve imaging-based diagnostic systems by reducing time and computational cost.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available the public must request access from the Alzheimer's Disease Neuroimaging Initiative for access to the imaging data.

AUTHOR CONTRIBUTIONS

DP led the writing and carried out the computational experiments. AB and LG conceived the idea behind the methodology. LG and XJ reviewed and provided feedback for the methods. JS and MS assisted with the clinical relevance of the paper. All authors contributed to the writing of the final manuscript.

FUNDING

This work was supported by the Huffington Research Fund and from MS's Adriana Blood Endowed Chair in Neurology and partially supported by the Cancer Prevention and Research Institute of Texas grant RP 170668. Additionally, one of Titan XP GPUs used for this research was donated by the NVIDIA Corporation.

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01053/full#supplementary-material

REFERENCES

Aisen, P. S., Cummings, J., Jack, C. R., Morris, J. C., Sperling, R., Frölich, L., et al. (2017). On the path to 2025: understanding the Alzheimer's disease continuum. *Alzheimers Res. Ther.* 9:60. doi: 10.1186/s13195-017-0283-5

Alzheimer's Association (2016). 2016 Alzheimer's disease facts and figures. *Alzheimers Dement.* 12, 459–509. doi: 10.1016/j.jalz.2016.03.001

Ancona, M., and Gross, M. (2017). "A unified view of gradient-based attribution methods for Deep Neural Networks," in *Neural Information Processing Systems*, 1–11.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Backstrom, K., Nazari, M., Gu, I. Y.-H., and Jakola, A. S. (2018). "An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images," in *IEEE International Symposium on Biomedical Imaging (ISBI 2018)* (Washington, DC), 149–153. doi: 10.1109/ISBI.2018.8363543

Barman, A., Inam, M. E., Lee, S., Savitz, S., Sheth, S., and Giancardo, L. (2019). "Determining ischemic stroke from CT-angiography imaging using symmetry-sensitive convolutional networks," in *IEEE International Symposium on Biomedical Imaging* (Venice), 1–5. doi: 10.1109/ISBI.2019.8759475

Bhagwat, N., Viviano, J. D., Voineskos, A. N., Chakravarty, M. M., and Initiative, A. D. N. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* 14:e1006376. doi: 10.1371/journal.pcbi.1006376

Caviness, V. S., Meyer, J., Makris, N., and Kennedy, D. N. (1996). MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J. Cogn. Neurosci.* 8, 566–587. doi: 10.1162/jocn.1996.8.6.566

Convit, A., de Asis, J., de Leon, M., Tarshish, C., De Santi, S., and Rusinek, H. (2000). Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiol. Aging* 21, 19–26. doi: 10.1016/S0197-4580(99)00107-4

de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., et al. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *NeuroImage* 167, 62–72. doi: 10.1016/j.neuroimage.2017.11.025

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2015). "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, 2962–2970.

Ghazi, M. M., Nielsen, M., Pai, A., Cardoso, M. J., Modat, M., Ourselin, S., et al. (2019). Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. *Med. Image Anal.* 53, 39–46. doi: 10.1016/j.media.2019.01.004

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., et al. (2018). NiftyNet: a deep-learning platform for medical imaging. *Comput. Methods Programs Biomed.* 158, 113–122. doi: 10.1016/j.cmpb.2018.01.025

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44

Gray, K. R., Wolz, R., Heckemann, R. A., Aljabar, P., Hammers, A., Rueckert, D., and the Alzheimer's Disease Neuroimaging Initiative (2012). Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage* 60, 221–229. doi: 10.1016/j.neuroimage.2011.12.071

Guo, X., Wang, Z., Li, K., Li, Z., Qi, Z., Jin, Z., et al. (2010). Voxel-based assessment of gray and white matter volumes in Alzheimer's disease. *Neurosci. Lett.* 468, 146–150. doi: 10.1016/j.neulet.2009.10.086

Hafkemeijer, A., Möller, C., Dopper, E. G. P., Jiskoot, L. C., Schouten, T. M., van Swieten, J. C., et al. (2015). Resting state functional connectivity differences between behavioral variant frontotemporal dementia and Alzheimer's disease. *Front. Hum. Neurosci.* 9:474. doi: 10.3389/fnhum.2015.00474

Huang, M., Yang, W., Feng, Q., Chen, W., and the Alzheimer's Disease Neuroimaging (2017). Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. *Nat. Sci. Rep.* 7, 1–13. doi: 10.1038/srep39880

Iglesias, J. E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., and Reuter, M. (2016). Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *NeuroImage* 141, 542–555. doi: 10.1016/j.neuroimage.2016.07.020

Josephs, K. A., Murray, M. E., Whitwell, J. L., Tosakulwong, N., Weigand, S. D., Petrucelli, L., et al. (2016). Updated TDP-43 in Alzheimer's disease staging scheme. *Acta Neuropathol.* 131, 571–585. doi: 10.1007/s00401-016-1537-1

Kantarci, K., Murray, M. E., Schwarz, C. G., Reid, R. I., Przybelski, S. A., Lesnick, T., et al. (2017). White-matter integrity on DTI and the pathologic staging of Alzheimer's disease. *Neurobiol. Aging* 56, 172–179. doi: 10.1016/j.neurobiolaging.2017.04.024

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations* (San Diego, CA), 1–15.

Korolev, I. O., Symonds, L. L., Bozoki, A. C., and Alzheimer's Disease Neuroimaging Initiative (2016). Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS ONE* 11:e0138866. doi: 10.1371/journal.pone.0138866

Lane, C. A., Hardy, J., and Schott, J. M. (2018). Alzheimer's disease. *Eur. J. Neurol.* 25, 59–70. doi: 10.1111/ene.13439

Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., and Anderson, R. M. (2017). A systematic review of longitudinal studies which measure Alzheimer's disease biomarkers. *J. Alzheimers Dis.* 59, 1359–1379. doi: 10.3233/JAD-170261

Li, Q., Wu, X., Xu, L., Chen, K., Yao, L., and Li, R. (2017). Multi-modal discriminative dictionary learning for Alzheimer's disease and mild cognitive impairment. *Comput. Methods Programs Biomed.* 150, 1–8. doi: 10.1016/j.cmpb.2017.07.003

Lillemark, L., Sørensen, L., Pai, A., Dam, E. B., and Nielsen, M. (2014). Brain region's relative proximity as marker for Alzheimer's disease based on structural MRI. *BMC Med. Imaging* 14:21. doi: 10.1186/1471-2342-14-21

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Long, X., Chen, L., Jiang, C., Zhang, L., and Alzheimer's Disease Neuroimaging Initiative (2017). Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS ONE* 12:e0173372. doi: 10.1371/journal.pone.0173372

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi: 10.1162/jocn.2009.21407

Marshall, G. A., Zoller, A. S., Lorius, N., Amariglio, R. E., Locascio, J. J., Johnson, K. A., et al. (2015). Functional activities questionnaire items that best discriminate and predict progression from clinically normal to mild cognitive impairment. *Curr. Alzheimer Res.* 12, 493–502. doi: 10.2174/1567205012050526115003

Nozadi, S. H., Kadoury, S., and the Alzheimer's Disease Neuroimaging Initiative (2018). Classification of Alzheimer's and MCI patients from semantically parcelled PET images: a comparison between AV45 and FDG-PET. *Int. J. Biomed. Imaging* 2018, 1–13. doi: 10.1155/2018/1247430

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement.* 10, 519–535. doi: 10.1016/j.dadm.2018.07.004

Peña-Nogales, Ó., Ellmore, T. M., de Luis-García, R., Suescun, J., Schiess, M. C., and Giancardo, L. (2018). Longitudinal connectomes as a candidate progression marker for prodromal Parkinson's disease. *Front. Neurosci.* 12:967. doi: 10.3389/fnins.2018.00967

Peng, J., An, L., Zhu, X., Jin, Y., and Shen, D. (2016). Structured sparse kernel learning for imaging genetics based Alzheimer's disease

diagnosis. *Med. Image Comput. Comput. Assist. Interv.* 9901, 70–78. doi: 10.1007/978-3-319-46723-8_9

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Chance, J. M., and Filos, S. (1982). Measurement of functional activities in older adults in the community. *J. Gerontol.* 37, 323–329. doi: 10.1093/geronj/37.3.323

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Reuter, M., Rosas, H. D., and Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *NeuroImage* 53, 1181–1196. doi: 10.1016/j.neuroimage.2010.07.020

Reuter, M., Schmansky, N. J., Rosas, H. D., and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* 61, 1402–1418. doi: 10.1016/j.neuroimage.2012.02.084

Rodrigues, F., Silveira, M., and the Alzheimer's Disease Neuroimaging Initiative. (2014). Longitudinal FDG-PET features for the classification of Alzheimer's disease. *IEEE Eng. Med. Biol. Soc.* 2014, 1941–1944. doi: 10.1109/EMBC.2014.6943992

Rondina, J. M., Ferreira, L. K., de Souza Duran, F. L., Kubo, R., Ono, C. R., Leite, C. C., et al. (2018). Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: a comparison across functional and structural imaging modalities and atlases. *NeuroImage Clin.* 17, 628–641. doi: 10.1016/j.nicl.2017.10.026

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., et al. (2018). Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. *NeuroImage* 183, 504–521. doi: 10.1016/j.neuroimage.2018.08.042

Schwab, S., Afyouni, S., Chen, Y., Han, Z., Guo, Q., Dierks, T., et al. (2018). Functional connectivity alterations of the temporal lobe and hippocampus in semantic dementia and Alzheimer's disease. *BioRxiv* 1–27. doi: 10.1101/322131

Spasov, S., Passamonti, L., Duggento, A., Liò, P., and Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 189, 276–287. doi: 10.1016/j.neuroimage.2019.01.031

Sun, Z., van de Giessen, M., Lelieveldt, B. P. F., Staring, M., and the Alzheimer's Disease Neuroimaging Initiative (2017). Detection of conversion from mild cognitive impairment to Alzheimer's disease using longitudinal brain MRI. *Front. Neuroinform.* 11:16. doi: 10.3389/fninf.2017.00016

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/CVPR.2015.7298594

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Venkatraghavan, V., Bron, E. E., Niessen, W. J., Klein, S., and the Alzheimer's Disease Neuroimaging Initiative (2019). Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage* 186, 518–532. doi: 10.1016/j.neuroimage.2018.11.024

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2017). Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved AD clinical trials. *Alzheimers Dement.* 13, e1–e85. doi: 10.1016/j.jalz.2016.11.007

Xu, Z., Shen, X., Pan, W., and Initiative the A. D. N. (2014). Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS ONE* 9:e102312. doi: 10.1371/journal.pone.0102312

Youn, Y. C., Jang, J.-W., Han, S.-H., Kim, H., Seok, J.-W., Byun, J. S., et al. (2017). 11C-PIB PET imaging reveals that amyloid deposition in cases with early-onset Alzheimer's disease in the absence of known mutations retains higher levels of PIB in the basal ganglia. *Clin. Interv. Aging* 12, 1041–1048. doi: 10.2147/CIA.S132884

Zhang, D., Shen, D., and Initiative A. D. N. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE* 7:e33182. doi: 10.1371/journal.pone.0033182

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Pena, Barman, Suescun, Jiang, Schiess, Giancardo and the Alzheimer's Disease Neuroimaging Initiative. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

BRIEF RESEARCH REPORT
published: 04 October 2019
doi: 10.3389/fnins.2019.01046

Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates

David Xing[1†], *Mehdi Aghagolzadeh*[2†], *Wilson Truccolo*[2,3,4], *Erwan Bezard*[5,6], *Gregoire Courtine*[7,8,9,10‡] *and David Borton*[1,3,4*‡]

[1] School of Engineering, Brown University, Providence, RI, United States, [2] Department of Neuroscience, Brown University, Providence, RI, United States, [3] Carney Institute for Brain Science, Brown University, Providence, RI, United States, [4] U.S. Department of Veterans Affairs, Center for Neurorestoration and Neurotechnology, Providence, RI, United States, [5] CNRS, Institut des Maladies Neurodégénératives, UMR 5293, Bordeaux, France, [6] Université de Bordeaux, Institut des Maladies Neurodégénératives, UMR 5293, Bordeaux, France, [7] Center for Neuroprosthetics and Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Geneva, Switzerland, [8] Department of Clinical Neuroscience, Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Lausanne, Switzerland, [9] Defitech Center for Interventional Neurotherapies (NeuroRestore), CHUV/UNIL/EPFL, Lausanne, Switzerland, [10] Department of Neurosurgery, CHUV, Lausanne, Switzerland

The dynamical systems view of movement generation in motor cortical areas has emerged as an effective way to explain the firing properties of populations of neurons recorded from these regions. Recently, many studies have focused on finding low-dimensional representations of these dynamical systems during voluntary reaching and grasping behaviors carried out by the forelimbs. One such model, the Poisson linear-dynamical-system (PLDS) model, has been shown to extract dynamics which can be used to decode reaching kinematics. However, few have investigated these dynamics, especially in non-human primates, during behaviors such as locomotion, which may involve motor cortex to a lesser degree. Here, we focused on unconstrained quadrupedal locomotion, and investigated whether unsupervised latent state-space models can extract low-dimensional dynamics while preserving information about hind-limb kinematics. Spiking activity from the leg area of primary motor cortex of rhesus macaques was recorded simultaneously with hind-limb joint positions during ambulation across a corridor, ladder, and on a treadmill at various speeds. We found that PLDS models can extract stereotyped low-dimensional neural trajectories from these neurons phase-locked to the gait cycle, and that distinct trajectories emerge depending on the speed and class of behavior. Additionally, it was possible to decode both the hind-limb kinematics and the gait phase from these inferred trajectories just as well or better than from the full neural population (18-80 neurons) with only 12 dimensions. Our results demonstrate that kinematics and gait phase during various locomotion tasks are well represented in low-dimensional latent dynamics inferred from motor cortex population activity.

Keywords: low dimensional dynamics, locomotion, non-human primate (NHP), poisson linear dynamical system, primary motor cortex (M1)

OPEN ACCESS

Edited by:
Reza Lashgari, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Kyle B. Gustafson, Independent Researcher, Washington, DC, United States
Mohammadi Ali Ahmadi-Pajouh, Amirkabir University of Technology, Iran
Yousef Salimpour, Johns Hopkins University, United States
Alexander Gail, Deutsches Primatenzentrum, Germany

***Correspondence:**
David Borton
david_borton@brown.edu

[†]*These authors have contributed equally to this work*

[‡]*These authors share senior authorship*

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Neuroscience

Received: *12 June 2019*
Accepted: *17 September 2019*
Published: *04 October 2019*

Citation:
Xing D, Aghagolzadeh M, Truccolo W, Bezard E, Courtine G and Borton D (2019) Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates. Front. Neurosci. 13:1046. doi: 10.3389/fnins.2019.01046

1. INTRODUCTION

With the advent of large scale intracortical recordings allowing for the simultaneous interrogation of dozens to hundreds of neurons, the study of the role of motor cortex in the generation of movement has been steadily moving toward investigation of cortical dynamics in the brain on the population level. Ensembles of cortical neurons are organized as recurrently connected networks, introducing shared variability among the constituent cells, in turn constraining firing activity to a lower-dimensional space (Yu et al., 2009; Truccolo et al., 2010; Afshar et al., 2011; Churchland et al., 2012; Cunningham and Yu, 2014; Sadtler et al., 2014; Gallego et al., 2017; Pandarinath et al., 2018). Dimensionality reduction techniques can be used to extract the coordinated neural activity of a population, and reveal structures that may be hidden at the isolated single-neuron level. Previous investigations have shown that for reaching movements with the arm, semi-oscillatory dynamics, inferred using jPCA, are a common feature underlying a variety of different reach movements (Churchland et al., 2012). Additionally, a common low-dimensional neural manifold underlying various wrist movements have be found using demixed principle component analysis, or dPCA (Gallego et al., 2018).

Besides dimensionality reduction, many of these techniques also employ dynamical systems models. These models address how the values at the current time step depend on the values at previous time steps, usually in the form of a temporal transition matrix. Poisson Linear-Dynamical-System (PLDS) is one such technique which employs both unsupervised dimensionality reduction as well as explicit temporal dynamics. PLDS maps low-dimensional latent states to the measured high-dimensional neural spiking activity through an observation point-process model and explicitly estimates the dynamics of these latent states as it evolves through time with a linear mapping (Truccolo et al., 2005; Macke et al., 2011). Unlike PCA and dPCA, PLDS models the low-dimensional space as a state-space in a linear dynamical system, thereby explicitly accounting for the temporal relationships in the population. Although these low-dimensional dynamics are inferred through an unsupervised process, they are able to retain relevant behavioral information. In reaching behaviors, explicit state-space models similar to PLDS have been shown to increase closed-loop BMI performance in cursor-control tasks (Kao et al., 2015) and decoding accuracy during forelimb reaching behaviors (Aghagolzadeh and Truccolo, 2014, 2016), demonstrating that only a small number of dimensions from the neural population space is needed to capture the movement kinematics during voluntary forelimb movements.

However, reaching actions with the arm are typically highly precise and have a strong voluntary control component, whereas locomotion movements have a higher degree of autonomy. For example, in felines, injection of the neurotoxins or ablation of the motor cortex does not affect the ability of the animal to walk along a flat surface, although their ability to step over a ladder and over obstacles is impaired (Beloozerova and Sirota, 1988; Drew et al., 1996). These results suggest that motor cortex may play a less active role in the control of limb movements during basic, unobstructed locomotion compared to movements that require top-down voluntary control. In non-human primates, inducing corticalspinal tract lesions showed that although there is some locomotion deficits post-lesion, these recover quickly, while dexterous foot grasping remain severely impaired, even after 3 months post-lesion (Courtine, 2005).

These findings suggest that the contribution of M1 to the control of movements might be different during locomotion compared to during reaching, although how exactly the role of M1 is changing between these two behaviors is still not well understood. It has been well known for several decades that M1 is active during walking, and that cortical neurons are phasically tuned to the gait cycle (Drew, 1988; Beloozerova and Sirota, 1998; Drew et al., 2008). Yakovenko and Drew recorded from corticospinal neurons during reaching and during walking over an obstacle in cats and found that the firing onset phase of certain neurons correlated with the onset phase of muscle activation for both types of movements (Yakovenko and Drew, 2015), suggesting similar encoding of movements in M1 for both types behaviors. However, a recently published study in mice found that the population-level structure of M1 neurons is disparate during reaching and lever pulling compared to simple treadmill walking (Miri et al., 2017). These contrasting conclusions demonstrate that there is still a lack of consensus on the role of motor cortex during locomotion-related activities, and it is still unclear whether M1 is contributing to the control of the limbs in a similar manner during walking as during directed reaching movements. In particular, it has yet to be shown whether the latent state-space models that capture movement parameters during precise arm reaching would also be able to capture hind-limb movements during the potentially less engaging act of locomotion. While previous studies have found low dimensional representations of motor cortex activity in non-human primates during simple treadmill walking (Foster et al., 2014; Yin et al., 2014), they have not shown that movement kinematics are preserved in these dynamics.

Here, we aim to determine whether the PLDS latent-state model is able to extract low-dimensional dynamics which are informative of the limb movements. We define informative as having the ability to decode hind-limb joint kinematics as well as gait phase, during various locomotion tasks such as treadmill, corridor, and horizontal ladder walking (**Figure 1A**). To test this, we used either firing rates of the full recorded population of single neurons or the corresponding low dimensional dynamics as input features into a Wiener filter decoder which attempts to reconstruct the kinematics from these inputs. We demonstrate that for PLDS inputs, only a small number of dimensions are necessary to decode limb kinematics and gait phase as accurately as the full neural population. As far as we are aware, this is the first employment of explicit state-space models during both basic locomotion along a treadmill and corridor, as well as during directed locomotion along evenly and unevenly spaced ladders in nonhuman primates. Furthermore, it is also the first demonstration that cortical ensemble dynamics robustly captures behavioral information such as limb kinematics and gait phase during these different ambulatory behaviors.

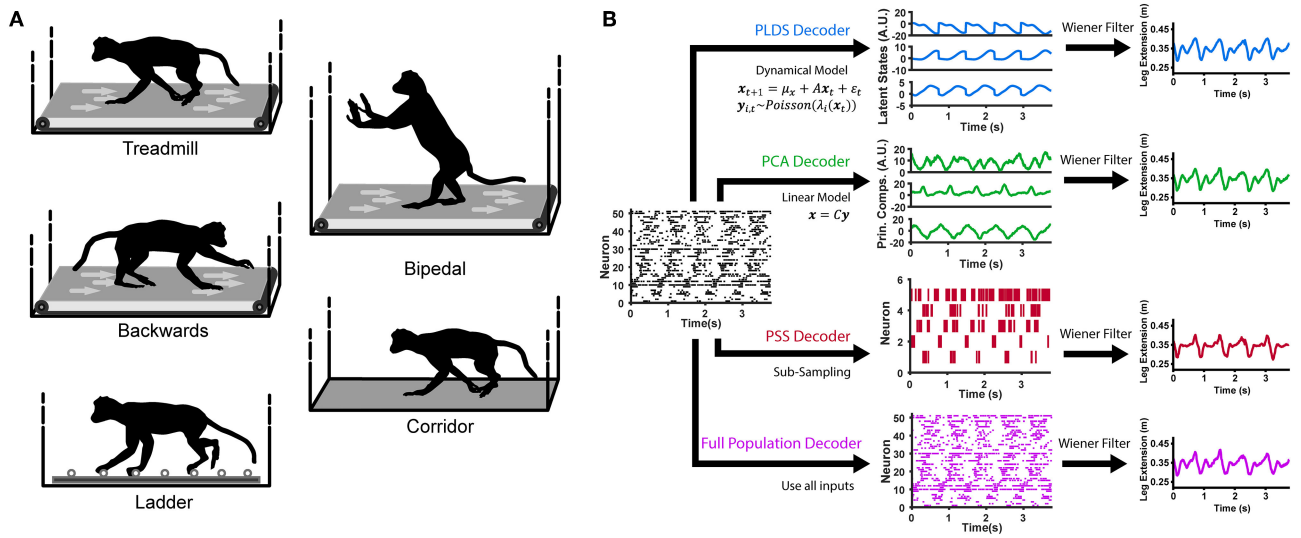


FIGURE 1 | Locomotion behaviors and neural decoding with latent state models. **(A)** Animals were trained to perform different locomotion tasks in a freely moving, untethered environment. The tasks include basic treadmill walking at various speeds, backwards treadmill walking, bipedal treadmill walking, corridor walking, and ladder walking. **(B)** Construction of the neural decoders. For decoders implementing dimensionality reduction (PLDS, PCA, PSS), rasters of neural spike counts for each gait cycle (left plot) undergo either an orthogonal linear transformation to extract principle components (center, top), a count-process filter based on a state-space model to extract low-dimensional latent trajectories (center, middle) or a greedy search algorithm to obtain an optimal subsample of neurons (center, bottom). These low-dimensional features are then used as inputs to a Wiener filter for reconstructing the estimate of a desired locomotion variable, such as the leg extension distance (right). The Full Population decoder (bottom path) performs decoding directly on all recorded normalized neural spike counts, binned at 100 ms. For decoding kinematics and gait, the trials of all tasks were combined and shuffled for each session.

2. METHODS

2.1. Surgery

Five male rhesus macaques between 5 and 8 years of age were implanted with 96-channel microelectrode arrays (Blackrock Microsystems, UT) in one hemisphere of the leg area of primary motor cortex (M1), located medially along the precentral gyrus (He et al., 1993). The details of the surgical implantation procedures have been described previously (Yin et al., 2014). Experiments complied with the European Union directive of September 22, 2010 (2010/63/EU) on the protection of animals used for scientific purposes in an AAALAC-accredited facility following acceptance of study design by the Institute of Lab Animal Science (Chinese Academy of Science, Beijing, China). Experiments were approved by the Institutional Animal Care and Use Committee of Bordeaux (CE50) under the license number 50120102-A.

2.2. Kinematic Data Collection

To obtain kinematic data, white reflective markers were painted over the shaved skin overlaying the right iliac crest (crest), greater trochanter (hip), lateral condyle (knee), lateral malleolus (ankle), 5th metatarsophalangeal (metatarsal), and outside tip of the fifth digit (toe tip). Marker locations were identified by feeling for the bony projections of the above anatomic landmarks under the skin; see Capogrosso et al. (2016) for a visualization of the marker locations. Videos were captured at 100 Hz from four high-speed cameras placed around the enclosures. Motion tracking software (Simi motion systems, Germany) was used to determine the 3D spatial coordinates of each marker, after calibrating the cameras in 3D space using known calibration objects at the beginning and end of each session. The origin of the coordinate system was set to the iliac crest of the animal. The horizontal axis was set to the direction of walking, while the vertical axis was set to the height off the ground. Joint angle was set as the inner angle between the two limb segments. The polar nature of the joint angles can confound error calculations (for example, 0° is closer to 350° than 180°, although 350–0 > 180–0), so instead of decoding the joint angles directly, we decoded the sine and cosine of each angle. Additionally, the total leg extension was calculated as the distance between the iliac crest and the metatarsal joint. Neural data, sampled at 30 kHz, was synchronized with the kinematic data using a camera-start trigger signal at the start of each trial of each recording session.

2.3. Tasks

Subjects carried out various behavioral motor tasks (**Figure 1A**). Monkeys were trained to walk in an enclosed treadmill at speeds of 1.1, 1.6, 2.4, 3.2, 4.0, 4.8, and 6.4 km/h (TRM trials). The enclosure was approximately one and a half meters long by one meter high by half a meter wide. It was constructed from clear Plexiglas, which is both visually and RF transparent. Spontaneously, they would switch to a bipedal gait, or start walking backwards (BIP and BACK trials, respectively). In addition, monkeys walked along a straight 3 m long corridor over a flat surface (CORR trials) and over a horizontal ladder with rungs either evenly spaced 35 cm apart or irregularly spaced (LAD trials). These tasks were self paced, with the

average corridor walking speed at 3.24 km/h (0.79 km/h standard deviation) and the average ladder walking speed at 3.49 km/h (0.79 m/s standard deviation). Monkeys were trained for at least 1 month to walk to the end of the corridor or ladder to receive a food reward in response to an auditory beep and flash of light.

Recording was carried out over 1–2 days for each monkey. Each day consisted of recording different randomly interleaved sessions for each task. Each treadmill session consisted of approximately 1 min of walking. These sessions were divided into epochs which contained only those time periods where the monkey was performing a single clear ambulatory behavior (e.g., epochs consisted of only bipedal walking, or only forward quadrupedal walking). Corridor and Ladder sessions consisted of walking back and forth along the corridor or ladder for a total of 20–30 rounds. These rounds contained only the steps performed at the middle of the corridor or ladder, to avoid the initiation and termination phase of locomotion.

2.4. Data Preprocessing

Neural data was obtained using the Blackrock Cerebus system, and saved onto a local computer through the Blackrock Central Suite program. The data was transmitted through a custom wireless headstage system (Yin et al., 2014), allowing for freely-moving recordings. The neural signal was band-passed at 500–5,000 Hz (second order Butterworth filter) and spikes were extracted by manual sorting using custom Matlab scripts (Laurens et al., 2013), which performed PCA on spike amplitudes and peak velocities (although in practice, any commonly accepted spike sorter can be used to obtain the firing rates). Neurons (the total number ranging between 18 and 80 for each session) were extracted from the multi-unit activity for each recording day (**Table 1**) and the spike times (counts) were binned into 10 ms intervals to match the 100 Hz kinematic data.

Each epoch was divided into trials consisting of a single gait cycle by manually marking the time point of foot-strike and toe-off. The stance phase of a single gait cycle was defined as the time period between the foot-strike and toe-off and represents 0–60% of the gait cycle while the swing phase was defined as the period between toe-off and the next foot-strike, and represents 60–100% of the gait cycle. The time-varying gait phase percentage was linearly interpolated from the foot-strike and toe-off time points (defined as the 0/100% and 60% mark, respectively).

Occasionally, neural data would become lost or corrupt for periods within a trial. Trials were manually inspected for data corruption, and if a trial was missing neural data, it was excluded from the analysis. All the trials shown in **Table 1** had complete neural data. Additionally, for two of the recording sessions, we did not record kinematics, and for the other seven sessions where kinematics were recorded, there were occasional trials where we were unable to obtain the kinematics (for example, due to video occlusion). **Table 1** contains the number of trials that contain usable kinematics for each task during each recording day. However, despite having incomplete kinematics for some of the trials, we were able to obtain gait phase data for all of the trials shown in **Table 1**, including in the two sessions where we were unable to obtain any kinematics. Therefore, our gait phase decoding analysis has a sample size of 9 (**Figure 3C**) while the kinematic decoding analysis has a sample size of 7 (**Figure 3D**).

2.5. Dimensionality Reduction Models

We used explicit state-space models to estimate the latent states of full population neural activity during locomotion. A LDS model assumes the neuronal ensemble activity as Gaussian linear observations, and uses expectation maximization (EM) learning to estimate the unknown model parameters and the latent states given only the observations (Macke et al., 2011; Aghagolzadeh and Truccolo, 2014). To account for the count process nature of ensemble spiking activity, we used a PLDS model, adding the assumption that the neural observations are conditionally Poisson given latent states (Truccolo et al., 2005; Aghagolzadeh and Truccolo, 2016). A Laplace approximation was used to compute the posterior density of the latent states given the neural observations. For decoding latent states from novel neural data, we used the mean of the state posterior density under the Laplace approximation as the estimate for the latent state (**Figure 1B**). The algorithm details have been described in Macke et al. (2011) and Aghagolzadeh and Truccolo (2016).

To compare PLDS with other dimensionality reduction techniques, we also tested decoders using low-dimensional inputs derived via principal component analysis (PCA), and also the activity of an optimal subsample of neurons from the full population, referred to as predictive subsampling (PSS). For PCA, the trials were concatenated across time and z-scored. The covariance matrix was computed, and the PCA projection matrix was constructed by eigenvalue decomposition—stacking the eigenvectors corresponding to the n largest eigenvalues, where n is chosen as the number of dimensions. We note that although PCA (and related methods) provide a low-dimensional representation of the ensemble activity, unlike PLDS, they do not explicitly account for any temporal dynamics in the latent states. PSS selects a subset of n neurons from the full population that optimize neural decoding through a greedy supervised learning algorithm, hence the name predictive. Details of this method can be found in a previous paper (Aghagolzadeh and Truccolo, 2016).

2.6. Comparison of Low-Dimensional Trajectories

To generate the example neural trajectories shown in **Figure 2**, we selected one recording session, Q1-day 1, which contained examples of all the tasks and treadmill speeds. In order to compute the average trajectories, the PLDS latent states were time-normalized to 0–100% of the gait cycle in steps of 1% using linear interpolation, with 0% representing the start of the trial (start of the stance phase) and 100% representing the last time point of the trial (end of the swing phase), while 60% was set to the transition between stance and swing. This resulted in 100 time points for each trial, and the values were averaged at each time point. The distance metric was calculated for each trial as the Mahalanobis distance between the values of that trial and the distribution of a selected reference trial type in the full 12 dimensional space. For the comparison across tasks, reference task was treadmill walking and for the comparison across speeds, the reference was the 6.4 km/h trials. The Mahalanobis distance values were averaged across trials for each of the tasks/speeds. The standard error of the mean (SEM) was also computed and shown for each task/speed.

TABLE 1 | Number of trials and neurons recorded for each recording day of each animal.

	Neurons	TRM trials	BIP trials	BACK trials	LAD trials	CORR trials	Trials with kinematics
Q1 Day1	56	96 (52)	43 (0)	4 (0)	35 (0)	18 (7)	59
Q1 Day2	80	91 (78)	61 (31)	4 (0)	14 (14)	5 (5)	128
Q2 Day1	63	129 (43)	12 (0)	0 (0)	6 (3)	6 (5)	51
Q2 Day2	51	217 (193)	22 (22)	0 (0)	22 (13)	4 (2)	230
Q3 Day1	46	77 (77)	6 (4)	0 (0)	10 (1)	5 (4)	86
Q3 Day2	41	79 (0)	0 (0)	0 (0)	16 (0)	12 (0)	0
Q4 Day1	51	224 (139)	0 (0)	0 (0)	0 (0)	0 (0)	139
Q4 Day2	18	51 (0)	0 (0)	0 (0)	3 (0)	20 (0)	0
Q5 Day1	39	37 (37)	0 (0)	0 (0)	26 (8)	9 (7)	52

We were unable to obtain kinematics in some of the trials; the number of trials that did contain kinematics for each task are displayed in parenthesis, and the total number of trials that contains kinematics across all tasks is displayed in the last column. All the trials shown here contain complete neural data and gait cycle data.

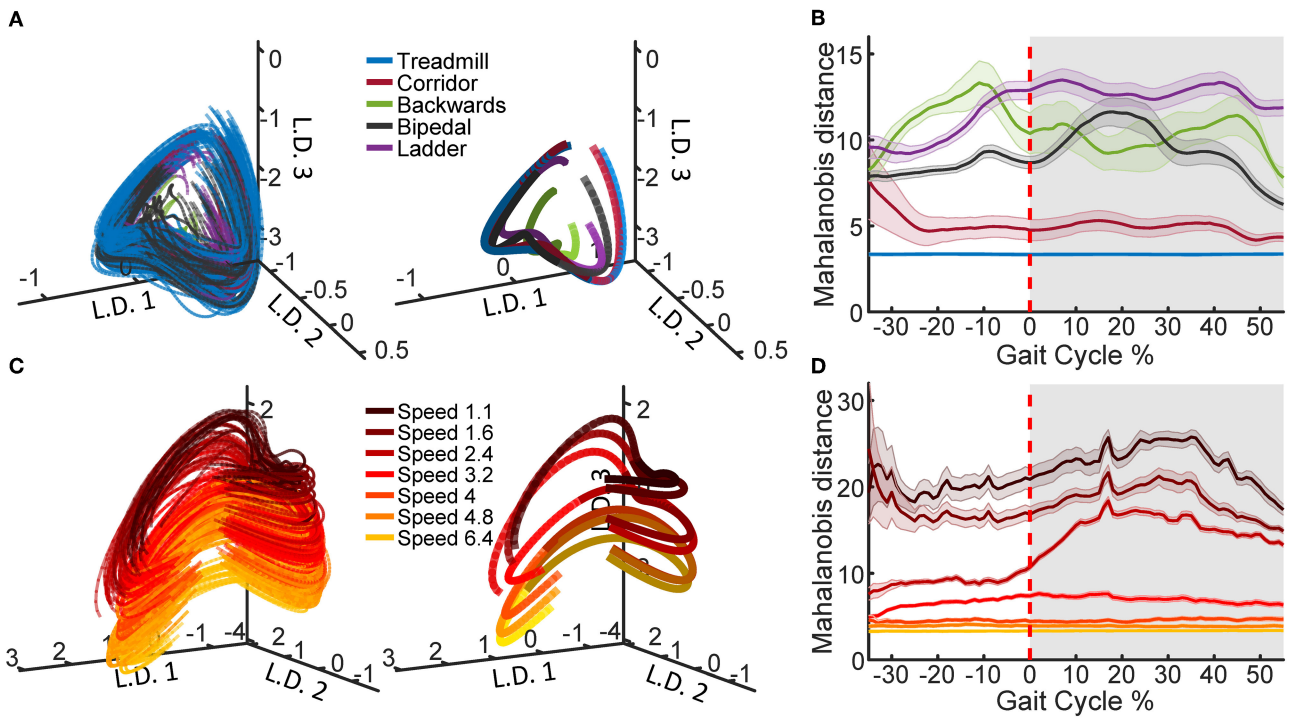


FIGURE 2 | Latent state space of neural population activity during locomotion. **(A)** Neural trajectories obtained from the PLDS model for the gait cycles of all tasks in animal Q1 on recording day 1. The trajectories in the first three latent dimensions (L.D.) for each individual gait cycle are shown (left), as well as task-averaged trajectories over the total gait cycle (right). **(B)** The Mahalanobis distance between the trajectories of each of the tasks and the trajectories of the treadmill task, calculated on all 12 dimensions. Shaded region denotes stance phase (0–60% of the gait cycle), shaded bars denote 1 S.E.M. **(C)** PLDS state-space trajectories in 3 dimensions for all treadmill walking trials at different locomotion speeds in animal Q2, session 2. Darker colors represent slower speeds; all speed values displayed in the legend are in km/h. As in **(A)**, individual gait cycles (left) and gait cycles averaged for each speed (right) are shown. For the average trajectories in both **(A,C)**, the darker shade represents the stance phase while the lighter shade represents the swing phase. **(D)** The Mahalanobis distance between the trajectories of each of the speeds and the trajectory of the fastest speed (6.4 km/h).

2.7. Frequency Analysis

To compare the smoothness of the decoder outputs, the power spectral density (PSD) of the decoded kinematics was computed for each of the four decoders, along with the real kinematics. For each trial, we estimated the PSD using Thomson's multitaper method (*pmtm()* function in Matlab), and the values were averaged across all the trials of all the sessions. This was done for each of the decoded kinematic variables.

2.8. Cross-Validation

We constructed four different decoders (which we will refer to as the Full Population, PSS, PCA and PLDS decoders) and

determined the performance of each through a 10-fold cross-validation paradigm (**Figure 1B**). The PSS, PCA and PLDS decoders employed a two-stage method where binned spike counts were projected onto a low-dimensional space, with dimensionality varying from 1 to 20. The coordinates of the neural data in this space were then utilized as input features for the Wiener filter described below. The Full Population decoder did not employ the dimensionality reduction stage and utilized the z-score normalized spike counts of all recorded neurons as input features (**Figure 1A**). The PLDS inputs assumes spike counts, and utilized the 10 ms bins, while the PCA, PSS, and full population utilized firing rates. The number of spikes is too sparse to estimate the firing rate using only 10 ms counts, so to estimate the firing rates at each 10 ms time step, the number of spikes in the current time bin as well as the previous nine time bins were summed and divided by 100 ms to get the rate in spikes/s. For all decoders, the time steps of the inputs was 10 ms.

Cross validation was performed on each recording session of each monkey. All of the trials of all five tasks were combined and randomized. The trials of all the tasks were then divided into 10 cross-validation blocks. Each block consisted of 5–24 (for decoding kinematics) and 7–27 (for decoding gait) trials which were used as the testing set while the remaining nine blocks were used as the training set. Decoder coefficients were calculated from the training sets using the least-squares regression algorithm between the neural data (Full Population and PSS) or latent states (PCA and PLDS) and the measured output signal (kinematics, or gait phase). The decoder was then used to estimate the output signals of the testing set, and the decoding accuracy was calculated as the coefficient of determination (R^2) between the estimated and the true signal as defined by equation 1. n is the number of points of the kinematic or gait variable in the trial, y_i is the actual variable value at point i, $\hat{y}_i$ is the estimated value from the decoder at point i, and $\bar{y}$ is the average value of the kinematic or gait variable for the trial.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

2.9. Decoder

A linear Wiener filter of order 10 (the filter order that gave the best decoding results under cross validation) was used to decode the kinematics and gait cycle phase from the neural data. The decoded signals included the horizontal and vertical positions of the hip, knee, ankle, metatarsal, and toe tip joint markers, the joint angles, the leg extension, and the percentage of the gait cycle phase. The decoder is described by the equation:

$$y[t] = \sum_{n=0}^{9} A_n * X[t-n] \quad (2)$$

where $y[t]$ is the output signal being decoded at time t, and $X[t-n]$ is a vector of the decoder inputs at lagged times $t-n$, for $n = 0, \ldots, 9$. A_n is the corresponding vector of regression coefficients computed from the training set.

Statistical tests comparing the R^2 values were carried out in Matlab (*signrank()* function).

3. RESULTS

3.1. Structure of Population Dynamics Varies Across Different Locomotion Tasks and Speeds

Using the PLDS model, we extracted cyclic, low-dimensional neural trajectories during locomotion. Visualizing in three dimensions, the trajectories follow similar, saddle-like rotations across different tasks, and across different walking speeds (**Figure 2A**). The Mahalanobis distance between all of the task trajectories against just the treadmill walking trajectories demonstrates that the corridor walking trajectories are the most similar to treadmill walking (mean distance = 5.0392), followed by bipedal, backwards, and ladder walking in various orders depending on the phase of the gait (mean distance = 9.12, 10.66, and 11.97, respectively, **Figure 2B**). However, any interpretation of the neural trajectories for backwards walking should be treated with caution due to the low number of trials available (**Table 1**). When training the PLDS model only on treadmill walking trials, there is a similar rotational structure across all walking speeds, however trajectories appear to separate along the third latent dimension (**Figure 2C**). The change in state-space position along this dimension is reflected in the increase in Mahalanobis distance between trials of different speeds as the difference in speed increases (**Figure 2D**). In general, the PLDS model was able to infer latent state trajectories that are closely phase-locked to the locomotor rhythm, and distinguishes relevant behavioral parameters such as task type and walking speed. We next describe how well kinematic and gait parameters could be decoded from these low-dimensional dynamics.

3.2. PLDS Latent State Trajectories Capture Limb Kinematics and Gait Phase

We used several neural features (full population firing rates, PLDS or PCA latent variables, or a predictive sub-sample (PSS) of the population firing rates) as inputs to a Wiener filter decoder and measured how well each input feature could decode various kinematic variables under cross-validation. To determine the total number of dimensions to use in our decoders, we measured the PLDS decoder performance for all the kinematic variables while varying the input dimensionality from 1 to 20. We found that the performance plateaued at approximately 12 dimensions (**Figure 3B**); we used this number of dimensions for the remainder of the analyses, as well as for calculating the Mahalanobis distances in **Figure 2**.

First, we investigated whether using PLDS as a feature extractor would improve or deteriorate decoding of hind-limb kinematics. An example trace of the decoded leg extension for one trial is shown in **Figure 3A**. For this example, the estimate from the PLDS decoder approximates the true signal more faithfully than the estimate from the Full Population decoder (R^2 = 0.8294 vs. R^2 = 0.5461), as well as compared to the PCA and PSS decoders (R^2 = 0.5355 and R^2 = 0.4980, respectively). We then compared the cross-validation results for the various decoders across all monkeys and recording sessions that have kinematics ($n = 7$) for each of the kinematic variables,

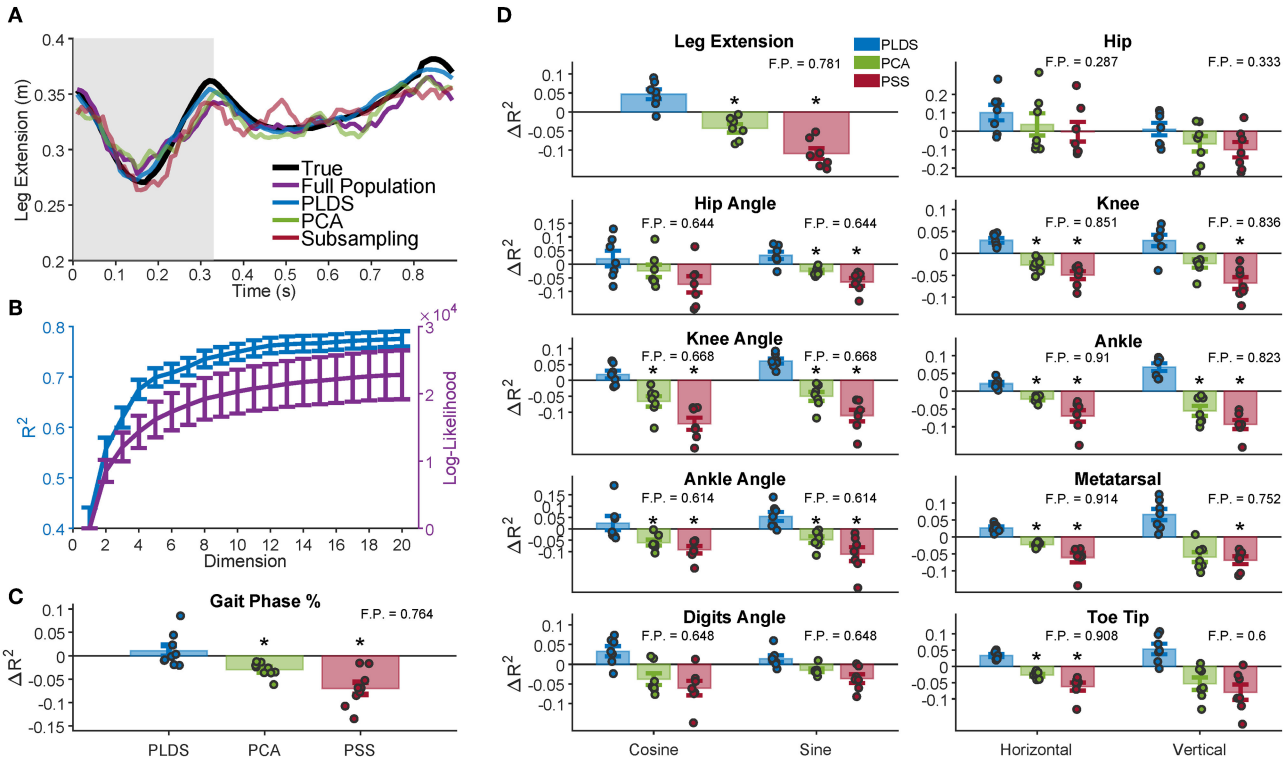


FIGURE 3 | Decoding of leg kinematics and gait phase. **(A)** A representative example of the limb extension kinematic variable during one gait cycle along with the decoded estimation for the four decoders. Gray background represents the swing period of the gait cycle, white background represents the stance period. In this example, the PLDS decoded signal was the most successful at reconstructing the true signal, with an R^2 of 0.8294, followed by decoding without any dimensionality reduction (Full Population: FP) with an R^2 of 0.5461. PCA and PSS decoding performed the worst, with R^2 of 0.5355 and 0.4980, respectively. The R^2 values for all of the animals and sessions are shown in **(D)**. **(B)** Dimensionality analysis for PLDS reveals a plateau in decoding performance (blue) at approximately 12 dimensions for the latent states (plateau defined as when the increase in R^2 <0.001), error bars 1 SEM. Log-likelihood (purple) also plateaus at around 12 dimensions. **(C)** Cross-validated decoding performance of dimensionality reduction techniques for decoding gait phase percentage. Bars represent the increase in R^2 over the Full Population decoder, with the inset number representing the average R^2 of the Full Population decoder. The full distribution of change in R^2 are displayed as dots over each bar. The PLDS decoder had the highest average R^2 for all the variables. Error bars: 1 S.E.M. Black stars denote significant difference from Full Population decoding R^2 (Wilcoxon Sign Rank Test; Benjamini-Hochberg false discovery rate correction for multiple comparisons, with FDR = 10%). **(D)** Same as in **(C)**, except for all analyzed kinematic variables rather than gait phase. Each dot represents a animal/recording session, and the bar plots show the average R^2 with error bars representing 1 S.E.M.

including the horizontal and vertical position of the joints, the joint angles (taking the sine and cosine to ensure valid R^2 values), and leg extension (two-tailed Wilcoxon signed-rank test, Benjamini-Hochberg false discovery rate correction for multiple comparisons, with FDR = 10%). The PLDS decoder performance was not significantly different from the full-population decoder for any of the tested kinematic variables despite having only 12 input features, and indeed for all variables, the average PLDS decoder R^2 was higher than the full population, although none of them reached significance when corrected for multiple testing (**Figure 3D**). However, the PCA decoder performed worse than the full population for many of the kinematic variables, as did the PSS decoder (see **Table S1** for *p*-values), leading us to conclude that PCA and PSS features, at the same number of dimensions, do not preserve the kinematic information as completely as those obtained with the PLDS model. From the sample trial in **Figure 3A**, the PLDS decoded signal appears much smoother than the other three estimations, which may account for the higher R^2. The power spectrum of the decoder outputs verifies that the PLDS decoder and the real kinematics have less power at high frequencies compared to the full population, PCA, and PSS decoders (**Figure S1**). Although the Wiener filter does provide some degree of smoothing by taking into account the previous 10 time bins in the decoding, the PLDS model smooths the intrinsic trajectories based on the activity of the whole ensemble, which may provide a more accurate estimation of the dynamics.

In addition to the above kinematic variables, we also investigated how well gait information, such as the phase of the gait cycle, could be estimated using each decoder. Trials consisted of a single gait cycle, and the phase of each cycle were standardized to a range of 0–100%. The phase was then estimated by each of the decoders, and the change in R^2 from the full population decoder is shown in **Figure 3C** for each of the recording sessions (n = 9). The decoding performance was similar to the decoding of kinematics; there was no significant difference in R^2 between the PLDS decoder and full population

decoder, and the PCA and PSS decoder performed worse than the full population decoder (see **Table S1** for *p*-values).

4. DISCUSSION

While the PLDS and PCA models are both dimensionality reduction techniques, a key difference of PLDS is it explicitly models the temporal dynamics using a linear dynamical system. Other techniques have been developed which also models these dynamics such as jPCA (Churchland et al., 2012), or Hypothesis-guided dimensionality reduction (HDR) (Lara et al., 2018). These other techniques introduces some constraints to the form of the dynamical system in order to test specific hypotheses about the structure of cortical dynamics. For example, jPCA limits the temporal transition matrix to the set of skew-symmetric matrices in order to extract rotational dynamics, while in Lara et al. (2018), HDR was used to divide the projected dimensions into linear dynamical systems states and dimensions orthogonal to those states that are condition-invariant (in order to find similarities between two different brain regions). We did not wish to impose any additional constraints about the form the dynamical systems so we used the general form PLDS model for this analysis.

Both the PCA and the PLDS model extracted neural trajectories that were oscillatory in low-dimensional space. PLDS explicitly models temporal dynamics of the latent states with a state transition matrix, resulting in smoother single-trial cyclical trajectories (**Figure 1B**, **Figure S1**). The structure of the trajectories for treadmill walking at different speeds are in agreement with previous studies, where the neural latent-states separated along one dimension as walking speeds increased while still conserving the rotational structure (Foster et al., 2014). We were also able to obtain qualitatively distinct neural trajectories during different tasks (**Figure 2A**), although again, rotational structure is preserved across all tasks.

The dimensionality of neural data required to represent hind-limb kinematics was approximately 12 dimensions, when empirically determined as the plateau in kinematic decoding performance (**Figure 3B**). Previous studies have estimated the intrinsic dimensionality in forelimb motor cortex during center out tasks to be around 10–20 dimensions (Yu et al., 2009; Sadtler et al., 2014; Vargas-Irwin et al., 2015). Despite the more constrained movements of the hind-limb during locomotion, the dimensionality between leg and arm area of M1 and between reaching and walking are surprisingly similar. Such similarity could suggest comparable levels of cortical involvement during these different behaviors (although the actual structure of the cortical activity could be different). In 1989, Georgopoulos and Grillner proposed the hypothesis that reaching movements in primates may have evolved out of precise gait adjustments during locomotion (Georgopoulos and Grillner, 1989), and others have suggested that similarities in neural and muscle activation onsets between those two types of movements support this hypothesis (Yakovenko and Drew, 2015). The similarities in neural dimensionality between voluntary reaching and walking would also be consistent with this view. However, due to the low sample size and the use of only one dimensionality reduction technique in this study, further experiments exploring additional models may be necessary before any definitive conclusions can be drawn on the dimensionality of leg-M1 population activity during locomotion.

It is well understood that neurons in motor cortex are correlated to muscle activity and kinematics during locomotion in cats (Beloozerova and Sirota, 1998; Drew et al., 2002; Drew and Marigold, 2015), rodents (Song et al., 2009; Rigosa et al., 2015; DiGiovanna et al., 2016; Miri et al., 2017), and non-human primates (Fitzsimmons et al., 2009; Foster et al., 2014; Yin et al., 2014). Here, we showed whether the extracted low dimensional dynamics from PLDS preserve the kinematic information that is present in the neural activity. In terms of the decoding performance, PLDS with only 12 dimensions was able to reconstruct all the kinematic variables in addition to the gait phase just as well as the full population decoder that contains 18–80 dimensions. There is some variance in the distribution of PLDS decoder improvements across animals and sessions (**Figure 3**). This could be due to the different populations of neurons that are recorded from in different subjects. These populations are not homogeneous and may represent the true underlying cortical dynamics by varying degrees. Other dimensionality reduction techniques such as PCA or PSS were unable to achieve the same decoding performance at the same number of dimensions, indicating that an explicit model of temporal dynamics of the low-dimensional states, such as PLDS, is crucial for decoding the kinematics accurately. These results reflect the improved decoding performance using linear dynamical systems models during cursor control from arm area (Yu et al., 2009; Kao et al., 2015) and during reaching and grasping behaviors (Aghagolzadeh and Truccolo, 2016), suggesting that low-dimensional dynamics play an important role in both types of movements. Our findings differ in that we did not see a statistically significant increase in performance when using latent state input features, whereas the arm decoding studies did see an improvement. However, given that the R^2 of every single kinematic variable was on average higher with the PLDS decoder compared to the full population decoder, this may be due to our low sample size and lack of statistical power. Additionally, the power is lowered by the use of parametric tests and the large number of kinematic variables tested.

We should also mention that our decoders utilized firing rate inputs, and thus carries the assumption of rate coding rather than temporal coding. Recent findings have suggested that temporal coding may play a larger role in motor control than previously thought (Srivastava et al., 2017). One future extension of this study would be to include models that utilize precise spike timings. Additionally, the Wiener filter decoder we employed is a linear decoder, and although it was able to reconstruct the kinematics fairly accurately, other non-linear models could be used to further improve decoding performance. Finally, we should mention that this study was limited to higher-level control areas such as the motor cortex, though lower-level structures such as brain-stem or spinal cord have been shown to also exhibit intrinsic population dynamics (Bruno et al., 2017).

In conclusion, our study investigates whether unsupervised dimensionality reduction can infer latent neural states reflecting ensemble dynamics, while preserving information about the kinematics and gait phase of the hind-limb during various locomotion tasks. We show that dynamical systems models, which have been shown to decode forelimb reaching kinematics, were able to extract robust, stereotyped low-dimensional state-space trajectories, and that these trajectories capture hind-limb movements during directed locomotion (e.g., ladder walking), as well as autonomous locomotion (e.g., basic treadmill and corridor walking). As far as we are aware, this is the first demonstration of explicit state-space models of neural dynamics robustly decoding kinematic and gait information during primate locomotion. These results also points to the potential of using PLDS in hind-limb BMIs, although direct testing in a closed-loop system would be required before any determination of the usefulness of PLDS as a feature extraction step can be made. Recently, newer techniques have been developed to extract neural dynamics using recurrent neural networks (Pandarinath et al., 2018) which enable extraction of non-linear dynamics and have been employed in arm reaching tasks. Such models may extract the underlying neural dynamics more accurately and may outperform PLDS in terms of decoding of kinematics. One potential extension to this study in the future would be to apply these non-linear models to hind-limb locomotion behaviors as well.

DATA AVAILABILITY STATEMENT

Data are available by directly contacting GC at gregoire.courtine@epfl.ch.

ETHICS STATEMENT

The animal study was reviewed and approved by Institutional Animal Care and Use Committee of Bordeaux (CE50), license number 50120102-A.

AUTHOR CONTRIBUTIONS

DB, EB, and GC performed data collection experiments. DX and MA performed decoding analysis. WT and DB conceived of and oversaw the project. EB and GC oversaw the original data collection. DX wrote the manuscript with MA. All authors contributed to the editing and finalization of the manuscript.

FUNDING

This work was sponsored by the Defense Advanced Research Projects Agency (DARPA) BTO under the auspices of Dr. Doug Weber and Alfred Emondi through the [Space and Naval Warfare Systems Center, Pacific OR DARPA Contracts Management Office] Grant/Contract No. D15AP00112; the DARPA program Neural Engineering Systems Design (NESD) (N666001-17-C-4013); as well as the International Research in Paraplegia Foundation (IRP, P152). This work was also supported by Marie Curie Fellowship, 331602 (e-WALK) to DB under the direct mentorship of GC. We would also like to acknowledge the Pablo J. Salame '88 Goldman Sachs endowed Associate Professorship of Computational Neuroscience at Brown University (WT), and the Howard Reisman '76 Family Graduate Fellowship Fund and the Charles A. Dana Graduate Fellowship Fund (DX). This work was supported in part by Merit Review Award # I01RX002835 from the United States (U.S.) Department of Veterans Affairs, Rehabilitation Research and Development Service. The contents of this manuscript do not represent the views of VA or the United States Government. Department of Veterans Affairs did not provide any direct support of the NHP work conducted for this project. These funding sources had no involvement in the design of this study, collection or analysis of data, nor in the authorship of this manuscript.

ACKNOWLEDGMENTS

The authors thank the Ecole Polytechnique Federale de Lausanne (EPFL) Center for Neuroprosthetics and the laboratory of Prof. Gregoire Courtine for contributing data used in the analysis presented. The authors would also like to thank Jean Laurens for his help with data collection and organization.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01046/full#supplementary-material

REFERENCES

Afshar, A., Santhanam, G., Yu, B. M., Ryu, S. I., Sahani, M., and Shenoy, K. V. (2011). Single-trial neural correlates of arm movement preparation. *Neuron* 71, 555–564. doi: 10.1016/j.neuron.2011.05.047

Aghagolzadeh, M., and Truccolo, W. (2014). "Latent state-space models for neural decoding," in *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* (Chicago, IL), 3033–3036.

Aghagolzadeh, M., and Truccolo, W. (2016). Inference and decoding of motor cortex low-dimensional dynamics via latent state-space models. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 272–282. doi: 10.1109/TNSRE.2015.2470527

Beloozerova, I. N., and Sirota, M. G. (1988). "Role of motor cortex in control of locomotion," in *Stance and Motion*, eds V. S. Gurfinkel, M. E. Ioffe, J. Massion, and J. P. Roll (Boston, MA: Springer), 163–176.

Beloozerova, I. N., and Sirota, M. G. (1998). Cortically controlled gait adjustments in the cat. *Ann. N. Y. Acad. Sci.* 860, 550–553. doi: 10.1111/j.1749-6632.1998.tb09101.x

Bruno, A. M., Frost, W. N., and Humphries, M. D. (2017). A spiral attractor network drives rhythmic locomotion. *eLife* 6:e27342. doi: 10.7554/eLife.27342

Capogrosso, M., Milekovic, T., Borton, D., Wagner, F., Moraud, E. M., Mignardot, J. B., et al. (2016). A brain-spine interface alleviating gait deficits after spinal cord injury in primates. *Nature* 539, 284–288. doi: 10.1038/nature20118

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., et al. (2012). Neural population dynamics during reaching. *Nature* 487, 51–56. doi: 10.1038/nature11129

Courtine, G. (2005). Performance of locomotion and foot grasping following a unilateral thoracic corticospinal tract lesion in monkeys (*Macaca mulatta*). *Brain* 128, 2338–2358. doi: 10.1093/brain/awh604

Cunningham, J. P., and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* 17, 1500–1509. doi: 10.1038/nn.3776

DiGiovanna, J., Dominici, N., Friedli, L., Rigosa, J., Duis, S., Kreider, J., et al. (2016). Engagement of the rat hindlimb motor cortex across natural locomotor behaviors. *J. Neurosci.* 36, 10440–10455. doi: 10.1523/JNEUROSCI.4343-15.2016

Drew, T. (1988). Motor cortical cell discharge during voluntary gait modification. *Brain Res.* 457, 181–187. doi: 10.1016/0006-8993(88)90073-X

Drew, T., Andujar, J. E., Lajoie, K., and Yakovenko, S. (2008). Cortical mechanisms involved in visuomotor coordination during precision walking. *Brain Res. Rev.* 57, 199–211. doi: 10.1016/j.brainresrev.2007.07.017

Drew, T., Jiang, W., Kably, B., and Lavoie, S. (1996). Role of the motor cortex in the control of visually triggered gait modifications. *Can. J. Physiol. Pharmacol.* 74, 426–442. doi: 10.1139/y96-043

Drew, T., Jiang, W., and Widajewicz, W. (2002). Contributions of the motor cortex to the control of the hindlimbs during locomotion in the cat. *Brain Res.* 40, 178–191. doi: 10.1016/S0165-0173(02)00200-X

Drew, T., and Marigold, D. S. (2015). Taking the next step: cortical contributions to the control of locomotion. *Curr. Opin. Neurobiol.* 33, 25–33. doi: 10.1016/j.conb.2015.01.011

Fitzsimmons, N. A., Lebedev, M. A., Peikon, I. D., and Nicolelis, M. A. (2009). Extracting kinematic parameters for monkey bipedal walking from cortical neuronal ensemble activity. *Front. Integr. Neurosci.* 3:3. doi: 10.3389/neuro.07.003.2009

Foster, J. D., Nuyujukian, P., Freifeld, O., Gao, H., Walker, R., I Ryu, S., et al. (2014). A freely-moving monkey treadmill model. *J. Neural Eng.* 11:046020. doi: 10.1088/1741-2560/11/4/046020

Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. (2017). Neural manifolds for the control of movement. *Neuron* 94, 978–984. doi: 10.1016/j.neuron.2017.05.025

Gallego, J. A., Perich, M. G., Naufel, S. N., Ethier, C., Solla, S. A., and Miller, L. E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* 9:4233. doi: 10.1038/s41467-018-06560-z

Georgopoulos, A. P., and Grillner, S. (1989). Visuomotor coordination in reaching and locomotion. *Science (New York, N.Y.)* 245, 1209–1210. doi: 10.1126/science.2675307

He, S. Q., Dum, R. P., and Strick, P. L. (1993). Topographic organization of corticospinal projections from the frontal lobe: motor areas on the lateral surface of the hemisphere. *J. Neurosci.* 13, 952–980. doi: 10.1523/JNEUROSCI.13-03-00952.1993

Kao, J. C., Nuyujukian, P., Ryu, S. I., Churchland, M. M., Cunningham, J. P., and Shenoy, K. V. (2015). Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nat. Commun.* 6:7759. doi: 10.1038/ncomms8759

Lara, A. H., Cunningham, J. P., and Churchland, M. M. (2018). Different population dynamics in the supplementary motor area and motor cortex during reaching. *Nat. Commun.* 9:2754. doi: 10.1038/s41467-018-05146-z

Laurens, J., Meng, H., and Angelaki, D. E. (2013). Computation of linear acceleration through an internal model in the macaque cerebellum. *Nat. Neurosci.* 16, 1701–1708. doi: 10.1038/nn.3530

Macke, J. H., Büsing, L., Cunningham, J. P., Yu, B. M., Shenoy, K., and Sahani, M. (2011). "Empirical models of spiking in neural populations," in *Advances in Neural Information Processing Systems 24: 25th Conference on Neural Information Processing Systems* (Granada), 1350–1358.

Miri, A., Warriner, C. L., Seely, J. S., Elsayed, G. F., Cunningham, J. P., Churchland, M. M., et al. (2017). Behaviorally selective engagement of short-latency effector pathways by motor cortex. *Neuron* 95, 683–696. doi: 10.1016/j.neuron.2017.06.042

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* 15, 805–815. doi: 10.1038/s41592-018-0109-9

Rigosa, J., Panarese, A., Dominici, N., Friedli, L., van den Brand, R., Carpaneto, J., et al. (2015). Decoding bipedal locomotion from the rat sensorimotor cortex. *J. Neural Eng.* 12:056014. doi: 10.1088/1741-2560/12/5/056014

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., et al. (2014). Neural constraints on learning. *Nature* 512, 423–426. doi: 10.1038/nature13665

Song, W., Ramakrishnan, A., Udoekwere, U. I., and Giszter, S. F. (2009). Multiple types of movement-related information encoded in hindlimb/trunk cortex in rats and potentially available for brain-machine interface controls. *IEEE Trans. Biomed. Eng.* 56(11 Pt 2), 2712–2716. doi: 10.1109/TBME.2009.2026284

Srivastava, K. H., Holmes, C. M., Vellema, M., Pack, A. R., Elemans, C. P. H., Nemenman, I., et al. (2017). Motor control by precisely timed spike patterns. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1171–1176. doi: 10.1073/pnas.1611734114

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* 93, 1074–1089. doi: 10.1152/jn.00697.2004

Truccolo, W., Hochberg, L. R., and Donoghue, J. P. (2010). Collective dynamics in human and monkey sensorimotor cortex: predicting single neuron spikes. *Nat. Neurosci.* 13, 105–111. doi: 10.1038/nn.2455

Vargas-Irwin, C. E., Brandman, D. M., Zimmermann, J. B., Donoghue, J. P., and Black, M. J. (2015). Spike train SIMilarity Space (SSIMS): a framework for single neuron and ensemble data analysis. *Neural Comput.* 27, 1–31. doi: 10.1162/NECO_a_00684

Yakovenko, S., and Drew, T. (2015). Similar motor cortical control mechanisms for precise limb control during reaching and locomotion. *J. Neurosci.* 35, 14476–14490. doi: 10.1523/JNEUROSCI.1908-15.2015

Yin, M., Borton, D. A., Komar, J., Agha, N., Lu, Y., Li, H., et al. (2014). Wireless neurosensor for full-spectrum electrophysiology recordings during free behavior. *Neuron* 84, 1170–1182. doi: 10.1016/j.neuron.2014.11.010

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* 102, 614–635. doi: 10.1152/jn.90941.2008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Xing, Aghagolzadeh, Truccolo, Bezard, Courtine and Borton. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

CORRECTION
published: 29 October 2020
doi: 10.3389/fnins.2020.604517

Corrigendum: Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates

David Xing[1†], *Mehdi Aghagolzadeh*[2†], *Wilson Truccolo*[2,3,4], *Erwan Bezard*[5,6], *Gregoire Courtine*[7,8,9,10‡] *and David Borton*[1,3,4*‡]

[1] School of Engineering, Brown University, Providence, RI, United States, [2] Department of Neuroscience, Brown University, Providence, RI, United States, [3] Carney Institute for Brain Science, Brown University, Providence, RI, United States, [4] U.S. Department of Veterans Affairs, Center for Neurorestoration and Neurotechnology, Providence, RI, United States, [5] CNRS, Institut des Maladies Neurodégénératives, UMR 5293, Bordeaux, France, [6] Université de Bordeaux, Institut des Maladies Neurodégénératives, UMR 5293, Bordeaux, France, [7] Center for Neuroprosthetics and Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Geneva, Switzerland, [8] Department of Clinical Neuroscience, Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Lausanne, Switzerland, [9] Defitech Center for Interventional Neurotherapies (NeuroRestore), CHUV/UNIL/EPFL, Lausanne, Switzerland, [10] Department of Neurosurgery, CHUV, Lausanne, Switzerland

OPEN ACCESS

Approved by:
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

***Correspondence:**
David Borton
david_borton@brown.edu

[†]These authors have contributed equally to this work

[‡]These authors share senior authorship

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Neuroscience

Received: 10 September 2020
Accepted: 10 September 2020
Published: 29 October 2020

Citation:
Xing D, Aghagolzadeh M, Truccolo W, Bezard E, Courtine G and Borton D (2020) Corrigendum: Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates. Front. Neurosci. 14:604517. doi: 10.3389/fnins.2020.604517

Keywords: low dimensional dynamics, locomotion, non-human primate (NHP), poisson linear dynamical system, primary motor cortex (M1)

A Corrigendum on

Low-Dimensional Motor Cortex Dynamics Preserve Kinematics Information During Unconstrained Locomotion in Nonhuman Primates

by Xing, D., Aghagolzadeh, M., Truccolo, W., Bezard, E., Courtine, G., and Borton, D. (2019). Front. Neurosci. 13:1046. doi: 10.3389/fnins.2019.01046

In the original article, we neglected to include the funder Marie Curie Fellowship, 331602 (e-WALK) to DB under the direct mentorship of GC.

The Funding section has been updated to read as follows:

This work was sponsored by the Defense Advanced Research Projects Agency (DARPA) BTO under the auspices of Dr. Doug Weber and Alfred Emondi through the [Space and Naval Warfare Systems Center, Pacific OR DARPA Contracts Management Office] Grant/Contract No. D15AP00112; the DARPA program Neural Engineering Systems Design (NESD) (N666001-17-C-4013); as well as the International Research in Paraplegia Foundation (IRP, P152). This work was also supported by Marie Curie Fellowship, 331602 (e-WALK) to DB under the direct mentorship of GC. We would also like to acknowledge the Pablo J. Salame '88 Goldman Sachs endowed Associate Professorship of Computational Neuroscience at Brown University (WT), and the Howard Reisman '76 Family Graduate Fellowship Fund and the Charles A. Dana Graduate Fellowship Fund (DX). This work was supported in part by Merit Review Award # I01RX002835 from the United States (U.S.) Department of Veterans Affairs, Rehabilitation Research and Development Service. The contents of this manuscript do not represent the views of VA or the United States Government. Department of Veterans Affairs did not provide any direct support of the NHP work conducted for this project. These funding sources had no involvement in the design of this study, collection or analysis of data, nor in the authorship of this manuscript.

Additionally, in the original article, the Data Availability Statement referenced the incorrect person/institution to contact for data access. A correction has been made to the Data Availability section to now read:

Data are available by directly contacting GC at gregoire.courtine@epfl.ch.

Finally, in the original article, the authorship did not reflect the contributions of members who supported the study. The authorship and the authorship contributions have been changed to now read:

David Xing[1†], Mehdi Aghagolzadeh[2†], Wilson Truccolo[2,3,4], Erwan Bezard[5,6], Gregoire Courtine[7,8,9,10‡] and David Borton[1,3,4*‡]

DB, EB, and GC performed data collection experiments. DX and MA performed decoding analysis. WT and DB conceived of and oversaw the project. EB and GC oversaw the original data collection. DX wrote the manuscript with MA. All authors contributed to the editing and finalization of the manuscript.

The authors apologize for these errors and state that they do not change the scientific conclusions of the article in any way. The original article has been updated.

Copyright © 2020 Xing, Aghagolzadeh, Truccolo, Bezard, Courtine and Borton. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 07 November 2019
doi: 10.3389/fnins.2019.01203

Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia

Raymond Salvador[1,2], Erick Canales-Rodríguez[1,2], Amalia Guerrero-Pedraza[3], Salvador Sarró[1,2], Diana Tordesillas-Gutiérrez[2,4], Teresa Maristany[5], Benedicto Crespo-Facorro[2,4], Peter McKenna[1,2] and Edith Pomarol-Clotet[1,2]*

[1] FIDMAG Hermanas Hospitalarias Research Foundation, Barcelona, Spain, [2] Centro de Investigación Biomédica en Red de Salud Mental, Madrid, Spain, [3] Hospital Benito Menni – Complex Assistencial en Salut Mental, Sant Boi de Llobregat, Spain, [4] Hospital Universitario Marqués de Valdecilla, Universidad de Cantabria, Santander, Spain, [5] Hospital Sant Joan de Déu, Barcelona, Spain

Magnetic resonance imaging (MRI) has been proposed as a source of information for automatic prediction of individual diagnosis in schizophrenia. Optimal integration of data from different MRI modalities is an active area of research aimed at increasing diagnostic accuracy. Based on a sample of 96 patients with schizophrenia and a matched sample of 115 healthy controls that had undergone a single multimodal MRI session, we generated individual brain maps of gray matter vbm, 1back, and 2back levels of activation (*n*back fMRI), maps of amplitude of low-frequency fluctuations (resting-state fMRI), and maps of weighted global brain connectivity (resting-state fMRI). Four unimodal classifiers (Ridge, Lasso, Random Forests, and Gradient boosting) were applied to these maps to evaluate their classification accuracies. Based on the assignments made by the algorithms on test individuals, we quantified the amount of predictive information shared between maps (what we call redundancy analysis). Finally, we explored the added accuracy provided by a set of multimodal strategies that included post-classification integration based on probabilities, two-step sequential integration, and voxel-level multimodal integration through one-dimensional-convolutional neural networks (1D-CNNs). All four unimodal classifiers showed the highest test accuracies with the 2back maps (80% on average) achieving a maximum of 84% with the Lasso. Redundancy levels between brain maps were generally low (overall mean redundancy score of 0.14 in a 0–1 range), indicating that each brain map contained differential predictive information. The highest multimodal accuracy was delivered by the two-step Ridge classifier (87%) followed by the Ridge maximum and mean probability classifiers (both with 85% accuracy) and by the 1D-CNN, which achieved the same accuracy as the best unimodal classifier (84%). From these results, we conclude that from all MRI modalities evaluated task-based fMRI may be the best unimodal diagnostic option in schizophrenia. Low redundancy values point to ample potential for accuracy improvements through multimodal integration, with the two-step Ridge emerging as a suitable strategy.

Keywords: multimodal integration, schizophrenia, machine learning, computer-aided diagnosis, convolutional neural network, lasso, ridge

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Domenico Zaca,
Siemens, Italy
Maria Eugenia Caligiuri,
University of Magna Graecia, Italy

****Correspondence:***
Raymond Salvador
rsalvador@fidmag.com;
rai.salvador@gmail.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *29 May 2019*
Accepted: *23 October 2019*
Published: *07 November 2019*

Citation:
Salvador R, Canales-Rodríguez E, Guerrero-Pedraza A, Sarró S, Tordesillas-Gutiérrez D, Maristany T, Crespo-Facorro B, McKenna P and Pomarol-Clotet E (2019) Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia. Front. Neurosci. 13:1203. doi: 10.3389/fnins.2019.01203

INTRODUCTION

In recent years, there has been growing interest in employing brain magnetic resonance imaging (MRI) datasets for medical diagnosis (Wang and Summers, 2012). Specifically, in the field of schizophrenia, a considerable number of studies have been carried out to evaluate the predictive power of machine learning algorithms based on MRI data (Wolfers et al., 2015; Arbabshirani et al., 2017). To improve the accuracy levels provided by unimodal data sources, some authors have explored ways to combine the information contained in images generated by different MRI modalities. These include methods with different levels of data integration and of a very different nature, ranging from simple post-classification majority-vote strategies to multimodal fusion techniques (Calhoun and Adali, 2009; Sui et al., 2013), including also multiple kernel learning (Peruzzo et al., 2015; Zu et al., 2016), multimodal Gaussian process classifiers (Young et al., 2013), and deep learning (Suk et al., 2014; Shi et al., 2018) among other techniques. Accordingly, multimodal MRI integration for clinical diagnosis is an open and dynamic area of research, but one that still requires investigation (Arbabshirani et al., 2017; Tulay et al., 2019).

Here, relying on two matched samples of patients with schizophrenia (N = 96) and healthy controls (N = 115) for which structural T1, task-based (*n*back task), and resting-state fMRI had been acquired in a single MRI session, we pursue three objectives: (i) to evaluate the differential discriminative power of brain maps derived from the different modalities; (ii) to quantify the degree to which the different types of images have similar or distinct predictive patterns; and (iii) to explore the added accuracy provided by a set of multimodal strategies based on different levels of data integration, including novel approaches such as a two-step data integration scheme and a one-dimensional-convolutional neural network (1D-CNN).

MATERIALS AND METHODS

Sample, Image Acquisition, and Preprocessing

One hundred fifteen healthy controls and 96 patients with a diagnosis of schizophrenia according to DSM-IV criteria underwent a single MRI session where images were acquired with three different modalities. Both groups were matched for age (mean = 36.5, SD = 10.6, and range = 18–63 in controls; mean = 36.3, SD = 10.9, and range = 16–65 in patients), gender (67% of males in both groups), and premorbid IQ as estimated using the Word Accentuation Test (Test de Acentuación de Palabras, TAP) (Del Ser et al., 1997) (mean TAP controls 23.17, mean TAP patients 22.37, t = 1.2373, p = 0.2176). MRI acquisitions included a T1 structural image, a resting-state fMRI sequence, and an fMRI acquisition obtained during performance of the *n*back task (a working memory task). Acquisition parameters as well as a detailed description of the preprocessing steps applied to the images can be found in previous reports: T1 (Salvador et al., 2017b), resting-state fMRI (Salvador et al., 2017a), and the *n*-back task (Fuentes-Claramonte et al., 2019). Images from all modalities were coregistered to the same standard MNI152 T1 2 mm template. For the functional images, this involved an initial linear registration to the individual T1 image followed by a non-linear transformation to the standard MNI template. A generic mask only containing gray matter voxels was generated by applying a threshold on the gray matter probabilistic MNI template available from the FieldMap SPM toolbox[1]. Finally, only voxels that contained data in all modalities were included in the analyses. All participants gave written informed consent prior to participation. All the study procedures had been previously approved by the local research ethical committee and adhered to the Declaration of Helsinki.

Brain Maps

From preprocessed images of the three different modalities, five different brain maps were generated. Individual gray matter voxel-based morphometry (GMVBM) maps were derived from the T1 images. GMVBM maps were selected from among other T1-related data formats because they had delivered optimal accuracy levels in a previous study (Salvador et al., 2017b).

We used images containing the individual regression coefficients for the two main contrasts of the *n*-bask task: 1back-vs.-baseline and 2back-vs.-baseline (Gevins and Cutillo, 1993; Fuentes-Claramonte et al., 2019). Finally, from the resting-state fMRI sequences, we derived two brain maps containing complementary information. On the one hand, we calculated maps of the amplitude of low-frequency fluctuations (ALFF) (Zang et al., 2007), and on the other hand, we generated weighted global brain connectivity (GBC) maps, which are functional connectivity maps based on averaging correlations between each voxel and all the remaining gray matter voxels (Cole et al., 2010; Salvador et al., 2016). Once the five different maps had been generated (**Figure 1**, top), their resolution was downsampled to 4 mm to reduce computational and memory-storage costs (leading to each final map containing 19,660 voxels).

Unimodal Machine Learning

Initially, we applied four different machine learning algorithms to evaluate the relative accuracy provided by the five brain maps. These algorithms included two linear additive classifiers: the Ridge and the Lasso logistic classifiers (Hastie et al., 2009). While the Ridge, with its L2 regularization, provides models in which all voxels had some weight, the Lasso (L1 regularization) provides sparse solutions with only few voxels contributing to the predictive model. Specifically, the functions contained in the glmnet package were applied through the R api (Friedman et al., 2010).

To lessen the linearity and additivity restrictions, two tree-based methods were further considered: a Random Forest and a Gradient Boosting algorithm (Hastie et al., 2009). For Random Forest, we considered 1000 trees and (number of voxels)$^{1/2}$ variables per split, and for Gradient Boosting we carried out an internal cross-validation for the selection of optimal regularization and depth values. R libraries randomForest and gbm were used for the calculations.

[1]https://www.fil.ion.ucl.ac.uk/spm/toolbox/fieldmap

A 10-fold cross-validation was used to obtain unbiased accuracy estimates of the different algorithms applied to the five brain maps by quantifying success rates in the test individuals. In each fold, prior to fitting the machine learning algorithms, linear models accounting for age and gender effects were fitted to the training sample. Later, coefficients of these models were applied to each corresponding test sample.

Redundancy Assessment

A simple index to quantify the degree to which diagnostic predictive traits in a type of brain map are also present in another map (what we call here redundancy) can be easily derived from the success rates achieved in test samples. If Prob(M1) and Prob(M2) are the probabilities of successfully classifying a specific individual using brain maps of type 1 and type 2, we can simply quantify this redundancy with Prob(M2| M1) (i.e., the conditional probability of predicting successfully with map 2 provided that map 1 predicted the right class).

Conditional probabilities will be constrained by two limiting scenarios:

(1) Total redundancy: All predictive features of map 1 are contained in map 2. Then

$$\text{Prob}(M2|M1) = 1 \quad (1)$$

(i.e., if map 1 classifies correctly an individual, then map 2 will also do it).

(2) Complete independence: Maps do not share any predictive traits. Then

$$\text{Prob}(M2|M1) = \text{Prob}(M2) \quad (2)$$

(i.e., the fact that map1 classifies correctly does not increase the probability of success with map2).

Since conditional probabilities will be in the [Prob(M2), 1] interval, a redundancy score (RSC) in the [0,1] interval will be given by

$$\text{RSC}(M2|M1) = (\text{Prob}(M2|M1) - \text{Prob}(M2))/(1 - \text{Prob}(M2)) \quad (3)$$

Note that, obviously, RSC(M2| M1) will usually differ from RSC(M1| M2). RSCs will be useful for evaluating the potential increase in predictive power provided by merging both brain maps in a single classification (i.e., while low RSC values will indicate potential benefits from merging, RSC values close to 1 will not, as the maps will contain very similar predictive information). In our study, estimates of RSC for each pair of maps were obtained from the relative frequencies of individuals being correctly classified with each, or both, brain maps.

Multimodal Integration

To take full advantage of the predictive power of the five brain maps, we followed three strategies with increasing levels of multimodal integration (**Figure 1**).

(1) Post-classification integration based on probabilities: After building unimodal classification algorithms based on the five brain maps and applying them to a test individual, this approach considered simple functions of the five outcome probabilities (delivered by the five brain maps) such as the mean probability or the maximum probability, finally classifying the individual according to this latter value. Alternatively, the log-odds of the same outcome probabilities but calculated for individuals in the training sets were considered as independent variables in a logistic model with diagnosis as the dependent variable. The fitted logistic model was then used to predict diagnosis for test individuals (using the log-odds of their outcome probabilities from unimodal algorithms as predictors).

(2) Two-step sequential integration: After applying unimodal machine learning on the training sets, the subset of "most relevant" voxels from each brain map was merged in a single dataset that, in turn, was used to train a final model. In Lasso and Ridge, relevance was defined by the magnitude of regression coefficients in the logistic model. Since Lasso is a sparse method,

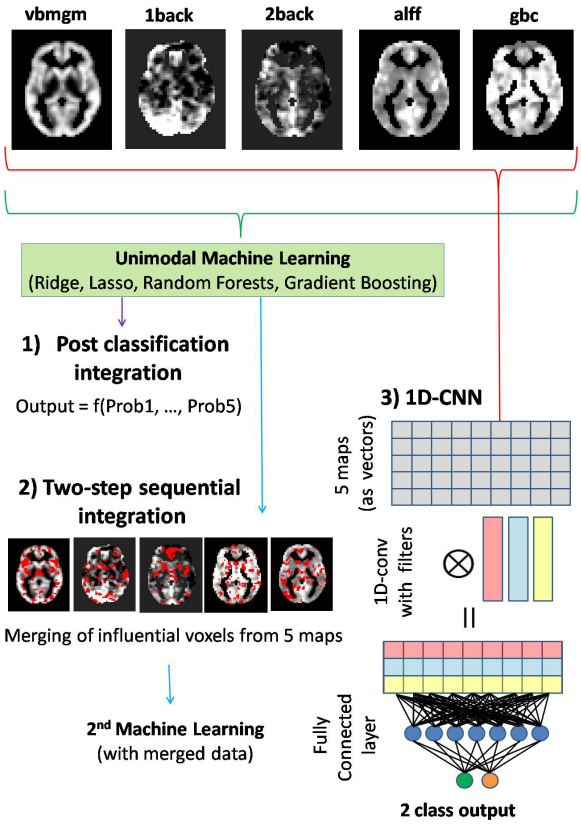


FIGURE 1 | Diagram showing the three multimodal integrative strategies starting from the five different brain maps pictured on top of the figure. (1) Post-classification integration based on functions of output probabilities delivered by unimodal classifiers. (2) Two-step sequential integration based on an initial unimodal classification used to select the most informative voxels from each map, and followed by a second classification only considering the values of these voxels as inputs of the algorithm. (3) Voxel-level multimodal integration with 1D-convolutional neural networks (1D-CNNs). In this last approach, one-dimensional convolutions are applied across brain maps generating new maps that are combinations of the original ones. This is followed by a second fully connected layer linked to the two-node output layer (patient–control labels).

we selected all voxels with non-null coefficients, while for Ridge 20% of the voxels with the largest coefficients (in absolute value) were retained. For both tree-based methods, importance scores were used to select voxels. As Gradient Boosting showed a high level of sparsity, all voxels with non-zero importance scores were selected. For Random Forests the same 20% criterion as in Ridge was used.

(3) Voxel-level multimodal integration with 1D-CNNs: Unlike previous approaches, applying 1D convolutions at the voxel level across brain maps may permit finding optimal, within voxel combinations of modalities. In our study a two-layer neural network with one 1D convolutional layer including 10 filters and a standard fully connected layer with 50 hidden units was considered (**Figure 1**). Other parameters used were as follows: number of iterations = 50, batch size = 4, learning rate = 0.001, momentum = 0.9, wd = 0.0005, activation function = Relu. Network training was carried out with the mxnet library[2].

The same 10-fold cross-validation scheme previously used for unimodal machine learning was applied in all analyses based on

[2]https://mxnet.apache.org/

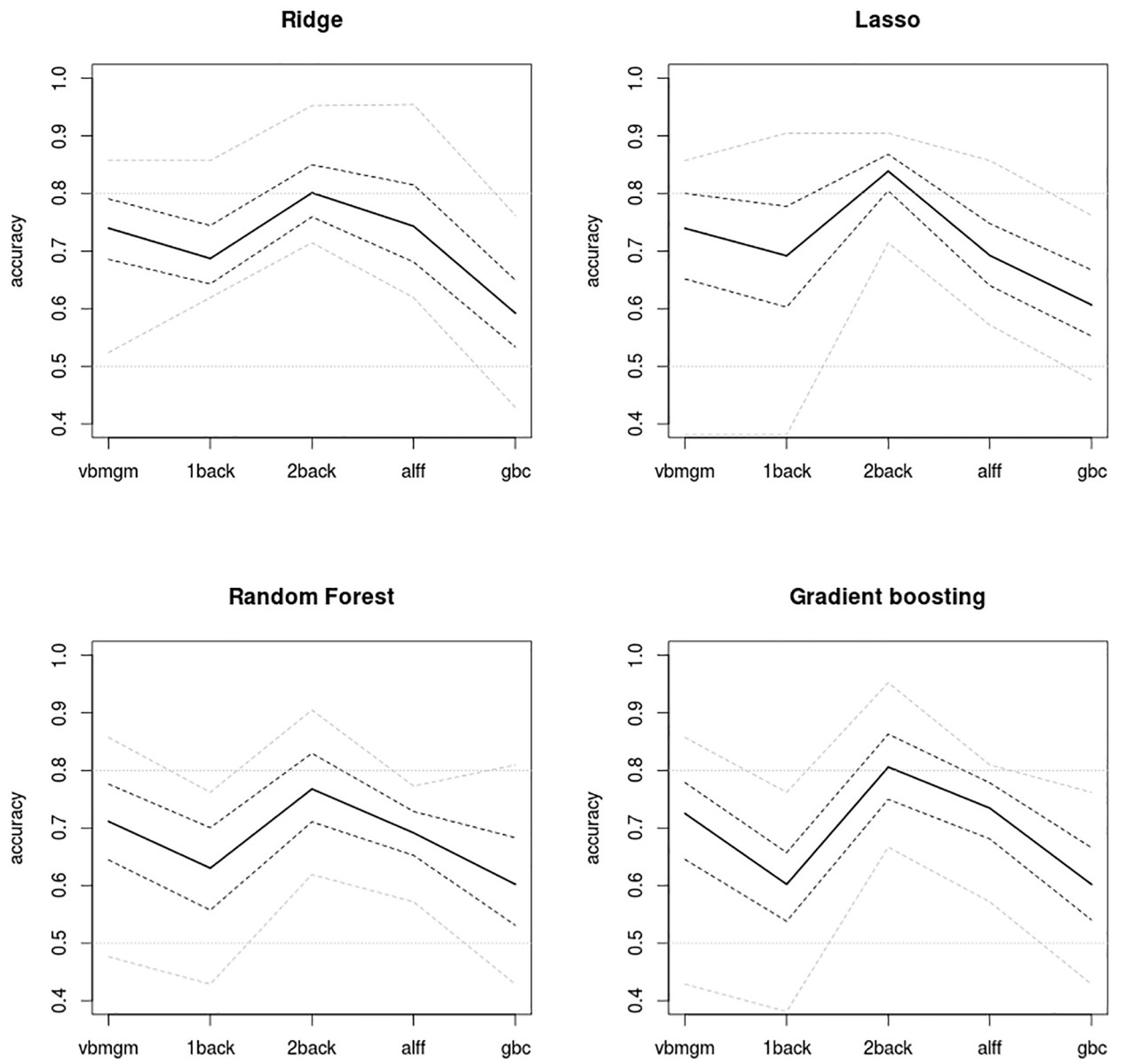


FIGURE 2 | Plots of accuracy levels achieved by the four unimodal algorithms applied to the five brain maps. Mean accuracies extracted from test samples are shown by a continuous line. 2back maps show the strongest predictive power using all algorithms and GBC maps deliver the weakest. Gray dashed lines, 95% bootstrap intervals of mean accuracies and gray dotted lines, maximum and minimum accuracies delivered by the 10-fold scheme. 0.5 accuracy indicates chance accuracy (no real predictive power).

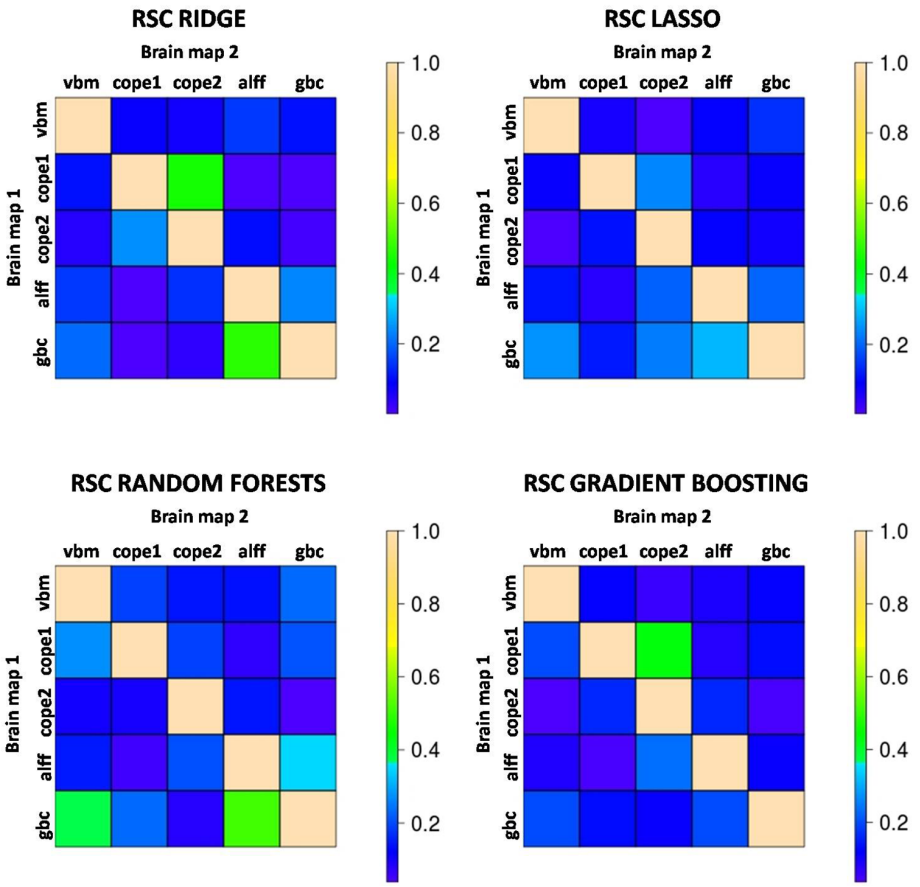


FIGURE 3 | Color-coded values for the redundancy scores (RSC) that quantify the degree to which predictive features of Brain map 2 are also present in Brain map 1. An RSC value close to 1 indicates that map 2 brings almost no predictive information apart from that contained in map 1 (high redundancy) while a value of RSC close to 0 indicates that map 1 contains hardly any of the predictive patterns present in map 2 (both maps convey independent information). The highest RSCs tend to occur between maps derived from the same image modality, although these hardly ever reach values >0.50. Most of the RSCs are well below this number, indicating very low levels of redundancy and potential increases in accuracy through multimodal integration. cope1: 1back maps; cope2: 2back maps.

multimodal integration in order to secure unbiased accuracies from the test data.

RESULTS

Unimodal Machine Learning

The main results from applying the different machine learning algorithms to the five brain maps are shown in **Figure 2**. In all four unimodal algorithms 2back maps showed the highest test accuracies (80% on average), achieving a maximum of 84% with Lasso. In contrast, GBC maps tended to deliver the lowest values (with an overall average of 60%). GMVBM (73%), 1back (65%), and ALFF (71%) maps had intermediate accuracy values. All accuracies are well >50%, indicating above-chance predictive power.

Redundancy Assessment

As shown in **Figure 3**, in general redundancy levels between brain maps were quite low, thus indicating that each brain map contained differential predictive information. As expected, the highest redundancies were observed between pairs of maps that came from the same MRI modality (i.e., 1back-2back maps and ALFF-GBC maps) but even these rarely reached RSCs of 0.5 [this only occurred for RSC(GBC|ALFF) = 0.51 with Random Forests]. In contrast, much lower redundancies were observed between maps derived from different MRI modalities, which in some cases showed almost complete independence in their predictive information, with RSC values as small as 0.002. Overall mean RSC was 0.14. In all, redundancy assessment indicated ample potential for accuracy improvements through multimodal integration. See the tables in the **Supplementary Material 1** for numerical values of the RSC and the conditional probabilities.

Multimodal Integration

Accuracy levels achieved by the first two multimodal integration strategies (i.e., probability based and two-step sequential integration) are shown in **Figure 4** together with the accuracies delivered by their respective best unimodal models and by the 1D-CNN classifier. Apart from the Ridge algorithm where

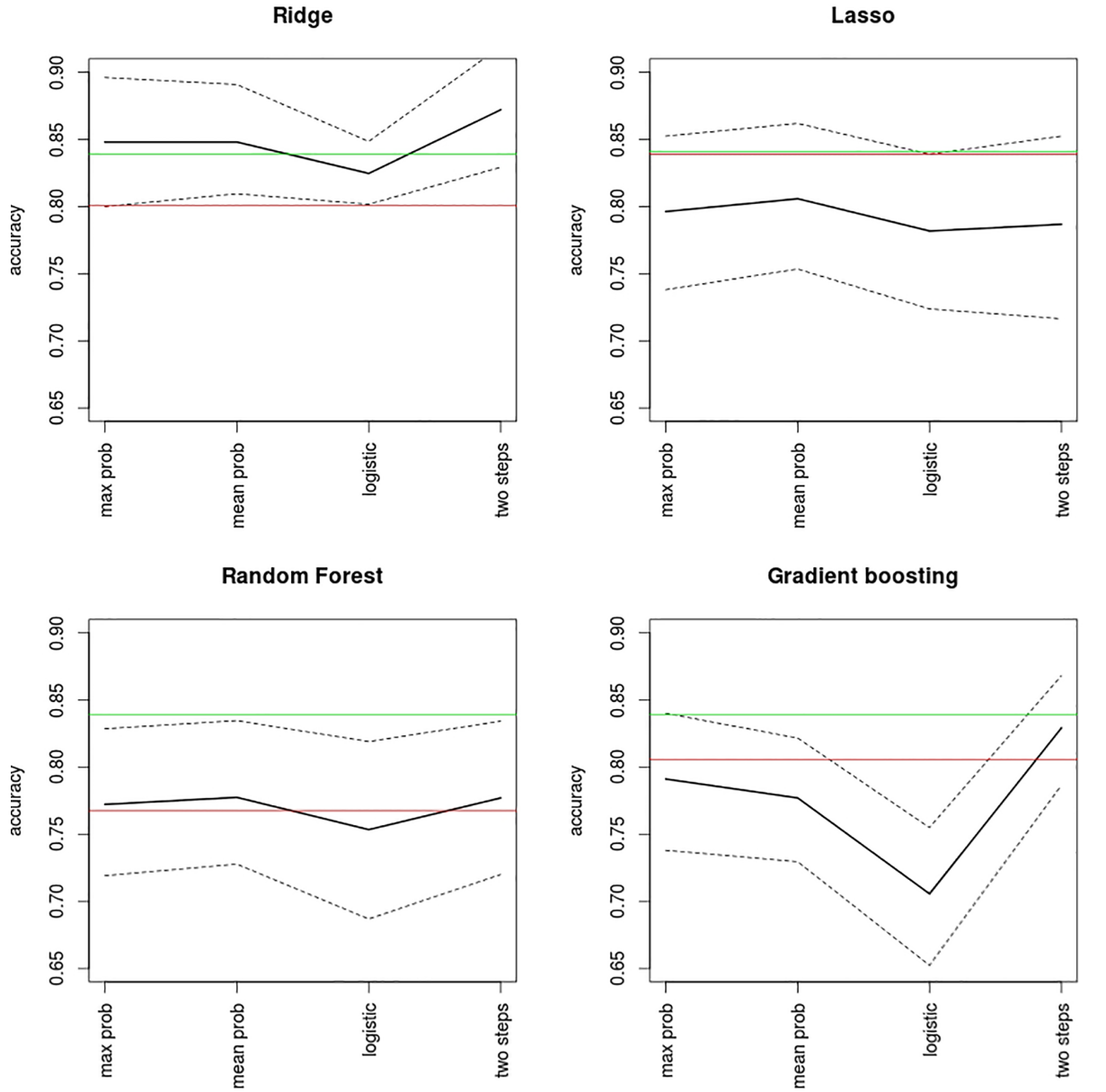


FIGURE 4 | Accuracy levels reported by probability-based and two-step multimodal integration approaches. Mean accuracy (black line) and its bootstrap 95% confidence levels (dashed lines) are shown for each classifier. Mean accuracies achieved by the best unimodal classification (red line) and by the 1D-CNN algorithm (green line) are also shown for the purposes of comparison. Max prob, maximum output probability algorithm; mean prob, mean output probability algorithm; logistic, logistic model on the probabilities.

the highest multimodal accuracies were achieved, all other multimodal accuracies were lower than that of the 1D-CNN classifier. Even so, the 1D-CNN classifier, with 84% accuracy, did not outperform the best unimodal result (i.e. Lasso on the 2back images, which also had 84% accuracy). The highest accuracies of all were obtained by the two-step Ridge classifier (87% accuracy) followed by the Ridge maximum and mean probability classifiers (both with 85% accuracy). From **Figure 4** it can also be concluded that, while Ridge is an algorithm that seems to adjust well to multimodal integration, Lasso is not – as with the later multimodal strategies provide clearly lower accuracies than the best unimodal classifier. A similar pattern, although not as clear-cut as in Lasso was observed for Gradient Boosting, while Random Forests delivered multimodal accuracies similar to those from the best unimodal classifier.

The anatomical overlap between voxels selected in the first step of the sequential Ridge algorithm (the best performing algorithm) was remarkably low (**Figure 5**). As expected, the largest overlaps occurred between maps from the same image modalities (i.e., 1back-2back 39.5%, ALFF-GBC 25.6%) while the

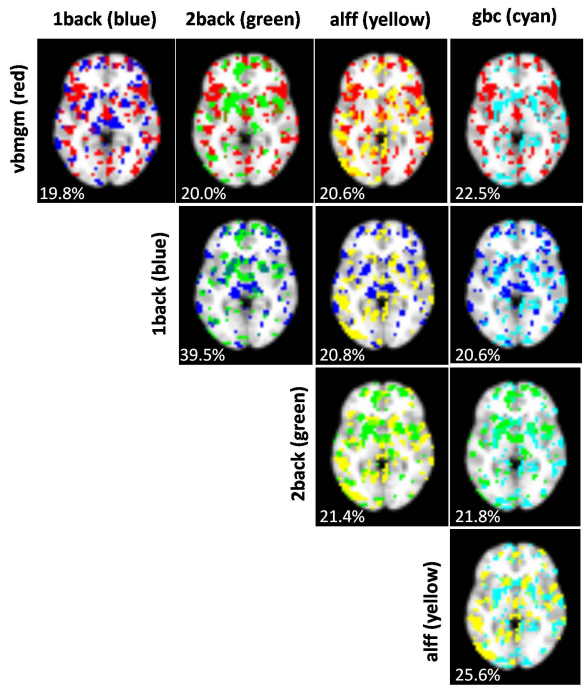


FIGURE 5 | Images showing, for each pair of brain maps, the degree of overlap between voxels selected in the first step of the sequential Ridge algorithm. Percentages of overlap are given for each pair. Chance overlap under a scenario of complete spatial independence is 4%.

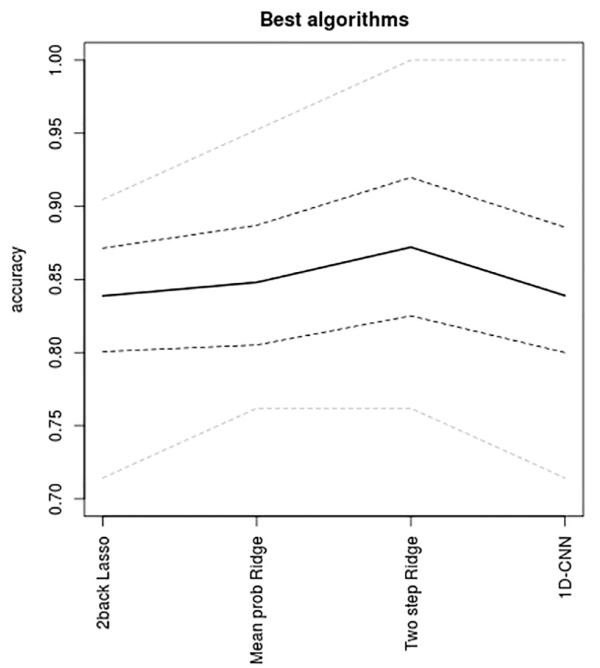


FIGURE 6 | Accuracies achieved by the best-performing algorithm in the unimodal setting and for the three multimodal integrative strategies, namely, unimodal Lasso applied to the 2back maps, the mean of the output probabilities from Ridge (which performed equally well as the maximum probability algorithm), the two-step sequential Ridge algorithm, and the one-dimensional convolutional neural network. Continuous line, mean accuracies; gray dashed lines, 95% bootstrap intervals of mean accuracies; gray doted lines, maximum and minimum accuracies delivered by the 10-fold scheme.

remaining pairs had values near 20%. Although this is clearly larger than the 4% overlap expected by complete chance (as the top 0.2 voxels were selected for each map), the observed levels of spatial overlap were, in general, low. This is in agreement with the low levels of redundancy reported in **Figure 3**. **Supplementary Figure 1** shows the distribution of values for the regression coefficients selected in the first step of the two-step Ridge classifier. As a summary of all findings **Figure 6** shows the highest test accuracy provided by the unimodal classifiers and by each one of the three multimodal integration strategies.

DISCUSSION AND CONCLUSION

The main results from the four unimodal classifiers have given the highest predictive power to the 2back maps, suggesting that task-based fMRI may have a more relevant role than other MRI data sources in diagnostic prediction. On the other hand, although multimodal classifiers have not led to accuracies much higher than those provided by the unimodal classifiers, the low redundancy levels observed between modalities indicate that multimodal integration is a potentially valuable strategy that should be pursued in future studies.

It is remarkable that a single brain map (the 2back map) allowed for a high accuracy between patients and controls, and this is quite likely to be behind the best accuracies reached by the two-step Ridge classifier (87%). This figure is higher than the accuracies reported in recent multimodal classification studies in schizophrenia such as the 83% of Wu et al. (2018) and substantially higher than the 75% of Cabral et al. (2016). However, these studies did not include task-based fMRI data as in ours. It is noteworthy that classification studies involving patients with schizophrenia in which task fMRI data were used for prediction have usually reported high accuracy levels (Wolfers et al., 2015; Arbabshirani et al., 2017) although some of these studies, especially the older ones, may have been unreliable due to small sample sizes.

Interestingly, the best-performing multimodal classifiers in our study were based on the Ridge logistic regression, which is a rather simple and inflexible algorithm (it implies both additivity and linearity). This is in contrast with the general trend in the broad machine learning community, which favors more flexible methods based on neural networks and the gradient boosting algorithm (Chollet, 2018). Although success of the Ridge classifier could be explained by the specific nature of the information contained in the images, it is also probable that its performance over more flexible methods is partly due to the fact that the latter usually require substantially larger amounts of data to exceed other methods. In this sense, our proposed 1D-CNN could have been better trained, its structure could have been enriched with the inclusion of several hidden layers, and it might have delivered higher accuracies if a larger dataset was available.

Of interest are the low levels of redundancy reported between the different brain maps. These low redundancies, together with the small degree of anatomical overlap observed between voxels of the different maps as selected in the first step of the sequential Ridge algorithm, point to valuable exclusive information and to a wide margin of potential improvement by multimodal integration. In spite of this, such improvement was not large (from 84 to 87%), reflecting the intrinsic difficulties of efficiently combining information from different sources. At the same time, it may have been unrealistic to expect accuracies significantly larger than those achieved as the patient sample used in the study was rather heterogeneous, including individuals with a wide range of clinical profiles and of illness durations, ranging from subjects on a first episode to subjects with many years of illness evolution.

The profile of the patient sample has also other implications for the applicability of the developed algorithms. Ideally, the development of an MRI-based diagnostic tool should be based on and aimed at the appropriate target population, which in the case of schizophrenia would be individuals at risk or undergoing a first episode of psychosis. To date, however, there are few studies of multimodal MRI integration entirely based on samples of first-episode patients. Two examples are the study by Peruzzo et al. (2015), which used small sample sizes or the more recent study by Ebdrup et al. (2018), which found no predictive power in the MRI datasets.

Finally, it is also appropriate to consider another limitation of the study. As explained in the section "Materials and Methods," downsampling to 4-mm voxels was made in order to reduce computational and memory-storage costs. Machine learning algorithms heavily rely on computational power, and their implementation may easily lead to the saturation of computational resources. In our study, this was especially relevant as, due to the multimodality, we were dealing with several datasets for each individual. Reduction of spatial resolution through voxel downsampling or ROI averaging may lead to some loss of relevant information. Still, it is difficult to evaluate the extent of this effect as downsampling differentially affects each modality and it has been shown to depend on the specific target groups (Gardumi et al., 2016; Mandelkow et al., 2017; Lancaster et al., 2018).

In summary, several conclusions may be drawn from our study. First, out of all MRI modalities evaluated, task-based fMRI appears to be the best option for unimodal diagnosis in schizophrenia. Secondly, the low levels of redundancy found between modalities suggest that multimodal integration is a potentially valuable strategy. Specifically, a simple and robust two-step algorithm based on the Ridge regression emerges as a suitable approach to multimodal diagnostic prediction in schizophrenia. Nevertheless, new studies based on samples of high-risk and first-episode patients will be required to develop valid multimodal MRI diagnostic tools in the disorder.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comité de Ética de la Investigación Clínica (CEIC-CEI) FIDMAG Hermanas Hospitalarias. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RS, PM, EP-C, and BC-F: definition and design of the study. AG-P, SS, and EP-C: recruitment of participants. TM, AG-P, and EC-R: MRI data acquisition. RS, DT-G, and EC-R: data analysis and development of algorithms. RS, PM, EP-C, and BC-F: manuscript writing and revision.

FUNDING

This work was supported by several grants funded by the Instituto de Salud Carlos III (co-funded by the European Regional Development Fund/European Social Fund "Investing in your future"): Miguel Servet Research Contract (CPII16/00018 to EP-C), Sara Borrell Research Contract (CD18/00029 to EC-R), Research Projects (PI14/01151 and PI18/00877 to RS, PI14/01148 to EP-C, PI18/00880 to PM, and PI17/01056 to DT-G), and an intramural grant (SAM18PI13 from CIBERSAM). It was also supported by the Generalitat de Catalunya: 2014SGR1573 to EP-C and SLT006/17/357 from the Departament de Salut to RS. The funding organizations played no role in the study design, data collection and analysis, or manuscript approval.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01203/full#supplementary-material

REFERENCES

Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage* 145, 137–165. doi: 10.1016/j.neuroimage.2016.02.079

Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V. D., Dwyer, D. B., Von Saldern, S., et al. (2016). Classifying schizophrenia using multimodal multivariate pattern recognition analysis: evaluating the impact of individual clinical profiles on the neurodiagnostic performance. *Schizophr. Bull.* 42(Suppl. 1), S110–S117. doi: 10.1093/schbul/sbw053

Calhoun, V. D., and Adali, T. (2009). Feature-based fusion of medical imaging data. *IEEE Trans. Inf. Technol. Biomed* 13, 711–720. doi: 10.1109/TITB.2008.923773

Chollet, F. O. (2018). *Deep Learning With Python*. New York, NY: Manning Publications Co.

Cole, M. W., Pathak, S., and Schneider, W. (2010). Identifying the brain's most globally connected regions. *Neuroimage* 49, 3132–3148. doi: 10.1016/j.neuroimage.2009.11.001

Del Ser, T., Gonzalez-Montalvo, J. I., Martinez-Espinosa, S., Delgado-Villapalos, C., and Bermejo, F. (1997). Estimation of premorbid intelligence in Spanish people with the word accentuation test and its application to the diagnosis of dementia. *Brain Cogn.* 33, 343–356. doi: 10.1006/brcg.1997.0877

Ebdrup, B. H., Axelsen, M. C., Bak, N., Fagerlund, B., Oranje, B., Raghava, J. M., et al. (2018). Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naive schizophrenia patients. *Psychol. Med.* 18, 1–10. doi: 10.1017/S0033291718003781

Friedman, J., Hastie, T., and Tibshirani, R. (2010). regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

Fuentes-Claramonte, P., Martin-Subero, M., Salgado-Pineda, P., Alonso-Lana, S., Moreno-Alcazar, A., Argila-Plaza, I., et al. (2019). Shared and differential default-mode related patterns of activity in an autobiographical, a self-referential and an attentional task. *PLoS One* 14:e0209376. doi: 10.1371/journal.pone.0209376

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., and Uludag, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *Neuroimage* 132, 32–42. doi: 10.1016/j.neuroimage.2016.02.033

Gevins, A., and Cutillo, B. (1993). Spatiotemporal dynamics of component processes in human working memory. *Electroencephalogr. Clin. Neurophysiol.* 87, 128–143. doi: 10.1016/0013-4694(93)90119-g

Hastie, T., Tibshirani, R., and Friedman, J. K. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. New York, NY: Springer.

Lancaster, J., Lorenz, R., Leech, R., and Cole, J. H. (2018). Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Front. Aging Neurosci.* 10:28. doi: 10.3389/fnagi.2018.00028

Mandelkow, H., de Zwart, J. A., and Duyn, J. H. (2017). Effects of spatial fMRI resolution on the classification of naturalistic movies. *Neuroimage* 162, 45–55. doi: 10.1016/j.neuroimage.2017.08.053

Peruzzo, D., Castellani, U., Perlini, C., Bellani, M., Marinelli, V., Rambaldelli, G., et al. (2015). Classification of first-episode psychosis: a multi-modal multi-feature approach integrating structural and diffusion imaging. *J. Neural. Transm.*122, 897–905. doi: 10.1007/s00702-014-1324-x

Salvador, R., Landin-Romero, R., Anguera, M., Canales-Rodriguez, E. J., Radua, J., Guerrero-Pedraza, A., et al. (2017a). Non redundant functional brain connectivity in schizophrenia. *Brain Imaging Behav.* 11, 552–564. doi: 10.1007/s11682-016-9535-4

Salvador, R., Radua, J., Canales-Rodriguez, E. J., Solanes, A., Sarro, S., Goikolea, J. M., et al. (2017b). Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS One* 12:e0175683. doi: 10.1371/journal.pone.0175683

Salvador, R., Vega, D., Pascual, J. C., Marco, J., Canales-Rodriguez, E. J., Aguilar, S., et al. (2016). Converging medial frontal resting state and diffusion-based abnormalities in borderline personality disorder. *Biol. Psychiatry* 79, 107–116. doi: 10.1016/j.biopsych.2014.08.026

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. (2018). Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's Disease. *IEEE J. Biomed Health Inform.* 22, 173–183. doi: 10.1109/JBHI.2017.2655720

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., et al. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia. *Neuroimage* 66, 119–132. doi: 10.1016/j.neuroimage.2012.10.051

Suk, H. I., Lee, S. W., Shen, D., Alzheimer's Disease Neuroimaging Initiative (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Tulay, E. E., Metin, B., Tarhan, N., and Arikan, M. K. (2019). Multimodal neuroimaging: basic concepts and classification of neuropsychiatric diseases. *Clin. EEG Neurosci.* 50, 20–33. doi: 10.1177/1550059418782093

Wang, S., and Summers, R. M. (2012). Machine learning and radiology. *Med. Image Anal.* 16, 933–951.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., and Marquand, A. F. (2015). From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* 57, 328–349. doi: 10.1016/j.neubiorev.2015.08.001

Wu, F., Zhang, Y., Yang, Y., Lu, X., Fang, Z., Huang, J., et al. (2018). Structural and functional brain abnormalities in drug-naive, first-episode, and chronic patients with schizophrenia: a multimodal MRI study. *Neuropsychiatr. Dis. Treat.* 14, 2889–2904. doi: 10.2147/NDT.S174356

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., Ourselin, S., et al. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neuroimage Clin.* 2, 735–745. doi: 10.1016/j.nicl.2013.05.004

Zang, Y. F., He, Y., Zhu, C. Z., Cao, Q. J., Sui, M. Q., Liang, M., et al. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain Dev.* 29, 83–91. doi: 10.1016/j.braindev.2006.07.002

Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., et al. (2016). Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. *Brain Imaging Behav.* 10, 1148–1159. doi: 10.1007/s11682-015-9480-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Salvador, Canales-Rodríguez, Guerrero-Pedraza, Sarró, Tordesillas-Gutiérrez, Maristany, Crespo-Facorro, McKenna and Pomarol-Clotet. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 10 December 2019
doi: 10.3389/fnins.2019.01321

Analyzing Neuroimaging Data Through Recurrent Deep Learning Models

Armin W. Thomas[1,2,3,4], *Hauke R. Heekeren*[2,4]*, *Klaus-Robert Müller*[1,5,6]* and *Wojciech Samek*[7]*

[1] Machine Learning Group, Technische Universität Berlin, Berlin, Germany, [2] Center for Cognitive Neuroscience Berlin, Freie Universität Berlin, Berlin, Germany, [3] Max Planck School of Cognition, Leipzig, Germany, [4] Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany, [5] Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea, [6] Max Planck Institute for Informatics, Saarbrücken, Germany, [7] Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

The application of deep learning (DL) models to neuroimaging data poses several challenges, due to the high dimensionality, low sample size, and complex temporo-spatial dependency structure of these data. Even further, DL models often act as *black boxes*, impeding insight into the association of cognitive state and brain activity. To approach these challenges, we introduce the DeepLight framework, which utilizes long short-term memory (LSTM) based DL models to analyze *whole-brain* functional Magnetic Resonance Imaging (fMRI) data. To decode a cognitive state (e.g., seeing the image of a house), DeepLight separates an fMRI volume into a sequence of axial brain slices, which is then sequentially processed by an LSTM. To maintain interpretability, DeepLight adapts the layer-wise relevance propagation (LRP) technique. Thereby, decomposing its decoding decision into the contributions of the single input voxels to this decision. Importantly, the decomposition is performed on the level of single fMRI volumes, enabling DeepLight to study the associations between cognitive state and brain activity on several levels of data granularity, from the level of the group down to the level of single time points. To demonstrate the versatility of DeepLight, we apply it to a large fMRI dataset of the Human Connectome Project. We show that DeepLight outperforms conventional approaches of uni- and multivariate fMRI analysis in decoding the cognitive states and in identifying the physiologically appropriate brain regions associated with these states. We further demonstrate DeepLight's ability to study the fine-grained temporo-spatial variability of brain activity over sequences of single fMRI samples.

Keywords: decoding, neuroimaging, fMRI, whole-brain, deep learning, recurrent, interpretability

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Liang Zhan,
University of Pittsburgh, United States
Gabriele Lohmann,
Max Planck Institute for Biological Cybernetics, Germany

***Correspondence:**
Hauke R. Heekeren
hauke.heekeren@fu-berlin.de
Klaus-Robert Müller
klaus-robert.mueller@tu-berlin.de
Wojciech Samek
wojciech.samek@hhi.fraunhofer.de

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *19 June 2019*
Accepted: *25 November 2019*
Published: *10 December 2019*

Citation:
Thomas AW, Heekeren HR, Müller K-R and Samek W (2019) Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. Front. Neurosci. 13:1321. doi: 10.3389/fnins.2019.01321

INTRODUCTION

Neuroimaging research has recently started collecting large corpora of experimental functional Magnetic Resonance Imaging (fMRI) data, often comprising many hundred individuals (e.g., Poldrack et al., 2013; Van Essen et al., 2013). By collecting these datasets, researchers want to gain insights into the associations between the cognitive states of an individual (e.g., while viewing images or performing a specific task) and the underlying brain activity, while also studying the variability of these associations across the population.

At first sight, the analysis of neuroimaging data thereby seems ideally suited for the application of deep learning (DL; LeCun et al., 2015; Goodfellow et al., 2016) methods, due to the availability of large and structured datasets. Generally, DL can be described as a class of representation-learning methods, with multiple levels of abstraction. At each level, the representation of the input data is transformed by a simple, but non-linear function. The resulting hierarchical structure of non-linear transforms enables DL methods to learn complex functions. It also enables them to identify intricate signals in noisy data, by projecting the input data into a higher-level representation, in which those aspects of the input data that are irrelevant to identify an analysis target are suppressed and those that are relevant are amplified. With this higher-level perspective, DL methods can associate a target variable with variable patterns in the input data. Importantly, DL methods can autonomously learn these projections from the data and therefore do not require a thorough prior understanding of the mapping between input data and analysis target (for a detailed discussion, see LeCun et al., 2015). For these reasons, DL methods seem ideally suited for the analysis of neuroimaging data, where intricate, highly variable patterns of brain activity are hidden in large, high-dimensional datasets and the mapping between cognitive state and brain activity is often unknown.

While researchers have started exploring the application of DL models to neuroimaging data (e.g., Plis et al., 2014; Suk et al., 2014; Nie et al., 2016; Sarraf and Tofighi, 2016; Mensch et al., 2018; Petrov et al., 2018; Yousefnezhad and Zhang, 2018), two major challenges have so far prevented broad DL usage: (1) Neuroimaging data are high dimensional, while containing comparably few samples. For example, a typical fMRI dataset comprises up to a few hundred samples per subject and recently up to several hundred subjects (e.g., Van Essen et al., 2013), while each sample contains several hundred thousand dimensions (i.e., voxels). In such analysis settings, DL models (as well as more traditional machine learning approaches) are likely to suffer from overfitting (by too closely capturing those dynamics that are specific to the training data, so that their predictive performance does not generalize well to new data). (2) DL models have often been considered as non-linear *black box models*, disguising the relationship between input data and decoding decision. Thereby, impeding insight into (and interpretation of) the association between cognitive state and brain activity.

To approach these challenges, we propose the DeepLight framework, which defines a method to utilize long short-term memory (LSTM) based DL architectures (Hochreiter and Schmidhuber, 1997; Donahue et al., 2015) to analyze whole-brain neuroimaging data. In DeepLight, each whole-brain volume is sliced into a sequence of axial images. To decode an underlying cognitive state, the resulting sequence of images is processed by a combination of convolutional and recurrent DL elements. Thereby, DeepLight successfully copes with the high dimensionality of neuroimaging data, while modeling the full spatial dependency structure of whole-brain activity (within and across axial brain slices). Conceptually, DeepLight builds upon the searchlight approach. Instead of moving a small searchlight beam around in space, DeepLight explores brain activity more in-depth, by looking through the full sequence of axial brain slices, before making a decoding decision. To subsequently relate brain activity and cognitive state, DeepLight applies the layer-wise relevance propagation (LRP; Bach et al., 2015; Lapuschkin et al., 2016) method to its decoding decisions. Thereby, decomposing these decisions into the contributions of the single input voxels to each decision. Importantly, the LRP analysis is performed on the level of a single input samples, enabling an analysis on several levels of data granularity, from the level of the group down to the level of single subjects, trials and time points. These characteristics make DeepLight ideally suited to study the fine-grained temporo-spatial distribution of brain activity underlying sequences of single fMRI samples.

Here, we will demonstrate the versatility of DeepLight, by applying it to an openly available fMRI dataset of the Human Connectome Project (Van Essen et al., 2013). In particular, to the data of an N-back task, in which 100 subjects viewed images of either body parts, faces, places or tools in two separate fMRI experiment runs (for an overview, see section Experiment Paradigm and **Figure S1**). Subsequently, we will evaluate the performance of DeepLight in decoding the four underlying cognitive states (resulting from viewing an image of either of the four stimulus classes) from the fMRI data and identifying the brain regions associated with these states. To this end, we will compare the performance of DeepLight to three representative conventional approaches to the uni- and multivariate analysis of neuroimaging data, with widespread application in the literature. In particular, we will compare DeepLight to the General Linear Model (GLM; Friston et al., 1994), searchlight analysis (Kriegeskorte et al., 2006) and whole-brain Least Absolute Shrinkage Logistic Regression (whole-brain Lasso; Grosenick et al., 2013; Wager et al., 2013). Note that the four analysis approaches differ in the number of voxels they include in their analyses. While the GLM analyses the data of single voxels independent of one another (univariate), the searchlight analysis utilizes the data of clusters of multiple voxels (multivariate) and the whole-brain lasso utilizes the data of all voxels in the brain (whole-brain). In this comparison, we find that DeepLight (1) decodes the cognitive states underlying the fMRI data more accurately than these other approaches, (2) improves its decoding performance better with growing datasets, (3) accurately identifies the physiologically appropriate associations between cognitive states and brain activity, and (4) identifies these associations on multiple levels of data granularity (namely, the level of the group, subject, trial and time point). We also demonstrate DeepLight's ability to study the temporo-spatial distribution of brain activity over a sequence of single fMRI samples.

METHODS

Experiment Paradigm

Hundred participants performed a version of the N-back task in two separate fMRI runs (for an overview, see **Figure S1** and Barch et al., 2013). Each of the two runs (260 s each) consisted of eight task blocks (25 s each) and four fixation blocks (15 s each). Within

each run, the four different stimulus types (body, face, place and tool) were presented in separate blocks. Half of the task blocks used a 2-back working memory task (participants were asked to respond "target" when the current stimulus was the same as the stimulus 2 back) and the other half a 0-back working memory task (a target cue was presented at the beginning of each block and the participants were asked to respond "target" whenever the target cue was presented in the block). Each task block consisted of 10 trials (2.5 s each). In each trial, a stimulus was presented for 2 s followed by a 500 ms interstimulus interval (ISI). We were not interested in identifying any effect of the N-back task condition on the evoked brain activity and therefore pooled the data of both N-back conditions.

FMRI Data Acquisition and Preprocessing

Functional MRI data of 100 unrelated participants for this experiment were provided in a preprocessed format by the Human Connectome Project (HCP S1200 release), WU Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Whole-brain EPI acquisitions were acquired with a 32 channel head coil on a modified 3T Siemens Skyra with TR = 720 ms, TE = 33.1 ms, flip angle = 52 deg, BW = 2,290 Hz/Px, in-plane FOV = $208 \times 180mm$, 72 slices, 2.0 mm isotropic voxels with a multi-band acceleration factor of 8. Two runs were acquired, one with a right-to-left and the other with a left-to-right phase encoding (for further methodological details on fMRI data acquisition, see Uğurbil et al., 2013).

The Human Connectome Project preprocessing pipeline for functional MRI data ("fMRIVolume"; Glasser et al., 2013) includes the following steps: gradient unwarping, motion correction, fieldmap-based EPI distortion correction, brain-boundary based registration of EPI to structural T1-weighted scan, non-linear registration into MNI152 space, and grand-mean intensity normalization (for further details, see Glasser et al., 2013; Uğurbil et al., 2013). In addition to the minimal preprocessing of the fMRI data that was performed by the Human Connectome Project, we applied the following preprocessing steps to the data for all decoding analyses: volume-based smoothing of the fMRI sequences with a 3 mm Gaussian kernel, linear detrending and standardization of the single voxel signal time-series (resulting in a zero-centered voxel time-series with unit variance) and temporal filtering of the single voxel time-series with a butterworth highpass filter and a cutoff of 128 s, as implemented in Nilearn 0.4.1 (Abraham et al., 2014). In line with previous work (Jang et al., 2017), we further applied an outer brain mask to each fMRI volume. We first identified those voxels whose activity was larger than 5% of the maximum voxel signal within the fMRI volume and then only kept those voxels for further analysis that were positioned between the first and last voxel to fulfill this property in the three spatial dimensions of any functional brain volume of our dataset. This resulted in a brain mask spanning $74 \times 92 \times 81$ voxels ($X \times Y \times Z$).

All of our preprocessing was performed by the use of Nilearn 0.4.1 (Abraham et al., 2014). Importantly, we did not exclude any TR of an experiment block of the four stimulus classes from the decoding analyses. However, we removed all fixation blocks from the decoding analyses. Lastly, we split the fMRI data of the 100 subjects contained in the dataset into two distinct training and test datasets (each containing the data of 70 and 30 randomly assigned subjects). All analyses presented throughout the following solely include the data of the 30 subjects contained in the held-out test dataset (if not stated otherwise).

Data Availability

The data that support the findings of this study are openly available at the ConnectomeDB S1200 Project page of the Human Connectome Project (https://db.humanconnectome.org/data/projects/HCP1200).

Baseline Methods

General Linear Model

The General Linear Model (GLM; Friston et al., 1994) represents a univariate brain encoding model (Naselaris et al., 2011; Kriegeskorte and Douglas, 2018). Its goal is to identify an association between cognitive state and brain activity, by predicting the time series signal of a voxel from a set of experiment predictor:

$$Y = X\beta + \epsilon \tag{1}$$

Here, Y presents a $T \times N$ dimensional matrix containing the multivariate time series data of N voxels and T time points. X represents the design matrix, which is composed of $T \times P$ data points, where each column represents one of P predictors. Typically, each predictor represents a variable that is manipulated during the experiment (e.g., the presentation times of stimuli of one of the four stimulus classes). β represents a $P \times N$ dimensional matrix of regression coefficients. To mimic the blood-oxygen-level dependent (BOLD) response measured by the fMRI, each predictor is first convolved with a hemodynamic response function (HRF; Lindquist et al., 2009), before fitting the β-coefficients to the data. After fitting, the resulting brain map of β-coefficients indicates the estimated contribution of each predictor to the time series signal of each of the N voxels. ϵ represents a $T \times N$ dimensional matrix of error terms. Importantly, the GLM analyzes the time series signal of each voxel independently and thereby includes a separate set of regression coefficients for each voxel in the brain.

Searchlight Analysis

The searchlight analysis (Kriegeskorte et al., 2006) is a multivariate brain decoding model (Naselaris et al., 2011; Kriegeskorte and Douglas, 2018). Its goal is to identify an association between cognitive state and brain activity, by probing the ability of a statistical classifier to identify the cognitive state from the activity pattern of a small clusters of voxels. To this end, the entire brain is scanned with a sphere of a given radius (the searchlight) and the performance of the classifier in decoding the cognitive states is evaluated at each location, resulting in a brain map of decoding accuracies. These decoding accuracies indicate

how much information about the cognitive state is contained in the activity pattern of the underlying cluster of voxels. Here, we used a searchlight radius of 5.6 mm and a linear-kernel Support Vector Machine (SVM) classifier (if not reported otherwise).

Given a training dataset of T data points $\left[y_t, x_t\right]_{t=1}^{T}$, where x_t represents the activity pattern of a cluster of voxels at time point t and y_t the corresponding label, the SVM (Cortes and Vapnik, 1995) is defined as follows:

$$\hat{y}\left(x\right) = sign\left[\sum_{t=1}^{T} \alpha_t \; y_t \; \gamma\left(x, x_t\right) + b\right] \tag{2}$$

Here, α_t and b are positive constants, whereas $\gamma\left(x, x_t\right)$ represents the kernel function. We used a linear kernel function, as implemented in Nilearn 0.4.1 (Abraham et al., 2014). We then defined the decoding accuracy achieved by the searchlight analysis as the maximum decoding accuracy that was achieved at any searchlight location in the brain. Similarly, we used the searchlight location that achieved the highest decoding accuracy to make decoding predictions (for example, to compute the confusion matrix presented in **Figure 2C**).

Whole-Brain Least Absolute Shrinkage Logistic Regression

The whole-brain Least Absolute Shrinkage Logistic Regression (or whole-brain lasso; Grosenick et al., 2013; Wager et al., 2013) represents a whole-brain decoding model (Naselaris et al., 2011; Kriegeskorte and Douglas, 2018). It identifies an association between cognitive state and brain activity, by probing the ability of a logistic model to decode the cognitive state from whole-brain activity (with one logistic coefficient β_i per voxel i in the brain). To reduce the risk of overfitting, resulting from the large number of model coefficients, the whole-brain lasso applies Least Absolute Shrinkage regularization to the likelihood function of the logistic model (Tikhonov, 1943; Tibshirani, 1996). Thereby, forcing the logistic model to perform automatic variable selection during parameter estimation, resulting in sparse coefficient estimates (i.e., by forcing many coefficient estimates to be exactly 0). In particular, the optimization problem of the whole-brain lasso can be defined as follows (again, N denotes the number of voxels in the brain, T the number of fMRI sampling time points and $\left[y_t, x_t\right]_{t=1}^{T}$ the set of class labels and voxel values of each fMRI sample):

$$\min_{\beta} \left\{ -\sum_{t=1}^{T} \left[y_t \; \log \sigma\left(\beta^T x_t\right) + \left(1 - y_t\right) \log\left(1 - \sigma\left(\beta^T x_t\right)\right)\right] + \lambda \sum_{i=1}^{N} \left|\beta_i\right| \right\} \tag{3}$$

Here, λ represents the strength of the L1 regularization term (with larger values indicating stronger regularization), whereas σ represents the logistic model:

$$\sigma\left(x\right) = \frac{1}{1 + e^{-x}} \tag{4}$$

For each voxel i in the brain, the resulting set of coefficient estimates β, indicates the contribution of the activity of this voxel to the decoding decision $\sigma\left(x_t\right)$ of the logistic model for a whole-brain fMRI sample x_t at time point t. Over the recent years, the whole-brain lasso, as well as closely related decoding approaches (e.g., McIntosh and Lobaugh, 2004; Ryali et al., 2010; Gramfort et al., 2013), have found widespread application throughout the neuroscience literature (e.g., Wager et al., 2013; Chang et al., 2015).

DeepLight Framework

Deep Learning Model

The DL model underlying DeepLight consists of three distinct computational modules, namely a feature extractor, an LSTM, and an output unit (for an overview, see **Figure 1**). First, DeepLight separates each fMRI volume into a sequence of axial brain slices. These slices are then processed by a convolutional feature extractor (LeCun and Bengio, 1995), resulting in a sequence of higher-level, and lower-dimensional, slice representations. These higher-level slice representations are fed to an LSTM (Hochreiter and Schmidhuber, 1997), integrating the spatial dependencies of the observed brain activity within and across axial brain slices. Lastly, the output unit makes a decoding decision, by projecting the output of the LSTM into a lower-dimensional space, spanning the cognitive states in the data. Here, a probability for each cognitive state is estimated, indicating whether the input fMRI volume belongs to each of these states. This combination of convolutional and recurrent DL elements is inspired by previous research, showing that it is generally well-suited to learn the spatial dependency structure of long sequences of input data (Donahue et al., 2015; McLaughlin et al., 2016; Marban et al., 2019). Importantly, the DeepLight approach is not dependent on any specific architecture of each of these three modules. The DL model architecture described in the following is exemplary and derived from previous work (Marban et al., 2019). Further research is needed to explore the effect of specific module architectures on the performance of DeepLight.

The feature extractor used here was composed of a sequence of eight convolution layers (LeCun and Bengio, 1995). A convolution layer consists of a set of kernels (or filters) w that each learn local features of the input image a. These local features are then convolved over the input, resulting in an activation map h, indicating whether a feature is present at each given location of the input:

$$h_{i,j} = g\left(\sum_{k=0}^{m}\sum_{l=0}^{n} w_{k,l} a_{(i-k),(j-l)}\right) \tag{5}$$

Here, b represents the bias of the kernel, while g represents the activation function. k and l represent the row and column index of the kernel matrix, whereas i and j represent the row and column index of the activation map.

Generally, lower-level convolution kernels (that are close to the input data) have small receptive fields and are only sensitive to local features of small patches of the input data (e.g., contrasts and orientations). Higher-level convolution kernels, on the other hand, act upon a higher-level representation of the input data,

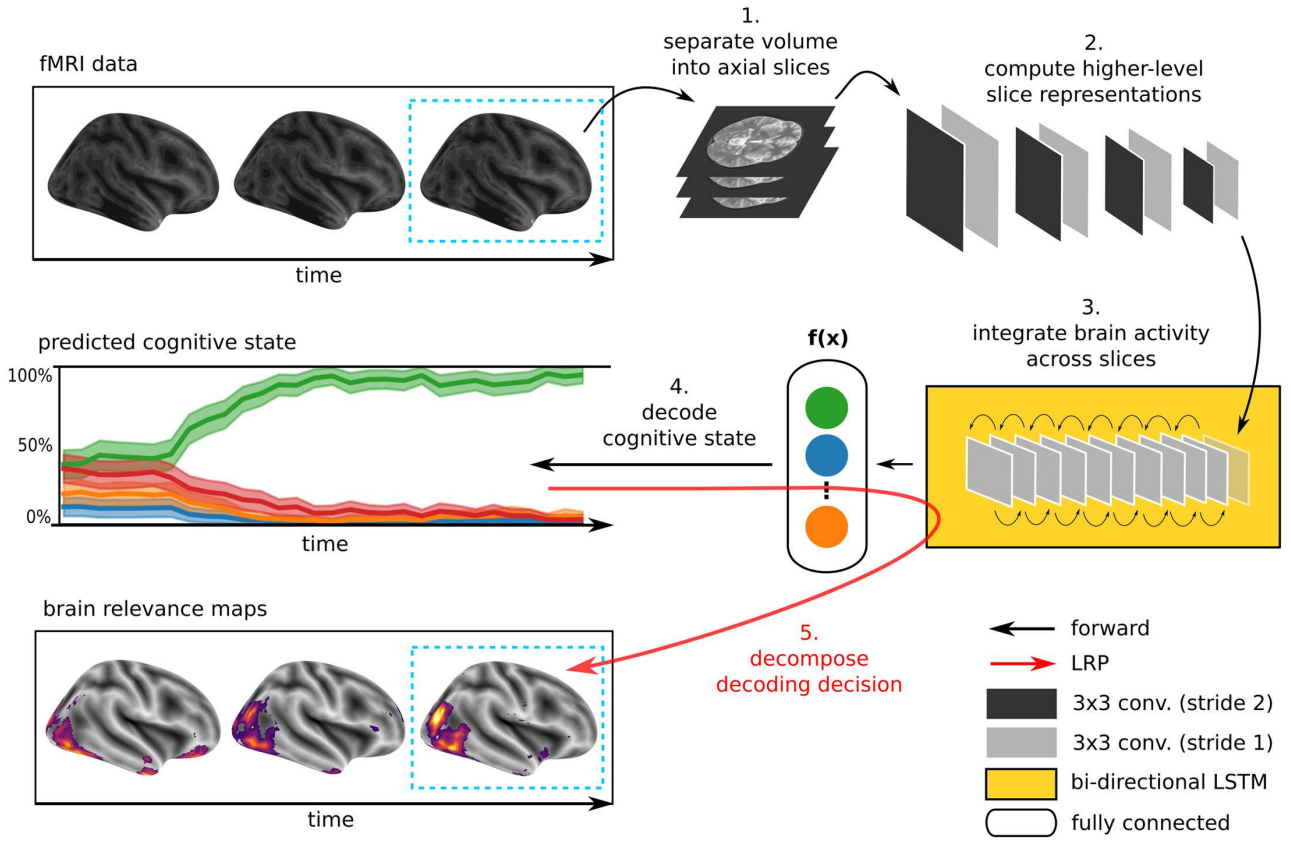


FIGURE 1 | Illustration of the DeepLight approach. A whole-brain fMRI volume is sliced into a sequence of axial images. These images are then passed to a DL model consisting of a convolutional feature extractor, an LSTM and an output unit. First, the convolutional feature extractor reduces the dimensionality of the axial brain slices through a sequence of eight convolution layers. The resulting sequence of higher-level slice representations is then fed to a bi-directional LSTM, modeling the spatial dependencies of brain activity within and across brain slices. Lastly, the DL model outputs a decoding decision about the cognitive state underlying the fMRI volume, through a softmax output layer with one output neuron per cognitive state in the data. Once the prediction is made, DeepLight utilizes the LRP method to decompose the prediction into the contributions (or relevance) of the single input voxels to the prediction. Thereby, enabling an analysis of the association between fMRI data and cognitive state.

which has already been transformed by a sequence of preceding lower-level convolution kernels. Higher-level kernels thereby integrate the information provided by lower-level convolution kernels, allowing them to identify larger and more complex patterns in the data. We specified the sequence of convolution layers as follows (see **Figure 1**): conv3-16, conv3-16, conv3-16, conv3-16, conv3-32, conv3-32, conv3-32, conv3-32 [notation: conv(kernel size) - (number of kernels)]. All convolution kernels were activated through a rectified linear unit function:

$$g(z) = max(0, z) \tag{6}$$

Importantly, all kernels of the even-numbered convolution layers were moved over the input fMRI slice with a stride size of one voxel and all kernels of odd-numbered layers with a stride size of two voxels. The stride size determines the dimensionality of the outputted slice representation. An increasing stride indicates more distance between the application of the convolution kernels to the input data. Thereby, reducing the dimensionality of the output representation at the cost of a decreasing sensitivity to differences in the activity patterns of neighboring voxels. Yet, the activity patterns of neighboring voxels are known to be highly correlated, leading to an overall low risk of information loss through a reasonable increase in stride size. To avoid any further loss of dimensionality between the convolution layers, we applied zero-padding. Thereby, adding zeros to the borders of the inputs to each convolution layer so that the outputs of the convolution layers have the same dimensionality as their inputs, if a stride of one voxel is applied, and only decrease in size, when a larger stride is used. The sequence of eight convolution layers thereby resulted in a 960-dimensional representation of each volume slice.

To integrate the information provided by the resulting sequence of slice representations into a higher-level representation of the observed whole-brain activity, DeepLight applies a bi-directional LSTM (Hochreiter and Schmidhuber, 1997), containing two independent LSTM units. Each of the two LSTM units iterates through the entire sequence of input slices,

but in reverse order (one from bottom-to-top and the other from top-to-bottom). An LSTM unit contains a hidden cell state C, storing information over an input sequence of length S with elements a_s and outputs a vector h_s for each input at sequence step s. The unit has the ability to add and remove information from C through a series of gates. In a first step, the LSTM unit decides what information from the cell state C is removed. This is done by a fully-connected logistic layer, the forget gate f:

$$f_t = \sigma\left(W_f a_s + U_f h_{s-1} + b_f\right) \tag{7}$$

Here, σ indicates the logistic function (see Equation 4), $[W, U]$ the gate's weight matrices and b the gate's bias. The forget gate outputs a number between 0 and 1 for each entry in the cell state C at the previous sequence step $s-1$. Next, the LSTM unit decides what information is going to be stored in the cell state. This operation contains two elements: the input gate i, which decides which values of C_s will be updated, and a tanh layer, which creates a new vector of candidate values C'_s:

$$i_s = \sigma\left(W_i a_s + U_i h_{s-1} + b_i\right) \tag{8}$$

$$C'_s = \tanh\left(W_c a_s + U_c h_{s-1} + b_c\right) \tag{9}$$

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{10}$$

Subsequently, the old cell state C_{s-1} is updated into the new cell state

$$C_s = f_s \cdot C_{s-1} + i_s \cdot C'_s \tag{11}$$

Lastly, the LSTM computes its output h_s. Here, the output gate o, decides what part of C_s will be outputted. Subsequently, C_s is multiplied by another *tanh* layer to make sure that h_s is scaled between -1 and 1:

$$o_s = \sigma\left(W_o a_s + U_o h_{s-1} + b_o\right) \tag{12}$$

$$h_s = o_s \cdot tanh\left(C_s\right) \tag{13}$$

Each of the two LSTM units in our DL model contained 40 output neurons. To make a decoding decision, both LSTM units pass their output for the last sequence element to a fully-connected softmax output layer. The output unit contains one neuron per cognitive state in the data and assigns a probability to each of the K (here, $K = 4$) states, indicating the probability that the current fMRI sample belongs to this state:

$$\sigma = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, with\ j = 1, ..., K \tag{14}$$

Layer-Wise Relevance Propagation in the DeepLight Framework

To relate the decoded cognitive state and brain activity, DeepLight utilizes the Layer-Wise Relevance Propagation (LRP; Bach et al., 2015; Montavon et al., 2017; Lapuschkin et al., 2019) method. The goal of LRP is to identify the contribution of a single dimension d of an input a (with dimensionality D) to the prediction $f(a)$ that is made by a linear or non-linear classifier f. We denote the contribution of a single dimension as its relevance R_d. One way of decomposing the prediction $f(a)$ is by the sum of the relevance values of each dimension of the input:

$$f(a) \approx \sum_{d=1}^{D} R_d \tag{15}$$

Qualitatively, any $R_d < 0$ can be interpreted as evidence against the presence of a classification target, while $R_d > 0$ denotes evidence for the presence of the target. Importantly, LRP assumes that $f(a) > 0$ indicates evidence for the presence of a target.

Let's assume the relevance $R_j^{(l)}$ of a neuron j at network layer l for the prediction $f(a)$ is known. We would like to decompose this relevance into the messages $R_{i\leftarrow j}^{(l-1,l)}$ that are sent to those neurons i in layer $l-1$ which provide the inputs to neuron j:

$$R_j^{(l)} = \sum_{i\epsilon(l)} R_{i\leftarrow j}^{(l-1,l)} \tag{16}$$

While the relevance of the output neuron at the last layer L is defined as $R_d^{(L)} = f(a)$, the dimension-wise relevance scores on the input neurons are given by $R_d^{(1)}$. For all weighted connections of the DL model in between (see Equations 5, 7, 8, 9, and 12), DeepLight defines the messages $R_{i\leftarrow j}^{(l-1,l)}$ as follows:

$$R_{i\leftarrow j}^{(l-1,l)} = \frac{z_{ij}}{z_j + \epsilon \cdot sign\left(z_j\right)} R_j^{(l)} \tag{17}$$

Here, $z_{ij} = a_i^{(l-1)} w_{ij}^{(l-1,l)}$ (w indicating the coefficient weight and a the input) and $z_j = \sum_i z_{ij}$, while ϵ represents a stabilizer term that is necessary to avoid numerical degenerations when z_j is close to 0 (we set $\epsilon = 0.001$).

Importantly, the LSTM also applies another type of connection, which we refer to as multiplicative connection (see Equations 11 and 13). Let z_j be an upper-layer neuron whose value in the forward pass is computed by multiplying two lower-layer neuron values z_g and z_s such that $z_j = z_g \cdot z_s$. These multiplicative connections occur when we multiply the outputs of a *gate* neuron, whose values range between 0 and 1, with an instance of the hidden cell state, which we will call *source* neuron. For these types of connections, we set the relevances of the gate neuron $R_g^{(l-1)} = 0$ and the relevances of the source neuron $R_s^{(l-1)} = R_j^{(l)}$, where $R_j^{(l)}$ denotes the relevances of the upper layer neuron z_j (as proposed in Arras et al., 2017). The reasoning behind this rule is that the gate neuron already decides in the forward pass how much of the information contained in the source neuron should be retained to make the classification. Even if this seems to ignore the values of the neurons z_g and z_s for the redistribution of relevance, these are actually taken into account when computing the value $R_j^{(l)}$ from the relevances

of the next upper-layer neurons to which z_j is connected by the weighted connections. We refer the reader to Samek et al. (2018) and Montavon et al. (2018) for more information about explanation methods.

In the context of this work, we decomposed the predictions of DeepLight for the actual cognitive state underlying each fMRI sample, as we were solely interested in understanding what DeepLight used as evidence in favor of the presence of this state. We also restricted the LRP analysis to those brain samples that the DL model classified correctly, because we can only assume that the DL model has learned a meaningful mapping between brain data and cognitive state, if it is able to accurately decode the cognitive state.

DeepLight Training

We iteratively trained DeepLight through backpropagation (Rumelhart et al., 1986) over 60 epochs by the use of the ADAM optimization algorithm as implemented in tensorflow 1.4 (Abadi et al., 2016). To prevent overfitting, we applied dropout regularization to all network layers (Srivastava et al., 2014), global gradient norm clipping (with a clipping threshold of 5; Pascanu et al., 2013), as well as an early stopping of the training (for an overview of training statistics, see **Figure S2**). During the training, we set the dropout probability to 50% for all network layers, except for the first four convolution layers, where we reduced the dropout probability to 30% for the first two layers and 40% for the third and fourth layer. Each training epoch was defined as a complete iteration over all samples in the training dataset (see section FMRI Data Acquisition and Preprocessing). We used a learning rate of 0.0001 and a batch size of 32. All network weights were initialized by the use of a normal-distributed random initialization scheme (Glorot and Bengio, 2010). The DL model was written in tensorflow 1.4 (Abadi et al., 2016) and the interprettensor library (https://github.com/VigneshSrinivasan10/interprettensor).

DeepLight Brain Maps

To generate a set of subject-level brain maps with DeepLight, we first decomposed the decoding decisions of DeepLight for each correctly classified fMRI sample of a subject with the LRP method (see section Layer-Wise Relevance Propagation in the DeepLight Framework). Importantly, we restricted the LRP analysis to those fMRI samples that were collected 5–15 s after the onset of the experiment block, as we expect the HRF (Lindquist et al., 2009) to be strongest within this time period. To then aggregate the resulting set of relevance maps for each decomposed fMRI sample within each cognitive state, we smoothed each relevance map with a 3 mm FWHM Gaussian kernel and averaged all relevance volumes belonging to a cognitive state, resulting in one brain map per subject and cognitive state. Group-level brain maps were then obtained, by averaging these subject-level brain maps for all subjects in the held-out test dataset within each cognitive state, resulting in one group-level brain map per cognitive state.

RESULTS

DeepLight Accurately Decodes Cognitive States From fMRI Data

A key prerequisite for the DeepLight analysis (as well as all other decoding analyses) is that it achieves reasonable performance in the decoding task at hand. Only then we can assume that it has learned a meaningful mapping from the fMRI data to the cognitive states and interpret the resulting brain maps as informative about these states.

Overall, DeepLight accurately decoded the cognitive states underlying 68.3% of the fMRI samples in the held-out test dataset (62.36, 69.87, 75.97, 65.09 for body, face, place and tool, respectively; **Figure 2A**). It generally performed best at discriminating the body and place (5.1% confusion in the held-out data), face and tool (7.8% confusion in the held-out data), body and tool (9.8% confusion in the held-out data) and face and place (10.4% confusion in the held-out data) stimuli from the fMRI data, while it did not perform as well in discriminating place and tool and body and face stimuli (15% confusion in the held-out data, respectively).

Note that DeepLight's performance in decoding the four cognitive states from the fMRI data varied over the course of an experiment block (**Figure 2B**). DeepLight performed best in the middle and later stages of the experiment block, where the average decoding accuracy reaches 80%. This finding is generally in line with the temporal evolution of the hemodynamic response function (HRF; Lindquist et al., 2009) measured by the fMRI (the HRF is known to be strongest 5–10 s after to the onset of the underlying neuronal activity).

To further evaluate DeepLight's performance in decoding the cognitive states from the fMRI data, we compared its performance in decoding these states to the searchlight analysis and whole-brain lasso. For simplicity, we sub-divided this comparison into a separate analysis on the group- and subject-level.

Group-Level

For the group-level comparison, we trained the searchlight analysis and whole-brain lasso on the data of all 70 subjects contained in the training dataset (for details on the fitting procedures, see section Parameter Estimation of the Baseline Methods in **Supplementary Information**). Subsequently, we evaluated their performance in decoding the cognitive states in the full held-out test data.

DeepLight clearly outperformed the other approaches in decoding the cognitive states. While the searchlight analysis achieved an average decoding accuracy of 60% (**Figure 2C**) and the whole-brain lasso an average decoding accuracy of 47.97% (**Figure 2D**), DeepLight improved upon these performances by 8.3% [$t_{(29)} = 5.80$, $p < 0.0001$] and 20.33% [$t_{(29)} = 13.39$, $p < 0.0001$], respectively.

All three decoding approaches generally performed best at discriminating face and place stimuli from the fMRI data (**Figures 2A,C,D**). Similar to DeepLight, the searchlight analysis and whole-brain lasso also performed well at discriminating body and place stimuli (3.3 and 12.2% confusion for the searchlight

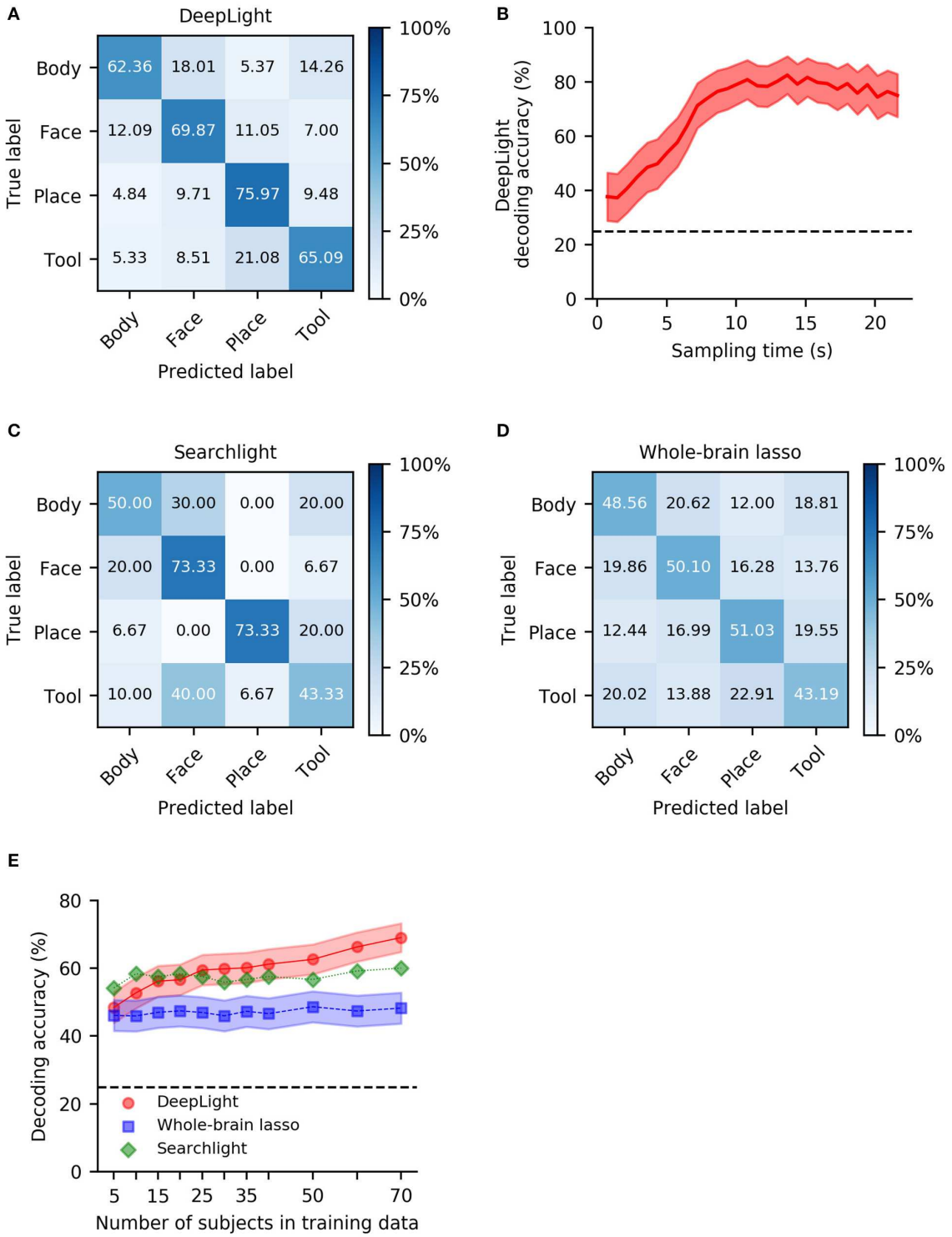


FIGURE 2 | Group-level decoding performance of DeepLight, the searchlight analysis and whole-brain lasso. **(A)** Confusion matrix of DeepLight's decoding decisions. **(B)** Average decoding performance of DeepLight over the course of an experiment block. **(C,D)** Confusion matrix for the decoding decisions of the group-level searchlight analysis **(C)** and whole-brain lasso **(D)**. **(E)** Average decoding accuracy of the searchlight (green), whole-brain lasso (blue) and DeepLight (red), when these are repeatedly trained on a subset of the subjects from the full training dataset. Black dashed horizontal lines indicate chance level.

analysis and whole-brain lasso, respectively, **Figures 2C,D**), while they also had more difficulties discriminating body and face stimuli from the fMRI data (25 and 20.2% confusion for the searchlight analysis and whole-brain lasso, respectively, **Figures 2C,D**).

A key premise of DL methods, when compared to more traditional decoding approaches, is that their decoding performance improves better with growing datasets. To test this, we repeatedly trained all three decoding approaches on a subset of the training dataset (including the data of 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, and 70 subjects), and validated their performance at each iteration on the full held-out test data (**Figure 2E**). Overall, the decoding performance of DeepLight increased by 0.27% [$t_{(10)} = 10.9$, $p < 0.0001$] per additional subject in the training dataset, whereas the performance of the whole-brain lasso increased by 0.03% [$t_{(10)} = 3.02$, $p = 0.015$] and the performance of the searchlight analysis only marginally increased by 0.04% [$t_{(10)} = 2.08$, $p = 0.067$]. Nevertheless, the searchlight analysis outperformed DeepLight in decoding the cognitive states from the data when only little training data were available (here, 10 or less subjects [$t_{(29)} = -4.39$, $p < 0.0001$]. The decoding advantage of DeepLight, on the other hand, came to light when the data of 50 or more subjects were available in the training dataset [$t_{(29)} = 3.82$, $p = 0.0006$]. DeepLight consistently outperformed the whole-brain lasso, when it was trained on the data of at least 10 subjects [$t_{(29)} = 5.32$, $p = 0.0045$].

Subject-Level

For the subject-level comparison, we first trained both, the searchlight analysis and whole-brain lasso on the fMRI data of the first experiment run of a subject from the held-out test dataset (for an overview of the training procedures, see section Parameter Estimation of the Baseline Methods in **Supplementary Information**). We then used the data of the second experiment run of the same subject to evaluate their decoding performance (by predicting the cognitive states underlying each fMRI sample of the second experiment run). Importantly, we also decoded the same fMRI samples with DeepLight. Note that DeepLight, in comparison to the other approaches, did not see any data of the subject during the training, as it was solely trained on the data of the 70 subjects in the training dataset (see section DeepLight Training).

DeepLight clearly outperformed the other decoding approaches, by decoding the cognitive states more accurately for 28 out of 30 subjects, when compared to the searchlight analysis (while the searchlight analysis achieved an average decoding accuracy of 47.2% across subjects, DeepLight improved upon this performance by 22.4%, with an average decoding accuracy of 69.3%, $t_{(29)} = 11.28$, $p < 0.0001$; **Figure 3A**), and for 29 out of 30 subjects, when compared to the whole-brain lasso (while the whole-brain lasso achieved an average decoding accuracy of 37% across subjects, DeepLight improved upon this performance by 32%; $t_{(29)} = 15.74$, $p < 0.0001$; **Figure 3B**).

To further ascertain that the observed differences in decoding performance between the searchlight and DeepLight did not result from the linearity contained in the Support Vector Machine (SVM; Cortes and Vapnik, 1995) of the searchlight analysis, we replicated our subject-level searchlight analysis, by the use of a non-linear radial basis function kernel (RBF; Cortes and Vapnik, 1995; Müller et al., 2001; Schölkopf and Smola, 2002) SVM (**Figure S3**). However, the decoding accuracies achieved by the RBF-kernel SVM were not meaningfully different from those of the linear-kernel SVM [$t_{(29)} = -1.75$, $p = 0.09$].

Lastly, we also compared the subject-level decoding performance of the whole-brain lasso to that of a recently proposed extension of this approach (TV-L1, for methodological details see Gramfort et al., 2013). The TV-L1 approach combines the Least Absolute Shrinkage Regularization (L1; see Equation 3) of the whole-brain lasso with an additional Total-Variation (TV) penalty (Michel et al., 2011), to better account for the spatial dependency structure of fMRI data. Yet, we found that the whole-brain lasso performed better at decoding the cognitive states from the subject-level fMRI data than TV-L1 [$t_{(29)} = 3.79$, $p = 0.0007$; see **Figure S4**].

DeepLight Identifies Physiologically Appropriate Associations Between Cognitive States and Brain Activity

Our previous analyses have shown that DeepLight has learned a meaningful mapping between the fMRI data and cognitive states, by accurately decoding these states from the data. Next, we therefore tested DeepLight's ability to identify the brain areas associated with the cognitive states, by decomposing its decoding decisions with the LRP method (see section DeepLight Framework). Subsequently, we compared the resulting brain maps of DeepLight to those of the GLM, searchlight analysis and whole-brain lasso. Again, we sub-divided this comparison into a separate analysis on the group- and subject-level. Note that due to the diverse statistical nature of the three baseline approaches, the values of their brain maps are on different scales and have different statistical interpretations (for methodological details, see section Baseline Methods). Further, all depicted brain maps in **Figures 4–6** are projected onto the inflated cortical surface of the FsAverage5 surface template (Fischl, 2012) for better visibility.

To evaluate the quality of the brain maps resulting from each analysis approach, we performed a meta-analysis of the four cognitive states with NeuroSynth (for details on NeuroSynth, see section NeuroSynth in **Supplementary Information** and Yarkoni et al., 2011). NeuroSynth provides a database of mappings between cognitive states and brain activity, based on the empirical neuroscience literature. Particularly, the resulting brain maps used here indicate whether the probability that an article reports a specific brain activation is different, when it includes a specific term (e.g., "face") compared to when it does not. With this meta-analysis, we defined a set of *regions-of-interest* (ROIs) for each cognitive state (as defined by the terms "body," "face," "place," and "tools"), in which we would expect the various analysis approaches to identify a positive association between the cognitive state and brain activity (for an overview, see **Figure 4A**). These ROIs were defined as follows: the upper parts of the middle and inferior temporal gyrus, the postcentral gyrus, as well as the right fusiform gyrus for the

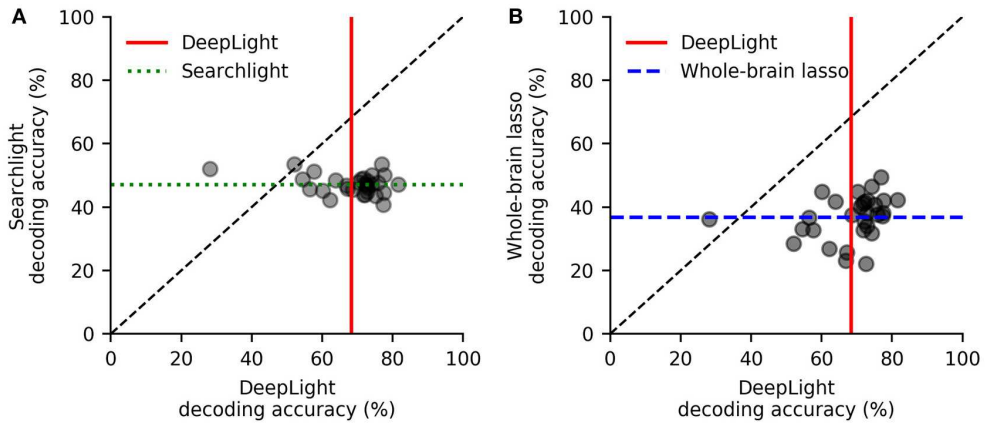


FIGURE 3 | Subject-level decoding performance comparison of DeepLight (red) to the searchlight analysis (**A**; green) and whole-brain lasso (**B**; blue). Black scatter points indicate the average decoding accuracy for a subject. Colored lines indicate the average decoding accuracy across all 30 test subjects.

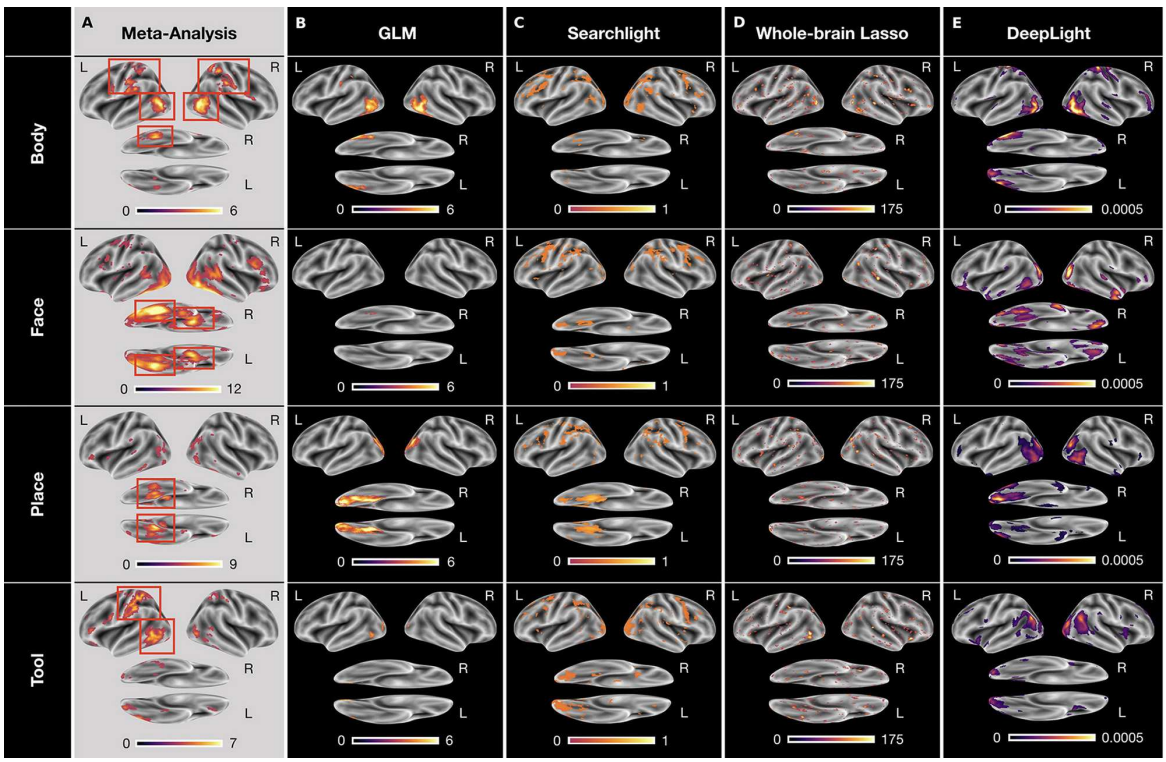


FIGURE 4 | Group-level brain maps for each cognitive state and analysis approach: **(A)** Results of a NeuroSynth meta-analysis for the terms "body," "face," "place," and "tools." The brain maps were thresholded at an expected false discovery rate of 0.01. Red boxes highlight the regions-of-interest for each cognitive state. **(B)** Results of the GLM group-level analysis. The brain maps of the GLM analysis were thresholded at an expected false discovery rate of 0.1. **(C–E)** Results of the group-level searchlight analysis **(C)**, whole-brain lasso **(D)**, and DeepLight **(E)**. The brain maps of the searchlight analysis, whole-brain lasso, and DeepLight were thresholded at the 90th percentile of their values. Note that the values of the brain maps are on different scales between analysis approaches, due to their different statistical nature. All brain maps are projected onto the inflated cortical surface of the FsAverage5 surface template (Fischl, 2012).

body state, the fusiform gyrus (also known as the fusiform face area FFA; Haxby et al., 2001; Heekeren et al., 2004) and amygdala for the face state, the parahippocampal gyrus (or parahippocampal place area PPA; Haxby et al., 2001; Heekeren et al., 2004) for the place state and the upper left middle and inferior temporal gyrus as well as the left postcentral gyrus for the tool state.

To ensure comparability with the results of the meta-analysis, we restricted all analyses of brain maps to the estimated positive associations between brain activity and cognitive states

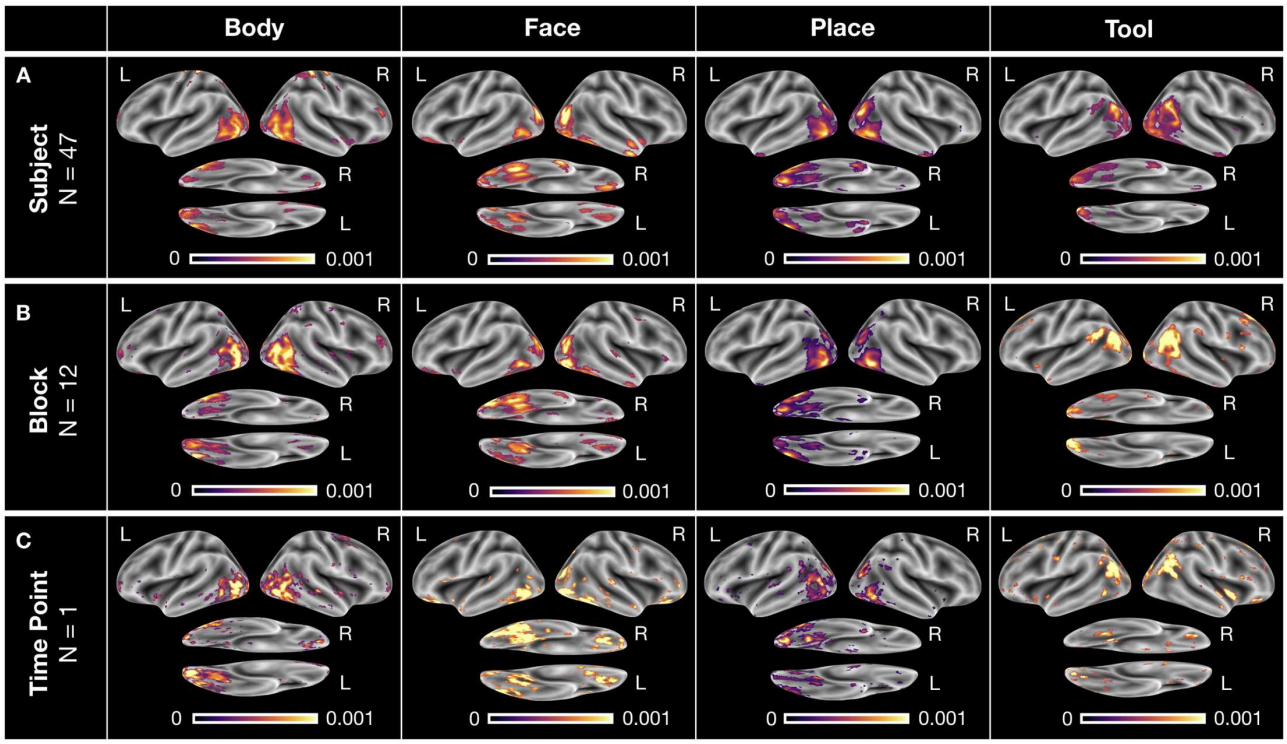


FIGURE 5 | Exemplary DeepLight brain maps for each of the four cognitive states on different levels of data granularity for a single subject. All brain maps belong to the subject with the highest decoding accuracy in the held-out test dataset. **(A)** Average relevance maps for all correctly classified TRs of the subject (with an average of 47 TRs per cognitive state). **(B)** Average relevance maps for all correctly classified TRs of the first experiment block of each cognitive state in the first experiment run (with an average of 12 TRs per cognitive state). **(C)** Exemplar relevance maps for a single TR of the first experiment block of each cognitive state in the first experiment run. All relevance maps were thresholded at the 90th percentile of their values and projected onto the inflated cortical surface of the FsAverage5 surface template (Fischl, 2012).

(i.e., positive relevance values as well as positive GLM and whole-brain lasso coefficients, see section Baseline Methods and section Parameter Estimation of the Baseline Methods in **Supplementary Information**). A negative Z-value in the meta-analysis indicates a lower probability that an article reports a specific brain activation when it includes a specific term, compared to when it does not include the term. A negative value in the meta-analysis is therefore conceptually different to negative values in the brain maps of our analyses (e.g., negative relevance values or negative whole-brain lasso coefficients). These can generally be interpreted as evidence against the presence of a cognitive state, given the specific set of cognitive states in our dataset (e.g., a negative relevance indicates evidence for the presence of any of the other cognitive states considered).

Group-Level

To determine the voxels that each analysis approach associated with a cognitive state, we defined a threshold for the values of each group-level brain map, indicating those voxels that are associated most strongly with the cognitive state. For the GLM analysis, we thresholded all *P*-values at an expected false discovery rate (Benjamini and Hochberg, 1995; Genovese et al., 2002) of 0.1 (**Figure 4B**). Similarly, for all decoding analyses, we thresholded each brain map at the 90th percentile of its values (**Figures 4C–E**). For the whole-brain lasso and DeepLight, the remaining 10 percent of values indicate those brain regions whose activity these approaches generally weight most in their decoding decisions. For the searchlight analysis, the remaining 10 percent of values indicate those brain regions in which the searchlight analysis achieved the highest decoding accuracy.

All analysis approaches correctly associated activity in the upper parts of the middle and inferior temporal gyrus with body stimuli. The GLM, whole-brain lasso and DeepLight also correctly associated activity in the right fusiform gyrus with body stimuli. Only DeepLight correctly associated activity in the postcentral gyrus with these stimuli. The GLM, whole-brain lasso and DeepLight further all correctly associated activity in the right FFA with face stimuli. None of the approaches, however, associated activity in the left FFA with face stimuli. Interestingly, the searchlight analysis did not associate the FFA with face stimuli at all. All analysis approaches also correctly associated activity in the PPA with place stimuli. Lastly, for tool stimuli, the GLM and whole-brain lasso correctly associated activity in the left inferior temporal sulcus with stimuli of this class. The searchlight analysis and whole-brain lasso only did so marginally. None of the approaches associated activity in the left postcentral gyrus with tool stimuli.

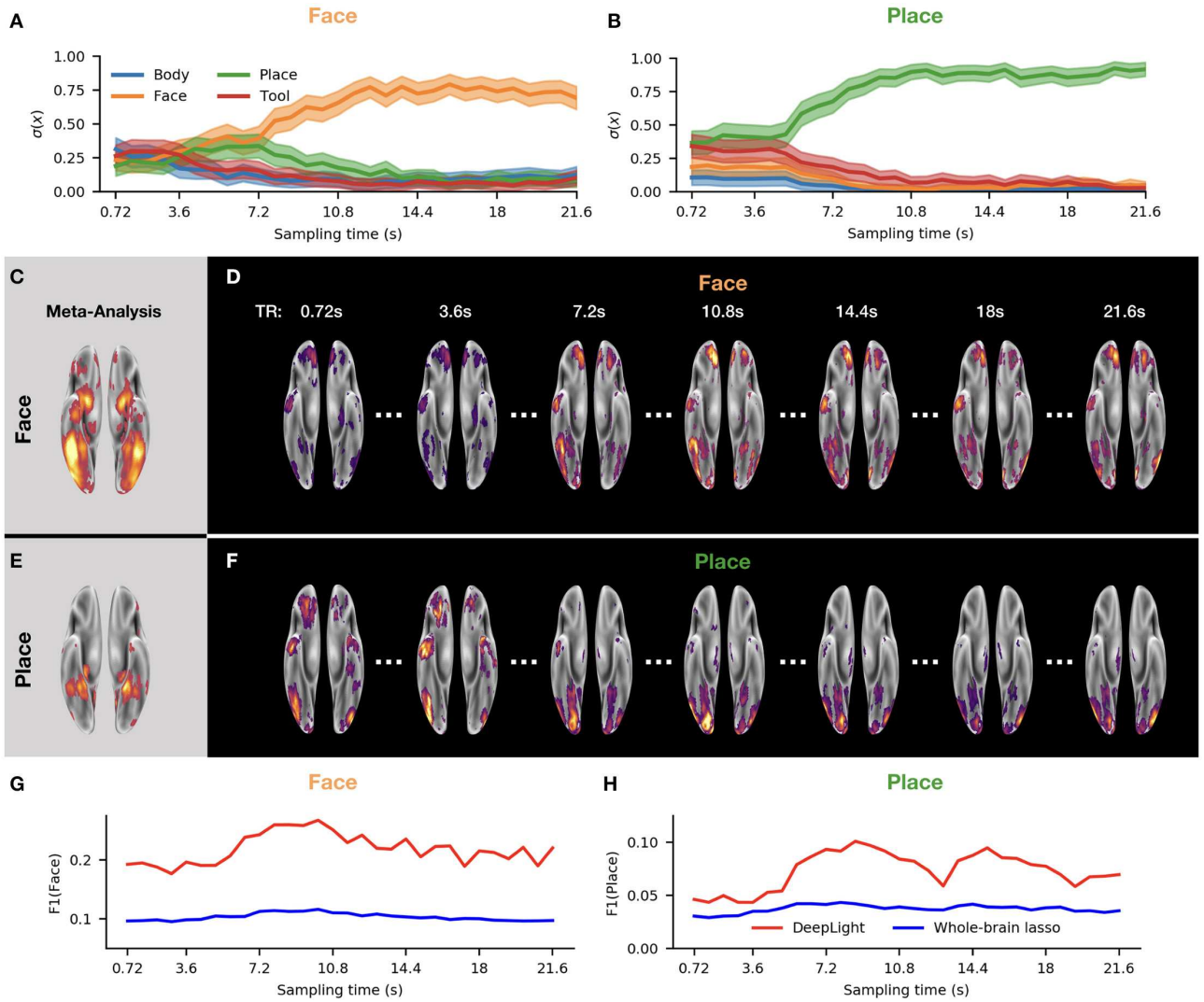


FIGURE 6 | DeepLight analysis of the temporo-spatial distribution of brain activity in the first experiment block of the face and place stimulus classes in the second experiment run of the held-out test dataset. **(A,B)** Average predicted probability that the fMRI data collected at each sampling time point belongs to each of the four cognitive states. **(C,E)** Results of a meta-analysis with the NeuroSynth database for the face and place stimulus classes (for details on the meta-analysis, see section NeuroSynth in **Supplementary Information**). **(D,F)** Group-level brain maps for seven fMRI sampling time points from the experiment block. Each group-level brain map at each time point is computed as an average over the relevance maps of each subject for this time point. Each group-level brain map is thresholded at the 90th percentile of its values. All brain maps are projected onto the inflated cortical surface of the FsAverage5 surface template (Fischl, 2012). **(G,H)** F1-score for each group-level brain map at each sampling time point of the experiment block. The F1-score quantifies the similarity between the group-level brain map and the results of the meta-analysis **(C,E)** (for further details on the F1-score, see section Subject-Level and section NeuroSynth in **Supplementary Information**). Red indicates the results of the F1-score comparison for the brain maps of DeepLight, whereas blue indicates the results of this comparison for the brain maps of the whole-brain lasso analysis (for further details on the F1-comparison for the whole-brain lasso analysis, see section DeepLight's Relevance Patterns Resemble Temporo-Spatial Variability of Brain Activity Over Sequences of Single fMRI Samples).

Overall, DeepLight's group-level brain maps accurately associated each of the ROIs with their respective cognitive states. Interestingly, DeepLight also associated a set of additional brain regions with the face and tool stimulus classes that were not identified by the other analysis approaches (see **Figure 4E**). For face stimuli, these regions are the orbitofrontal cortex and temporal pole. While the temporal pole has been shown to be involved in the ability of an individual to infer the desires, intentions and beliefs of others (*theory-of-mind*; for a detailed review, see Olson et al., 2007), the orbitofrontal cortex has been associated with the processing of emotions in the faces of others (for a detailed review, see Adolphs, 2002). For tool stimuli, DeepLight additionally utilized the activity of the temporoparietal junction (TPJ) to decode these stimuli. The TPJ has been shown to be associated with the ability of an individual to discriminate self-produced actions and the actions produced

by others and is generally regarded of as a central hub for the integration of body-related information (for a detailed review, see Decety and Grèzes, 2006). Although it is not clear why only DeepLight associated these brain regions with the face and tool stimulus classes, their assumed functional roles do not contradict this association.

Subject-Level

The goal of the subject-level analysis was to test the ability of each analysis approach to identify the physiologically appropriate associations between brain activity and cognitive state on the level of each individual.

To quantify the similarity between the subject-level brain maps and the results of the meta-analysis, we defined a similarity measure. Given a target brain map (e.g., the results of our meta-analysis), this measure tests for each voxel in the brain whether a source brain map (e.g., the results of our subject-level analyses) correctly associates this voxel's activity with the cognitive state (true positive), falsely associates the voxel's activity with the cognitive state (false positives) or falsely does not associate the voxel's activity with the cognitive state (false negatives). Particularly, we derived this measure from the well-known F1-score in machine learning (see section F1-Score in **Supplementary Information** as well as Goutte and Gaussier, 2005). The benefit of the F1-score, when compared to simply computing the ratio of correctly classified voxels in the brain, is that it specifically considers the brain map's precision and recall and is thereby robust to the overall size of the ROIs in the target brain map. Here, precision describes the fraction of true positives from the total number of voxels that are associated with a cognitive state in the source brain map. Recall, on the other hand, describes the fraction of true positives from the overall number of voxels that are associated with a cognitive state in the target brain map. Generally, an F1-score of 1 indicates that the brain map has both, perfect precision and recall with respect to the target, whereas the F1-score is worst at 0.

To obtain an F1-score for each subject-level brain map (for details on the estimation of subject-level brain maps with the three baseline analysis approaches, see section Parameter Estimation of the Baseline Methods in **Supplementary Information**), we again thresholded each individual brain map. For the GLM, we defined all voxels with $P > 0.005$ (uncorrected) as not associated with the cognitive state and all others as associated with the cognitive state. For the searchlight analysis, whole-brain lasso and DeepLight, we defined all voxels with a value below the 90th percentile of the values within the brain map as not associated with the cognitive state and all others as associated with the cognitive state.

Overall, DeepLight's subject-level brain maps had meaningfully larger F1-scores for the body, face and place stimulus classes, when compared to those of the GLM [$t_{(29)} = 10.46$, $p < 0.0001$ for body stimuli, **Figure S5A**; $t_{(29)} = 13.04$, $p < 0.0001$ for face stimuli, **Figure S5D**; $t_{(29)} = 9.26$, $p < 0.0001$ for place stimuli, **Figure S5G**], searchlight analysis [$t_{(29)} = 13.26$, $p < 0.0001$ for body stimuli, **Figure S5B**; $t_{(29)} = 8.57$, $p < 0.0001$ for face stimuli, **Figure S5E**; $t_{(29)} = 4.25$, $p = 0.0002$, for place stimuli, **Figure S5H**], and whole-brain lasso [$t_{(29)} = 20.93$, $p < 0.0001$ for body stimuli, **Figure S5C**; $t_{(29)} = 48.32$, $p < 0.0001$ for face stimuli, **Figure S5F**; $t_{(29)} = 22.43$, $p < 0.0001$, for place stimuli, **Figure S5I**]. For tool stimuli, the GLM and searchlight generally achieved higher subject-level F1-scores than DeepLight [$t_{(29)} = -8.19$, $p < 0.0001$, **Figure S5J**; $t_{(29)} = -4.39$, $p = 0.0001$, **Figure S5K** for the GLM and searchlight, respectively], whereas DeepLight outperformed the whole-brain lasso analysis [$t_{(29)} = 18.31$, $p < 0.0001$, **Figure S5L**].

To ascertain that the results of this comparison were not dependent on the thresholds that we chose, we replicated the comparison for each combination of the 85th, 90th, and 95th percentile threshold for the brain maps of the searchlight analysis, whole-brain lasso and DeepLight, as well as a P-threshold of 0.05, 0.005, 0.0005, and 0.00005 for the brain maps of the GLM. Within all combinations of percentile values and P-thresholds, the presented results of the F1-comparison were generally stable (see **Tables S3–S6**).

DeepLight Accurately Identifies Physiologically Appropriate Associations Between Cognitive States and Brain Activity on Multiple Levels of Data Granularity

DeepLight's ability to correctly identify the physiological appropriate associations between cognitive states and brain activity is exemplified in **Figure 5**. Here, the distribution of relevance values for the four cognitive states is visualized on three different levels of data granularity of an exemplar subject (namely, the subject with the highest decoding accuracy in **Figures 3A,B**): First, on the level of the overall distribution of relevance values of each cognitive state of this subject (**Figure 5A**; incorporating an average of 47 TRs per cognitive state), then on the level of the first experiment block of each cognitive state in the first experiment run (**Figure 5B**; incorporating an average of 12 TRs per cognitive state) and lastly on the level of a single brain sample of each cognitive state (**Figure 5C**; incorporating a single TR per cognitive state).

On all three levels, DeepLight utilized the activity of a similar set of brain regions to identify each of the four cognitive states. Importantly, these regions largely overlap with those identified by the DeepLight group-level analysis (**Figure 4E**) as well as the results of the meta-analysis (**Figure 4A**).

DeepLight's Relevance Patterns Resemble Temporo-Spatial Variability of Brain Activity Over Sequences of Single fMRI Samples

To further probe DeepLight's ability to analyze single time points, we next studied the distribution of relevance values over the course of a single experiment block (**Figure 6**). In particular, we plotted this distribution as a function of the fMRI sampling-time over all subjects for the first experiment block of the face and place stimulus classes in the second experiment run. We restricted this analysis to the face and place stimulus classes, as the neural networks involved in processing face and place stimuli, respectively, have been widely characterized (see, for example Haxby et al., 2001 as well as Heekeren et al.,

2004). For a more detailed overview, we also created two videos for the two experiment blocks depicted in **Figure 6** (**Supplementary Videos 1**, **2**). These videos display the temporal evolution of relevance values for each fMRI sample in the original fMRI sampling time of the face (**Supplementary Video 1**) and place (**Supplementary Video 2**) experiment blocks.

In the beginning of the experiment block, DeepLight was generally uncertain which cognitive state the observed brain samples belonged to, as it assigned similar probabilities to each of the cognitive states considered (**Figures 6A,B**). As time progressed, however, DeepLight's certainty increased and it correctly identified the cognitive state underlying the fMRI samples. At the same time, it started assigning more relevance to the target ROIs of the face and place stimulus classes (**Figures 6C–F**), as indicated by the increasing F1-scores resulting from a comparison of the brain maps at each sampling time point with the results of the meta-analysis (**Figures 6G,H**; all brain maps were again thresholded at the 90th percentile for this comparison). Interestingly, the relevances started peaking in the target ROIs 5s after the onset of the experiment block. The temporal evolution of the relevances thereby mimics the hemodynamic response measured by the fMRI (Lindquist et al., 2009).

To further evaluate the results of this analysis, we replicated it by the use of the whole-brain lasso group-level decoding model (see section Baseline Methods and section Parameter Estimation of the Baseline Methods in **Supplementary Information**). In particular, we multiplied the fMRI samples of all test subjects collected at each sampling time point with the coefficient estimates of the whole-brain lasso group-level model. Subsequently, we averaged the resulting weighted fMRI samples within each sampling time point depicted in **Figures 6G,H** and computed an F1-score for a comparison of the resulting average brain maps with the results of the meta-analysis (as described in section Subject-Level). Interestingly, we found that the F1-scores of the whole-brain lasso analysis varied much less over the sequence of fMRI samples and were throughout lower than those of DeepLight. Thereby, indicating that the brain maps of the whole-brain lasso analysis exhibit comparably little variability over the course of an experiment block with respect to the target ROIs defined for the face and place stimulus classes.

DISCUSSION

Neuroimaging data have a complex temporo-spatial dependency structure that renders modeling and decoding of experimental data a challenging endeavor. With DeepLight, we propose a new data-driven framework for the analysis and interpretation of whole-brain neuroimaging data that scales well to large datasets and is mathematically non-linear, while still maintaining interpretability of the data. To decode a cognitive state, DeepLight separates a whole-brain fMRI volume into its axial slices and processes the resulting sequence of brain slices by the use of a convolutional feature extractor and LSTM. Thereby, accounting for the spatially distributed patterns of whole-brain brain activity within and across axial slices. Subsequently, DeepLight relates cognitive state and brain activity, by decomposing its decoding decisions into the contributions of the single input voxels to these decisions with the LRP method. Thus, DeepLight is able to study the associations between brain activity and cognitive state on multiple levels of data granularity, from the level of the group down to the level of single subjects, trials and time points.

To demonstrate the versatility of DeepLight, we have applied it to an openly available fMRI dataset of 100 subjects viewing images of body parts, faces, places, and tools. With these data, we have shown that the DeepLight (1) decodes the underlying cognitive states more accurately from the fMRI data than conventional means of uni- and multivariate brain decoding, (2) improves its decoding performance better with growing datasets, (3) accurately identifies the physiologically appropriate associations between cognitive states and brain activity, (4) can study these associations on multiple levels of data granularity, from the level of the group down to the level of single subjects, trials and time points, and (5) can capture the temporo-spatial variability of brain activity over sequences of single fMRI samples.

Transferring DeepLight to Other fMRI Datasets

The DeepLight architecture used here is exemplary. Future research is needed to evaluate how the specific architectural choices for its three sub-modules (the convolutional feature extractor, LSTM unit and softmax output layer; see section DeepLight Framework) will effect its performance. In the following, we will briefly outline how the proposed architecture can be transferred to the analysis of other fMRI datasets with different spatial resolution and decoding targets. Importantly, online minimal changes are necessary in order to adapt DeepLight's architecture for the analysis of such fMRI datasets.

DeepLight first processes an fMRI volume within each axial slice, by computing a higher-level, and lower-dimensional, representation of the slices with the convolutional feature extractor. Here, the spatial sensitivity of DeepLight to the fine-grained activity differences of neighboring voxels within each slice is determined by the stride size applied by the convolution layers. The stride size indicates the distance between the application of the convolution kernels to the axial slices of the fMRI volume. Generally, a larger stride decreases DeepLight's sensitivity for fine-grained differences in the activity of neighboring voxels, as it increases the distance between the applications of the convolution kernels to the input slice. Reversely, a smaller stride size increases DeepLight's sensitivity for the fine-grained activity differences of neighboring voxels, as it decreases the distance between the applications of the convolution kernels. For example, when analyzing fMRI volumes that have a lower spatial resolution than the ones used here, containing fewer voxels per axial slice (and thereby less information about the distribution of brain activity within each slice), we would recommend to decrease the stride size for more of DeepLight's convolution

layers, in order to best leverage the information contained in these voxels.

After the application of the convolutional feature extractor, DeepLight integrates the information of the resulting higher-level slice representations, by the use of a bi-directional LSTM. Here, each of the two LSTM units iterates through the entire sequence of slice representations, before forwarding its output. The proposed DeepLight architecture therefore does not require any modification in order to accommodate fMRI datasets with a different number of axial slices per volume, as it generalizes to any sequence length.

Further, the number of neurons in the softmax output layer is directly determined by the number of decoding targets considered in the data (one output neuron per decoding target). In the case of a continuous decoding target (for example, by predicting a subject's score in a cognitive test), the softmax output layer can be replaced with a linear regression layer. The LRP decomposition approach (see section Layer-Wise Relevance Propagation in the DeepLight Framework) also applies to continuous output variables (for further details on the application of the LRP approach to continuous output variables, see Bach et al., 2015 and Montavon et al., 2017).

Lastly, recent exploratory empirical work has shown that even for more complex fMRI decoding analyses, encompassing up to 400 subjects and 20 distinct cognitive states (see Thomas et al., 2019), DeepLight does not require more than 64 neurons per layer. We would therefore not recommend to increase the number of neurons further, as this will also lead to an overall increased risk of overfitting.

Comparison to Baseline Methods

General Linear Model

The GLM is conceptually different from the other neuroimaging analysis approaches considered in this work. It aims to identify an association between cognitive state and brain activity, by modeling (or predicting) the time series signal of a single voxel as a linear combination of a set of experiment predictors (see section Baseline Methods). It is thereby limited in three meaningful ways that do not apply to DeepLight: First, the time series signal of a voxel is generally very noisy. The GLM treats each voxel's signal as independent of one another, thereby, not leveraging the evidence that is shared across the time series signal of multiple voxels. Second, even though the linear combination of a set of experiment predictors might be able to explain variance in the observed fMRI data, it does not necessarily provide evidence that this exact set of predictors is encoded in the neuronal response. Generally, the same linear model (in terms of its predictions) can be constructed from many different (even random) sets of predictors (for a detailed discussion of this "feature fallacy," see Kriegeskorte and Douglas, 2018). The results of the GLM analysis thereby indicate that the measured neuronal response is highly structured and that this structure is preserved across individuals, whereas the labels assigned to its predictors might be arbitrary. Third, the performance of the GLM in predicting the response signal of a voxel is typically not evaluated on independent data, which leaves unanswered how well its results generalize to new data.

Searchlight Analysis

DeepLight generally outperformed the searchlight analysis in decoding the cognitive states from the fMRI data. In small datasets (here, the data of 10 or less subjects), however, the performance of the searchlight analysis was superior. In contrast to DeepLight, the searchlight analysis decodes a cognitive state from single cluster of only few voxels. Its input data, as well as the number of parameters in its decoding model, are thereby considerably smaller, leading to an overall lower risk of overfitting. Yet, this advantage comes at the cost of additional constraints that have to be considered when considering both approaches. If a cognitive state is associated with the activity of a small brain region only, the searchlight analysis will generally be more sensitive to the activity of this region than DeepLight, as it has learned a decoding model that is specific to the activity of the region. If, however, the cognitive state is not identifiable by the activity of a single brain region only, but solely in conjunction with the activity of another spatially distinct brain region, the searchlight analysis will not be able to identify this association, due to its narrow spatial focus. DeepLight, on the other hand, will generally be less sensitive to the specifics of the activity of a local brain region, but perform better in identifying a cognitive state from spatially wide-spread brain activity. When choosing between both approaches, one should therefore consider whether the assumed associations between brain activity and cognitive state specifically involve the activity of a local brain region only, or whether the cognitive state is associated with the activity of spatially distinct brain regions.

Whole-Brain Lasso

In contrast to DeepLight, the whole-brain lasso analysis is based on a linear decoding model. It assigns a single coefficient weight to each voxel in the brain and makes a decoding decision by computing a weighted sum over the activity of an input fMRI volume. Importantly, due to the strong regularization that is applied to the coefficients during the training, many coefficients equal 0. The resulting set of coefficients thereby resembles a brain mask, defining a set of fixed brain regions whose activity the whole-brain lasso utilizes to decode a cognitive state. DeepLight, on the other hand, utilizes a hierarchical structure of non-linear transforms of the fMRI data. It projects each fMRI volume into a more abstracted, higher-level space. This abstracted (and more flexible) view enables DeepLight to better account for the variable patterns of brain activity underlying a cognitive state (within and across individuals). This ability is exemplified in **Figure 6**, as well as **Supplementary Videos 1**, **2**, where we visualize the variable patterns of brain activity that DeepLight associates with the face and place stimulus classes throughout an experiment block. The relevance patterns of DeepLight mimic the hemodynamic response and peak in the ROIs 5–10 s after the onset of the experiment block. Importantly, we find that the whole-brain lasso does not exhibit such temporo-spatial variability.

Disentangling Temporally Distinct Associations Between Cognitive State and Brain Activity

DeepLight's ability to identify a cognitive state through variable patterns of brain activity makes it ideally suited for the analysis

of the fine-grained spatial distribution of brain activity over temporal sequences of fMRI samples. For example, Hunt and Hayden (2017) recently raised the question whether the neural networks involved in reward-based decision making can be subdivided into a set of spatially distinct and temporally discrete network components, or whether the underlying networks act in parallel, with highly recurrent activity patterns. Answering this question is difficult with conventional approaches to the analysis of neuroimaging data, such as the baseline methods included in this paper. These often learn a fixed mapping between brain activity and cognitive state, by aggregating over the information provided by a sequence of fMRI samples (e.g., by estimating a single coefficient weight for each voxel from a sequence of fMRI data). The resulting brain maps thereby only indicate whether there exist spatially distinct brain regions that are associated with a cognitive state, without providing any insight whether the activity patterns are temporally discrete. While these methods can be adapted to specifically account for the temporal differences in the activity patterns of these regions (e.g., by analyzing different time points independent of one another), these adaptations often require specific hypotheses about the studied temporal differences (e.g., by needing to specify the different time points to analyze). DeepLight, on the other hand, operates purely data-driven and is thereby able to autonomously identify an association between spatially distinct patterns of brain activity and a cognitive state at temporally discrete time points.

Integrative Analysis of Multimodal Neuroimaging Data

DeepLight is not bound to fMRI data, but can be easily extended to other neuroimaging modalities. One such complementary modality, with a higher temporal, but lower spatial resolution, is the Electroencephalography (EEG). While a plethora of analysis approaches have been proposed for the integrative analysis of EEG and fMRI data, these often incorporate restrictive assumptions to enable the integrative statistical analysis of these two data types, with clearly distinct spatial, temporal and physiological properties (for a detailed review, see Jorge et al., 2014). DeepLight, on the other hand, represents a data-driven analysis framework. By providing both, EEG and fMRI data as separate inputs to the DL model, DeepLight could learn the fine-grained temporal structure of brain activity from the EEG data, while utilizing the fMRI data to localize the spatial brain regions underlying this activity. Recently, researchers have already demonstrated the usefulness of interpretable DL methods for the analysis of EEG data (Sturm et al., 2016).

Extending DeepLight

Lastly, we would like to highlight several possible extensions of the DeepLight approach, resulting from its flexible and modular architecture. First, DeepLight can be extended to specifically account for the temporo-spatial distribution of brain activity over sequences of fMRI samples, by the addition of another recurrent network layer. This layer would process each of the higher-level whole-brain representations resulting from the currently proposed architecture. This extension would enable DeepLight to more specifically account for the temporal distribution of brain activity. Second, DeepLight can be extended to the integrative analysis of neuroimaging data from multiple cognitive tasks and experiments. For example, by adding one neuron to the output layer for each cognitive state from each task. This extension would enable a more thorough analysis of the differences (and similarities) between the associations of cognitive state and brain activity across multiple tasks and experiments.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The scanning protocol involving human participants were reviewed and approved by Washington University in St. Louis's Human Research Protection Office (HRPO), IRB# 201204036. No experimental activity involving the human subjects took place at the authors' institutions. The patients/participants provided their written informed consent to participate in this study. Only de-identified, publicly released data were used in this study.

AUTHOR CONTRIBUTIONS

AT, HH, K-RM, and WS conceived of DeepLight. AT, K-RM, and WS planned all data analyses. AT implemented all visualizations of DeepLight and the experimental procedures, performed all formal data analyses, wrote all software that was used in the data analyses and that is underlying DeepLight, and wrote the original draft of the manuscript. HH, K-RM, and WS reviewed and edited the manuscript. This work was supervised by HH, K-RM, and WS.

FUNDING

This work was supported by the German Federal Ministry for Education and Research through the Berlin Big Data Centre (01IS14013A), the Berlin Center for Machine Learning (01IS18037I) and the TraMeExCo project (01IS18056A). Partial funding by the German Research Foundation (DFG) is acknowledged (EXC 2046/1, project-ID: 390685689). K-RM was also supported by the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at arXiv:1810.09945 (see Thomas et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01321/full#supplementary-material

Supplementary Video 1 | DeepLight analysis of the temporo-spatial distribution of brain activity in the first experiment block of the face stimulus class in the second experiment run of the held-out test dataset. **(Right)** DeepLight's group-level brain maps for each sampling time point of the first experiment block of the face stimulus class in the second experiment run of the held-out test dataset. All brain maps are projected onto the inflated cortical surface of the FsAverage5 surface template. **(Left)** For each sampling time point, three metrics are displayed: DeepLight's softmax output prediction for each decoding target, DeepLight's average decoding accuracy, and the F1-score of a comparison of the shown group-level brain map with the results of a NeuroSynth meta-analysis for the term "face"' (for further details on the F1-score, see section Subject-Level and section NeuroSynth in **Supplementary Information**). Higher F1-scores generally indicate stronger similarity.

Supplementary Video 2 | DeepLight analysis of the temporo-spatial distribution of brain activity in the first experiment block of the place stimulus class in the second experiment run of the held-out test dataset. **(Right)** DeepLight's group-level brain maps for each sampling time point of the first experiment block of the place stimulus class in the second experiment run of the held-out test dataset. All brain maps are projected onto the inflated cortical surface of the FsAverage5 surface template. **(Left)** For each sampling time point, three metrics are displayed: DeepLight's softmax output prediction for each decoding target, DeepLight's average decoding accuracy, and the F1-score of a comparison of the shown group-level brain map with the results of a NeuroSynth meta-analysis for the term "place"' (for further details on the F1-score, see sections Subject-Level and NeuroSynth in **Supplementary Information**). Higher F1-scores generally indicate stronger similarity.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *OSDI* (Savannah, GA) Vol. 16, 265–283.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8:14. doi: 10.3389/fninf.2014.00014

Adolphs, R. (2002). Neural systems for recognizing emotion. *Curr. Opin. Neurobiol.* 12, 169–177. doi: 10.1016/S0959-4388(02)00301-X

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)* (Copenhagen: Association for Computational Linguistics), 159–168.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage* 80, 169–189. doi: 10.1016/j.neuroimage.2013.05.033

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., and Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* 13:e1002180. doi: 10.1371/journal.pbio.1002180

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Decety, J., and Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Res.* 1079, 4–14. doi: 10.1016/j.brainres.2005.12.115

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. (2015). "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 2625–2634.

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15, 870–878. doi: 10.1006/nimg.2001.1037

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Sardinia), 249–256.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*, Vol. 1. Cambridge: MIT press.

Goutte, C., and Gaussier, E. (2005). "A probabilistic interpretation of pre- cision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval* (Santiago de Compostela: Springer), 345–359.

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). "Identifying predictive regions from fmri with tv-l1 prior," in *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (Washington, DC: IEEE), 17–20.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., and Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with graphnet. *Neuroimage* 72, 304–321. doi: 10.1016/j.neuroimage.2012.12.062

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature* 431:859. doi: 10.1038/nature02966

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hunt, L. T., and Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nat. Rev. Neurosci.* 18:172. doi: 10.1038/nrn.2017.7

Jang, H., Plis, S. M., Calhoun, V. D., and Lee, J.-H. (2017). Task-specific feature extraction and classification of fmri volumes using a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks. *Neuroimage* 145, 314–328. doi: 10.1016/j.neuroimage.2016.04.003

Jorge, J., Van der Zwaag, W., and Figueiredo, P. (2014). Eeg–fmri integration for the study of human brain function. *Neuroimage* 102, 24–34. doi: 10.1016/j.neuroimage.2013.05.114

Kriegeskorte, N., and Douglas, P. K. (2018). Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* 55, 167–179. doi: 10.1016/j.conb.2019.04.002

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., and Samek, W. (2016). The layer-wise relevance propagation toolbox for artificial neural networks. *J. Mach. Learn. Res.* 17, 1–5. Available online at: http://www.jmlr.org/papers/v17/15-618.html

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* 10:1096. doi: 10.1038/s41467-019-08987-4

LeCun, Y., Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, eds M. A. Arbib (Cambridge, MA: MIT Press), 3361.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *Neuroimage* 45, S187–S198. doi: 10.1016/j.neuroimage.2008.10.065

Marban, A., Srinivasan, V., Samek, W., Fernandez, J., and Casals, A. (2019). A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery. *Biomed. Signal Process. Control*. 50, 134–150. doi: 10.1016/j.bspc.2019.01.011

McIntosh, A. R., and Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage* 23, S250–S263. doi: 10.1016/j.neuroimage.2004.07.020

McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1325–1334.

Mensch, A., Mairal, J., Thirion, B., and Varoquaux, G. (2018). Extracting universal representations of cognition across brain-imaging studies. *arXiv preprint arXiv:*1809.06035.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *IEEE Trans. Med. Imaging* 30, 1328–1340. doi: 10.1109/TMI.2011.2113378

Montavon, G., Bach, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.* 65, 211–222. doi: 10.1016/j.patcog.2016.11.008

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transac. Neural Netw.* 12, 181–201. doi: 10.1109/72.914517

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Nie, D., Zhang, H., Adeli, E., Liu, L., and Shen, D. (2016). "3d deep learning for multi- modal imaging-guided survival time prediction of brain tumor patients," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 212–220.

Olson, I. R., Plotzker, A., and Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain* 130, 1718–1731. doi: 10.1093/brain/awm052

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning* (Atlanta, GA), 1310–1318.

Petrov, D., Kuznetsov, B. A., van Erp, T. G., Turner, J. A., Schmaal, L., Veltman, D., et al. (2018). "Deep learning for quality control of subcortical brain 3d shape models," in International Workshop on Shape in Medical Imaging, ShapeMI 2018 held in conjunction with 21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018 (Granada: Springer Verlag), 268–276.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fmri data: the openfmri project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323:533. doi: 10.1038/323533a0

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fmri data. *Neuroimage* 51, 752–764. doi: 10.1016/j.neuroimage.2010.02.040

Samek, W., Wiegand, T., and Müller, K.-R. (2018). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *ITU J. ICT Discov*. 1, 39–48.

Sarraf, S., and Tofighi, G. (2016). Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:*1603.08631.

Schölkopf, B., and Smola, A. J. (2002). *Learning With Kernels: Support Vector Machines, Regularization, Optimization*, and *Beyond*. Cambridge, MA: MIT press.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958. Available online at: http://jmlr.org/papers/v15/srivastava14a.html

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R. (2016). Interpretable deep neural networks for single-trial eeg classification. *J. Neurosci. Methods*. 274, 141–145. doi: 10.1016/j.jneumeth.2016.10.008

Suk, H.-I., Lee, S.-W., Shen, D., and the Alzheimers Disease Neuroimaging Initiative. (2014). Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *Neuroimage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Thomas, A. W., Heekeren, H. R., Müller, K.-R., and Samek, W. (2018). Analyzing neuroimaging data through recurrent deep learning models. *arXiv preprint arXiv:1810.09945.*

Thomas, A. W., Müller, K. R., and Samek, W. (2019). "Deep transfer learning for whole-brain FMRI analyses," in *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, eds L. Zhou, D. Sarikaya, S. M. Kia, S. Speidel, A. Malpani, D. Hashimoto, M. Habes, T. Löfstedt, K. Ritter, and H.Wang (Cham: Springer), 59–67.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tikhonov, A. N. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* 39, 195–198.

Uğurbil, K., Xu, J., Auerbach, E. J., Moeller, S., Vu, A. T., Duarte-Carvajalino, J. M., et al. (2013). Pushing spatial and temporal resolution for functional and diffusion mri in the human connectome project. *Neuroimage* 80, 80–104. doi: 10.1016/j.neuroimage.2013.05.012

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fmri-based neurologic signature of physical pain. *N. Eng. J. Med.* 368, 1388–1397. doi: 10.1056/NEJMoa1204471

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8:665. doi: 10.1038/nmeth.1635

Yousefnezhad, M., and Zhang, D. (2018). Anatomical pattern analysis for decoding visual stimuli in human brains. *Cognit. Comput.* 10, 284–295. doi: 10.1007/s12559-017-9518-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Thomas, Heekeren, Müller and Samek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fnins.2019.01313

Using Deep Learning and Resting-State fMRI to Classify Chronic Pain Conditions

Alex Novaes Santana[1], Ignacio Cifre[2], Charles Novaes de Santana[3] and Pedro Montoya[1]**

[1] Research Institute of Health Sciences (IUNICS-IdISBa), University of the Balearic Islands, Palma, Spain, [2] Facultat de Psicologia, Ciències de l'Educació i de l'Esport, Blanquerna, Universitat Ramon Llull, Barcelona, Spain, [3] Landscape Ecology, Institute of Terrestrial Ecosystems, ETH Zürich, Zurich, Switzerland

Chronic pain is known as a complex disease due to its comorbidities with other symptoms and the lack of effective treatments. As a consequence, chronic pain seems to be under-diagnosed in more than 75% of patients. At the same time, the advance in brain imaging, the popularization of machine learning techniques and the development of new diagnostic tools based on these technologies have shown that these tools could be an option in supporting decision-making of healthcare professionals. In this study, we computed functional brain connectivity using resting-state fMRI data from one hundred and fifty participants to assess the performance of different machine learning models, including deep learning (DL) neural networks in classifying chronic pain patients and pain-free controls. The best result was obtained by training a convolutional neural network fed with data preprocessed using the MSDL probabilistic atlas and using the dynamic time warping (DTW) as connectivity measure. DL models had a better performance compared to other less costly models such as support vector machine (SVM) and RFC, with balanced accuracy ranged from 69 to 86%, while the area under the curve (ROC) ranged from 0.84 to 0.93. Also, DTW overperformed correlation as connectivity measure. These findings support the notion that resting-state fMRI data could be used as a potential biomarker of chronic pain conditions.

Keywords: chronic pain, machine learning, classification, rs-fMRI, deep-learning, DTW

OPEN ACCESS

Edited by:
Reza Lashgari, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Daniel Lopez Martinez, Google Brain Health, United States
Daniel Callan, National Institute of Information and Communications Technology (Japan), Japan

****Correspondence:***
Alex Novaes Santana
alex.santana@gmail.com
Pedro Montoya
pedro.montoya@uib.es

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *14 August 2019*
Accepted: *25 November 2019*
Published: *17 December 2019*

Citation:
Santana AN, Cifre I, de Santana CN and Montoya P (2019) Using Deep Learning and Resting-State fMRI to Classify Chronic Pain Conditions. Front. Neurosci. 13:1313. doi: 10.3389/fnins.2019.01313

INTRODUCTION

Pain is a subjective psychological phenomenon that emerges from brain activity but can be influenced by several and different aspects of human behavior and cognition (attention, learning, beliefs, etc.) (Albe-Fessar et al., 1985; Montoya et al., 2004). Indeed, pain is a complex problem in which biological, psychological, and social factors may play a relevant role in its maintenance over time (Bevers et al., 2016). This multidimensional aspect of pain requires that measurement of pain should include not only subjective ratings, but also psychological and neurophysiological events related to pain perception. Accordingly, research on cognitive pain has been conducted using several approaches and tools such as questionnaires (Pagé et al., 2015; Traeger et al., 2016), Quantitative Sensory Tests (QST) (Cruz-Almeida and Fillingim, 2014; Cámara et al., 2016), genetic factors (Diatchenko et al., 2005; Ablin and Buskila, 2015; Ultsch et al., 2016),

patterns of physical activity (Hidalgo et al., 2012, 2014; Paraschiv-Ionescu et al., 2012), EEG (Pinheiro et al., 2016), neuroimaging (Davis et al., 2012; Schmidt-Wilcke, 2015), and more recently functional near-infrared spectroscopy (fNIRS) (Lopez-Martinez et al., 2018, 2019).

Chronic pain is characterized by symptoms such as pain that lasts more than 3–6 months (Merskey et al., 1994; Wolfe et al., 2016), as well as by fatigue, sleep disturbance, cognitive and mood changes (Gatchel et al., 2010). As a result, chronic pain may limit individual daily activities, leading to disability and reduced quality of life. Several studies have shown that chronic pain syndromes are also associated with alterations in the functional connectivity of BOLD signals (Baliki et al., 2008). For instance, chronic back pain (CBP) patients seem to have reduced deactivations in specific brain regions of the default mode network (DMN), such as mPFC, amygdala, and PCC. Furthermore, studies in patients with neuropathic pain (NP) have confirmed reduced DMN connectivity together with enhanced resting-state functional connectivity among several pain related areas (Cauda et al., 2009). These findings are in line with previous data suggesting that these alterations could be the neurophysiological mechanisms involved in cognitive and behavioral impairments associated with chronic pain (Apkarian et al., 2004; Cauda et al., 2009). The analyses of resting-state fMRI have suggested that the brain of chronic pain patients differ from that of healthy subjects by showing significant reductions of functional connectivity of the DMN, together with significant enhancements of several pain-related areas. Pain-related brain structures also presented significant changes in functional connectivity when comparing fibromyalgia (FM) patients and healthy controls (Cifre et al., 2012). FM patients show changes in regions involved in facilitating and reducing pain processing. Other similar studies using fMRI also found alterations in brain connectivity for patients with chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS) (Johnson et al., 2015; Lin et al., 2017), migraine (Lovati et al., 2016), and other (Davis and Moayedi, 2013; Napadow and Harris, 2014; Truini et al., 2016; Zamorano et al., 2017). The intensity of these brain alterations is also correlated with pain intensity reported by CP patients (O'Shea et al., 2015). Also, patterns of functional brain connectivity have been widely investigated as a potential biomarker for classification and prediction of a variety of other neurological and psychiatric syndromes such as attention deficit hyperactivity disorder (ADHD), schizophrenia, and Mild Cognitive Impairment (Du et al., 2018).

Despite recent findings about neurophysiological mechanisms (central sensitization, brain plasticity), chronic pain remains under-diagnosed and under-treated. In some cases, more than 75% of the patients do not receive an accurate diagnosis (Kress et al., 2015; Dodick et al., 2016). Furthermore, the assessment of knowledge, attitudes, and beliefs (KAB) about chronic pain among primary care providers revealed that even those who participate in continuous education programs still may have inappropriate beliefs and attitudes about pain and its treatment (Lalonde et al., 2014). Therefore, one possible solution would be to provide clinicians with objective methods to support their decision about pain diagnosis and treatment. In this regard, a multidimensional framework and operational diagnostic criteria for the major chronic pain conditions were proposed by the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION), the US Food and Drug Administration, and the American Pain Society (APS) (Dworkin et al., 2016). This framework is divided into five dimensions: (1) Core Diagnostic Criteria, (2) Common Features, (3) Common Medical and Psychiatric Comorbidities, (4) Neurobiological, Psychosocial, and Functional Consequences, and finally (5) Putative Mechanisms, Risk Factors, and Protective Factors. Following this framework, other studies proposed evidence-based diagnostic criteria for specific chronic pain conditions (Dampier et al., 2017; Paice et al., 2017; Widerström-Noga et al., 2017; Zhou et al., 2018; Arnold et al., 2019; Freeman et al., 2019; Ohrbach and Dworkin, 2019). The majority of these criteria are composed by the patient historical data, self-reported information via questionnaires, and psychophysical tests that determine pain features and pain thresholds. Although pain neuroimaging was not presented in these criteria, it has been used in clinical trials with a focus on diagnostic properties of different conditions, including FM and chronic back pain (Smith et al., 2017).

Due to the complexity and the great number of features, the patterns of functional brain connectivity patterns are commonly analyzed using multivariate analysis such as Support Vector Machines (SVM), Logistic Regression (LR), and Least Absolute Shrinkage and Selection Operator (LASSO). In fact, the majority of the models mentioned in Du et al. (2018) made use of SVM. Deep learning (DL) comprehends a family of machine learning algorithms that use a set of processing layers to extract and transform features from data with multiple levels of abstraction. These algorithms have improved the image and speech recognition, compared with traditional machine learning algorithms (e.g., SVM and LASSO) (Krizhevsky et al., 2012; LeCun et al., 2015). This performance puts DL models in focus as a promising approach to classify brain images, especially for single subject prediction (Vieira et al., 2017). Furthermore, some DL architectures have been specially designed to learn from brain functional connectivity networks (Kawahara et al., 2017; Meszlényi et al., 2017). In the context of chronic pain, functional brain images were also used as an information source to multivariate pattern analysis in the attempt to classify or predict chronic pain syndromes (Callan et al., 2014; Sundermann et al., 2014; López-Solà et al., 2017). In these works, only traditional models such as SVM, LR, and k-Nearest Neighbors (k-NN) were trained to differentiate healthy control subjects from patients with FM (Sundermann et al., 2014; López-Solà et al., 2017) and from chronic back pain (CBP) (Callan et al., 2014). This restraint to traditional algorithms motivated us to apply DL to chronic pain classification. Moreover, the majority of such previous studies used a certain type of stimulus during the acquiring process, which requires a more adequate environment and a team of technicians to be executed. When a resting-state protocol was applied, the results had a lower performance with accuracy indices below 80%. Also, most of the

studies focused on identifying only one chronic pain syndrome (usually CBP or FM).

Our objective in this work was to evaluate the performance of a set of DL algorithms in the classification problem of chronic pain syndromes and to compare it with the performance of traditional classifiers. Moreover, we will analyze how different brain parcelations and connectivity measures affect the classification performance. In order to achieve that, we analyzed the data using four different parcelations including ROI and group-ICA based parcelations, and two different measures of functional brain connectivity, such as correlations and Dynamic Time Warping.

MATERIALS AND METHODS

Participants

The participants of the study were ninety-eight healthy controls (age: 40.85 ± 23.7) and sixty chronic pain patients (age: 45.65 ± 15.23). Sixty-four females (age: 39.33 ± 20.5) and thirty-four males (age: 43.76 ± 23.89) participants compose the healthy control group. A limitation of this study is that those male participants were presented only on the control group. The chronic pain group (CP) was composed of thirty-six FM patients and twenty-four chronic back pain patients. All chronic pain participants suffered from persistent pain for more than 6 months, also they were diagnosed following the IASP's criteria (Merskey et al., 1994) and (Wolfe et al., 2016) for FM. There was no significant difference in age between the groups (T = 1.56, p = 0.11). No participant used opiates, gabapentin, or pregabalin for pain treatment. Four patients occasionally used non-steroidal anti-inflammatory drugs (NSAIDs) and/or paracetamol. Medication for non-pain related disorders involved in birth control and female hormonal drugs. Three CP took benzodiazepine (1–5 mg per day), of which one also took serotonin reuptake inhibitors.

Image Acquisition

BOLD resting-state functional magnetic resonance images (rsfMRI) from one hundred and fifty-three subjects were used in this work. All data were collected using a GE 3T scanner (General Electric Signa HDx, GE Healthcare, Milwaukee, WI, United States). A total of 240 whole-brain echo-planar images were acquired in a time of 10 min with the following scanner parameters: 32 slices per volume, 3 mm of slice thickness, 4 mm of space between slices, repetition time equals 2500 ms, echo time equals 35 ms, 90° flip angle, matrix dimensions equal 64 × 64, and FOV equals 240. The structural images were collected using three similar protocols. These protocols differed exclusively by the number of slices per volume; where the values of 292, 220, 256, and 248 slices were used for 24, 44, 48, and 37 of all dataset, respectively. The other parameters were configured as follows: 1 mm of slice thickness, repetition time equals 7796 ms, echo time equals 2984 ms, 12° flip angle, matrix dimensions equal 256 × 256, and FOV equals 256. Scanner noise was passively reduced by −36 dB using in-ear hearing protection. In addition, MRI foam-cushions were placed over the ears to restrict head motion and further to reduce the impact of scanner noise.

Image Preprocessing

Image preprocessing was performed using the Neuroimaging in Python Pipelines and Interfaces (Nipype) (Gorgolewski et al., 2011). The first five image volumes were excluded prior to image preprocessing. Then, spikes were removed using an algorithm from the Analysis of Functional NeuroImages (AFNI)[1] software suite. Subsequently, the processes of slice-time correction, realigning, and coregistration were performed using the Statistical Parametric Mapping software package version 12 (SPM12). Also, the time-course signal-to-noise ratio (TSNR) of each time series was calculated together with a Nipype's algorithm to detect artifacts based on RapidArt[2]. This algorithm uses intensity and motion parameters to infer outliers. Finally, all images were normalized to the standard Montreal Neurological Institute (MNI) stereotactic space with a bias regularization of 10^{-4} and a FWHM of Gaussian smoothness of bias of 60, resulting in a voxel size of 2 mm.

Brain Functional Parcelation

We tested four different parcelation maps in order to identify how the number of parcels affects the classification process. Two parcelations were part of the UK Biobank Imaging Study[3], where about five thousands resting-state functional MRI data points were collected from different participants. Both parcelations were the result of a group independent components analysis (group-ICA) (Miller et al., 2016). These parcelations were composed of 25 and 100 components. After the authors excluded the components that are not neuronally driven, the parcelations presented 21 and 55 components, respectively. Only this final sets of 21 and 55 components were used in our study.

The third parcelation was obtained from an analysis called Multi-Subject Dictionary Learning (MSDL) using resting-state functional MRI from 20 subjects with eyes closed (Varoquaux et al., 2011). The final map presented a parcelation with 39 components. The fourth and last parcelation was the Willard functional atlas, containing 499 functional regions of interest (Richiardi et al., 2015). All brain parcelations can be visualized in **Figure 1**.

Extracting Connectivity Measures

The brain functional network of each participant was determined using the suite Nilearn (Abraham et al., 2014). This process followed a pipeline of three tasks. The first task was to clean and to exclude bad images: images that presented more than 3 "bad" slices detected by the artifacts detection algorithm were removed from the dataset. For the images remained, the time-series were detrended; the movement confounds were removed based on a projection on the orthogonal of the signal space (Friston et al., 1994; Lindquist et al., 2019); then, the images were standardized. The second task was to extract the brain activity time-series from the brain regions defined by the parcelation map. For each region, the least-squares solution was calculated, and a new time-series representing the region was obtained.

[1]https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dDespike.html

[2]http://www.nitrc.org/projects/rapidart/

[3]https://www.fmrib.ox.ac.uk/ukbiobank/

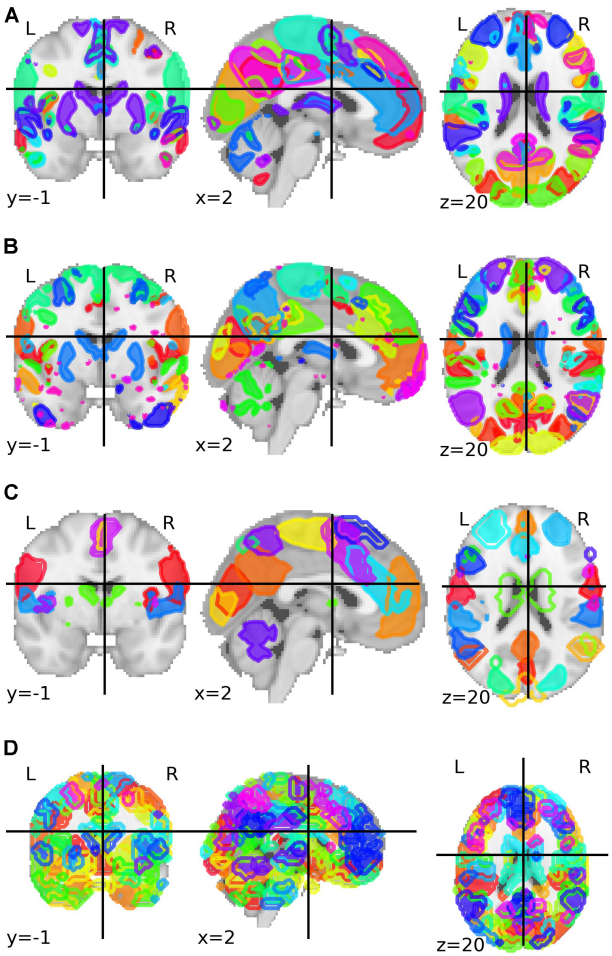


FIGURE 1 | Group-ICA based parcelation from Biobank UK with **(A)** 21 and **(B)** 55 components after the authors excluded the components that are not neuronally driven. **(C)** Parcelation with 39 parcels and based on a multi-subject dictionary learning process (MSDL). **(D)** Willard functional atlas with 499 regions of interest.

In the last task, a functional network was constructed using the brain parcels as nodes. The connectivity between each parcel was calculated using the representative time-series of each parcel, pair-wisely. Correlation and dynamic time warping distance (DTW) were used to measure the connectivity value between these time-series. DTW is a similarity measure for time-series, where one time-series can be "compressed" or "stretched" in time in order to find the best alignment with another time-series. It was first designed for speech recognition (Sakoe and Chiba, 1978) and was significantly more accurate than other similarity measures (Wang et al., 2013). In our experiments, the DTW was calculated using the FastDTW algorithm (Salvador and Chan, 2007). Finally, after calculating the connectivity matrix, we normalized it. We calculated the z-score of each matrix cell in regards with all cells. At this point, each image was converted into eight matrices, one for each combination of four parcelations (Biobank_UK_25, Biobank_UK_100, MSDL, and Willard) with two measures (Correlation and DTW).

Classifiers

Four classifiers were compared to determine the one that presented the best results in a binary classification problem: chronic pain patients against controls. Three of the four classifiers used convolution neural network approaches. The classifier BrainNetCNN was defined by Kawahara and collaborators (Kawahara et al., 2017) and it proposes three new conventional filters. Those filters are adapted to adjacent matrices that represent any kind of neural network. The edge-to-edge filter computes a weighted response over neighboring edges for a given edge, while the edge-to-node filter computes a weighted response over neighboring edges for a given node. The third filter, called node-to-graph, applies a one-dimensional convolution to calculate a scalar based on a weighted combination of the nodes. Also, this architecture includes three fully connected hidden layers, characterizing it as a deep neural network.

The second classifier was created by modifying the BrainNetCNN. Batch Normalization layers were inserted between the BrainNetCNN layers to create a new classifier, called BrainNetCNNBatch. This classifier normalizes the activations of the previous layer at each batch, improving the performance and stability, and allowing to reach higher learning ratios (LRs) (Ioffe and Szegedy, 2015). In a similar work, Meszlényi et al. (2017) presented a convolutional neural network architecture to classify rsfMRI. The presented architecture was our third classifier, which had a sequence of two one-dimensional convolutional layers followed by a fully connected layer and a softmax layer with two outputs. See **Supplementary Figure 1** for a detailed scheme of these convolution neural networks.

The fourth classifier was an automated machine learning toolkit called TPOT[4]. This classifier uses genetic programming to find an optimized machine-learning pipeline (Olson et al., 2016a,b). This toolkit allowed us to test different classical machine learning models and feature engineering processes with a reduced computational cost.

Training and Evaluation

Before the training, the missing data were dropped, which together with the previous exclusion criteria of a maximum of 3 "bad" slices per image resulted in a dataset of 140 participants. Next, the dataset was split into training and testing dataset, keeping the original proportion of healthy controls and chronic pain patients. Training and testing datasets had 98 and 42 participants, respectively. Due to this reduced number of participants, we opted to augment the training data by creating five thousand Local Synthetic Instances (LSI) (Brown et al., 2015). This split 98/42 is the independent fold selection. These 42 subjects were separated in the beginning, before any data training or feature engineering. On the other hand, the 98 subjects from training dataset were split in the same proportion, if the model is a CCNN we did a single split (train/validation), otherwise (TPOT case) we applied a k-fold cross-validation approach. All scores presented in this paper are related to the independent fold. For CCNN, we opted to use only one split (train/validation/test) because of the computation cost. Also,

[4]http://epistasislab.github.io/tpot/

during the data augmentation, the training and validation data were augmented separately to avoid any leak of information between the two sets.

Overfitting is a common problem during the development of a new machine learning model. It directly affects the model's generalization, in other words, the ability to predict new scenarios (inputs) correctly. Overfitting can be detected by evaluating the model error during training and test. When the difference between training and test errors increase, it is a sign that the model is starting to overfit the training dataset. One approach to avoid this scenario is to use regularizers. Regularization is defined as any modification to a learning algorithm that is intended to reduce its test error but not its training error (Goodfellow et al., 2016, p. 177). Three kinds of regularizers are commonly used in neural network-based classifiers: Dropout, L1, and L2 regularizers.

In this work, different combinations of Dropout, L1, and L2 regularizers were tested. Dropout had values of [0, 0.3, 0.5, 0.8], these values represent the probability of a node be excluded from the neural network. This behavior generates sub-networks that are trained which permits we apply inexpensive bagged ensemble methods with neural networks. L1 and L2 are also called Lasso regression and Ridge regression, respectively. These two regularizers add a penalization term to the model's loss function. To control how this term affects the loss function, a constant λ is multiplied to it. For λ we had values of [0, 0.1, and 0.01]. When gamma assumes a value of zero, this regularizer is deactivated. The range values of dropout and λ for L1 and L2 were chosen arbitrarily. Also, when not specified in the original architecture description, the regularization was applied between the fully connected layers.

The LR was another parameter that was optimized and tested with values of [0.1, 0.01, and 0.001]. Thus, each combination of parcelation, connectivity measure (Correlation and DTW), neural network architecture, learning ration, and regularizers created a classification experiment, combining into a total of 2592 classification experiments. These experiments were trained using the five thousand synthetic instances and evaluated using forty-two participants from the testing set.

For the experiments using TPOT, we applied a five stratified k-fold cross-validation process. Also, the genetic algorithm was configured to retain 50 individuals in every generation while running it for ten generations. Three scores were used as metrics to be optimized by TPOT: balanced accuracy, the area under the receiver operating curve (AUC), and log loss. Combining these three metrics with the eight connectivity matrices, we executed 24 classification experiments using TPOT. **Figure 2** describes the entire process of acquisition, preprocessing, processing, learning, and evaluation.

RESULTS

In order to evaluate and compare the experiments, we used three scores: balanced accuracy; the area under the receiver operating curve (AUC); and cross-entropy loss (log loss). Balanced accuracy is defined as the average of *recall* obtained on each class, which in turn is the proportion of actual *positives* that are predicted as *positives*. Imbalance groups do not affect this accuracy score. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). For both balanced accuracy and AUC, the values can range between 0 and 1; values close to 1 indicate better classifiers, and the scores are calculated using the predicted classes. On the other hand, log loss can range from 0 to $+\infty$, where values close to 0 are better scores. It is calculated using the probability of an instance to belong to a target class. It is defined as the negative log-likelihood of the true labels given a probabilistic classifier's predictions. The log loss can be interpreted as a measure of *certainty*, where a classifier that predicts a wrong class with a high probability is punished (Bishop, 2006).

Correlation vs. Dynamic Time Warping

Analyzing all 2616 experiments, the first observed result was that DTW outperformed correlation in balanced accuracy and AUC (**Figure 3**). This result was found for the average score, as well for the best scores of each group. This result is in accordance with (Meszlényi et al., 2016). For the log loss, the difference between experiments using DTW and Correlation was not significant (**Table 1**). Thus, all the subsequent results will only summarize the experiments using DTW.

Parcelation Size vs. Performance

Examining the performance of the brain parcelations, we observed that the number of parcels had no significant relation with the values of accuracy, AUC, and log loss (**Figure 4**). However, the parcelation seems to affect the performance of classifier separately. We can observe that both Ann4Brains architectures decrease (slightly) the performance with bigger parcelations. Meanwhile, the architecture proposed by Meszlényi et al. (2017) improves its performance with the increment of parcels. TPOT seems to be stable in relation to the parcelation size. Once TPOT finds the best pipeline in a set of different types of classifiers and/or dimensionality reduction techniques, it can adapt to the parcelation size.

Best Classifier for Chronic Pain

In **Figure 5**, we have a box plot with results of all experiments grouped by parcelation and architecture. TPOT clearly shows the best mean values for all metrics and parcelation. Since TPOT results are only for the best pipeline found by the genetic algorithm, higher mean (or lower in log loss case) was expected. Among the DL architectures, the Ann4Brains based architectures had better results compared with the architecture proposed by Meszlényi et al. (2017).

The best accuracy of classifying chronic pain was obtained using Ann4brains architecture and its variation with batch normalization (see **Supplementary Figure 2**). These classifiers exhibited accuracy values of 0.868. In both cases, the MSDL parcelation was used. Ann4brainsBatch also presented the best results for AUC and log loss, with values of 0.935 and 0.344, respectively. In that case, the BiobankUK_25 parcelation was used. In **Table 2**, we summarized the best results obtained by

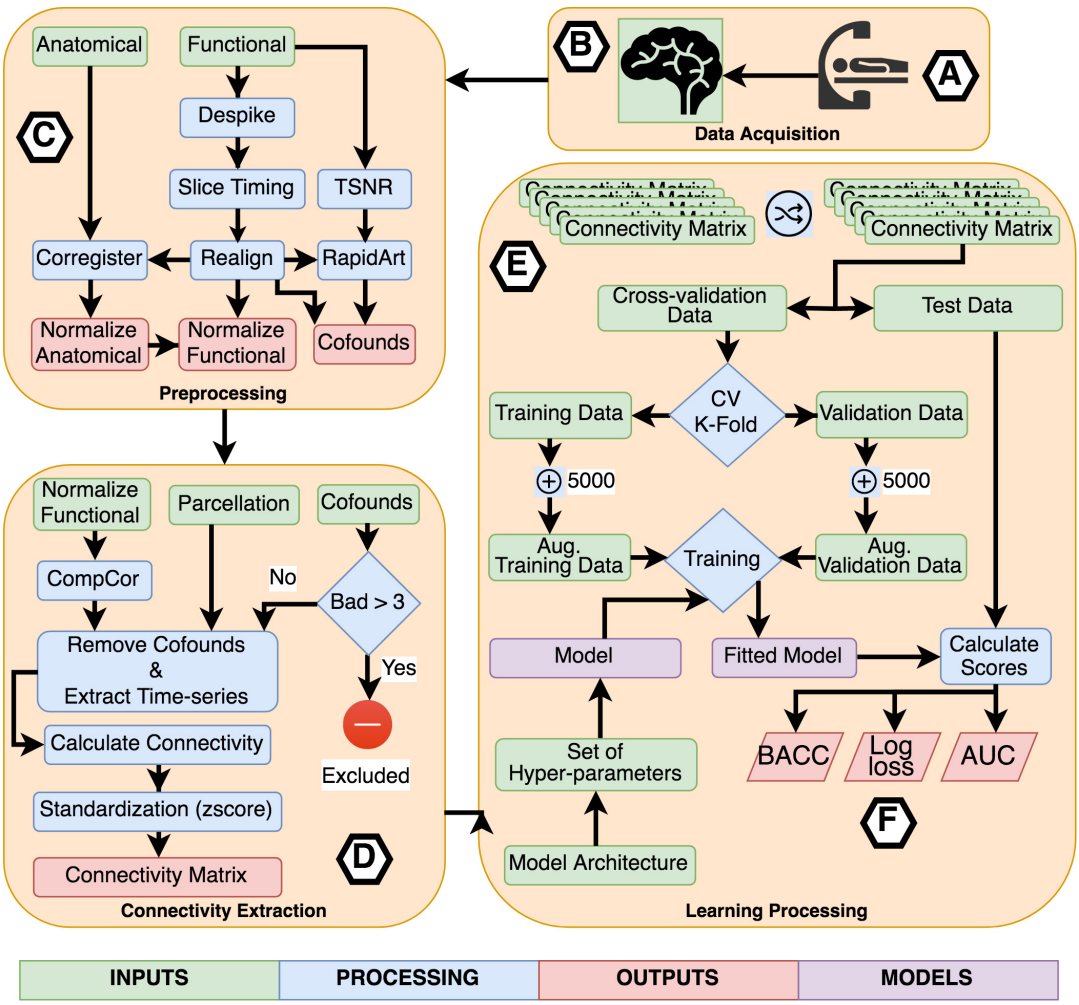


FIGURE 2 | Flowchart describing the entire process of acquisition, preprocessing, processing, learning, and evaluation. **(A)** Participants were scanned in a resting-state protocol and **(B)** functional and structural images were collected. **(C)** Each subject's images were preprocessed separately applying standard procedures like time-slicing, realigning, coregistration, artifacts detection, and normalization. **(D)** A connectivity matrix was created for each subject combining a normalized functional image, a set of confounds, a parcelation, and a connectivity measure. **(E)** The learning processing starts shuffling the list of subject's connectivity matrix, we preserved the shuffling index to be replicated with the other inputs set combining the different parcelations and connectivity measures. The next step was to select the independent fold used to compare the different fit models. The remainder data was used in the cross-validation process. All models were trained using a k-fold cross-validation scheme with $k = 2$ to CNN based models and $k = 5$ for TOPT models. Inside the cross-validation process, from each testing and validation's dataset, five thousand synthetic connectivity matrices were created. These synthetic data were used to train and validate the model, which is a combination of the model architecture and a specific set of hyper-parameters, including regularization. For the last, **(F)** the fitted model was evaluated using the independent fold and three scores: Balanced Accuracy (BACC), Log loss, and the Area Under the receiver-operating characteristics curve. These scores were used to compare the performance of architectures, parcelations, and connectivity measures.

each classifier and parcelation. All these results were obtained using the DTW as a connectivity measure. The experiments using correlation had the best accuracy of 0.826 and the best AUC of 0.905 using TPOT and BiobankUK_100, which reinforce the results found in see section "Correlation vs. Dynamic Time Warping."

A limitation of this study is that the chronic pain group presented only female participants. To exclude the possibility that our models were learning to classify the groups according to participants' sex rather than to the presence or not of chronic pain, all models were assessed by using sex as the target variable.

Supposing that the models were classifying the groups based on participants' sex. It would be expected that the models had a better performance if evaluate using the sex label as the classification target. Evaluating this possibility using the independent fold, the mean result for balanced accuracy, log loss, and AUC were 0.67, 3.76, and 0.73, respectively. More specifically, the best model using DTW, MSDL, and Ann4BrainsBatch had, respectively, a balanced accuracy, a log loss, and AUC of 0.70, 1.69, and 0.78 using the sex label. Comparing with the chronic pain prediction, the higher values of log loss is an indication that if the models were learning the sex difference it would be

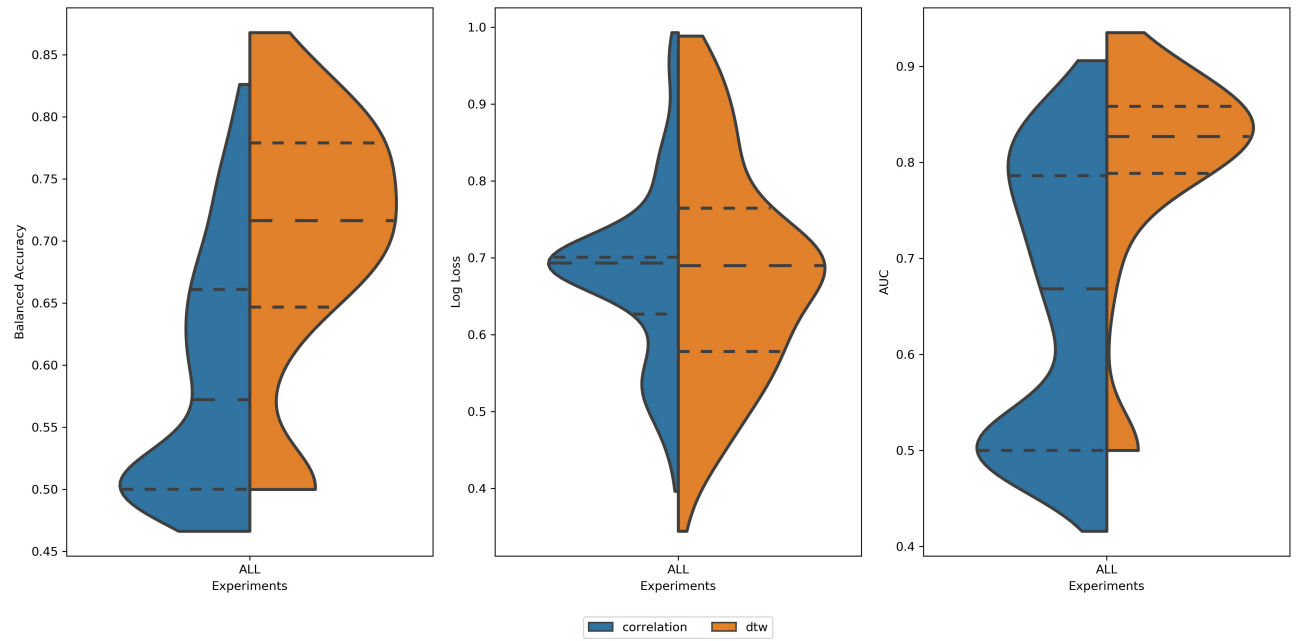


FIGURE 3 | Distribution of values of **(left)** Balanced Accuracy, **(center)** Log Loss, and **(right)** AUC for all 2616 experiments using correlation (blue area) and DTW (orange area). All tails were truncated in maximum and minimum values while the dash lines delimit the quartiles. In the case of Balanced Accuracy and AUC, the distributions for Correlation and DTW differ with p-value < 0.05.

TABLE 1 | Mean values of balanced accuracy, log loss, and AUC for experiments using correlation and DTW as a measure of connectivity.

	Balanced accuracy	Log loss	AUC
Correlation	0.62 ± 0.08	0.61 ± 0.10	0.73 ± 0.12
DTW	0.69 ± 0.09	0.60 ± 0.13	0.80 ± 0.9
t-index (p-value)	−1.89 (3.19e−66)	−0.95 (3.39e−01)	−16.3 (1.01e−51)

predicting the wrong classes with high probabilities. In other words, the models are uncertain about the sex label. These results gave us the confidence that our models were not affected by the sex group limitation.

DISCUSSION

In this paper, we presented a classifier for chronic pain conditions using resting-state fMRI and one convolutional neural network architecture. This classifier, with a DL approach, outperformed traditional machine learning techniques. Also, in the search to find the best classifier, we performed a set of experiments to understand how parcelation and connectivity measures affect the classification process. These experiments revealed that while the parcelation does not affect significatively the classification, the use of DTW significantly increase the classifier performance.

The findings of this study clearly show that functional brain images in association with DL can be used to differentiate chronic pain patients from pain-free controls. The best model using the Ann4brain architecture and MSDL parcelation had a balanced accuracy of 86.8% while the AUC was equal to 0.918. In the literature, other attempts to classify chronic pain using brain images have been made. Chronic low back pain (Baliki et al., 2011; Callan et al., 2014; Ung et al., 2014), and fibromyalgia (Sundermann et al., 2014; Robinson et al., 2015; López-Solà et al., 2017) are the most common syndromes that were studied, while temporomandibular disorder (Harper et al., 2016) and knee osteoarthritis (Baliki et al., 2011) are also present in previous studies.

A tool for classifying individuals with chronic back pain (CBP) was proposed by Callan and collaborators (Callan et al., 2014). They used voxel-level differences from 13 CBP patients and 13 pain-free controls during periods of both resting and electrical stimulation. By using Sparse Logistic Regression (SLR), they reported an accuracy of 92.3%. A similar accuracy percentage (93%) has been achieved by using a combination of three classifiers based on SVM and LR in 37 FM patients and 35 pain-free controls when painful stimuli were applied (López-Solà et al., 2017). Moreover, an SVM classifier was able to accurately classify 10 patients with myofascial-type temporomandibular disorders and 10 matched pain-free controls during painful pressure stimulation (Harper et al., 2016). Compared to previous studies, the present work focused on the comparison of different methods to classify participants into chronic pain patients and healthy controls based exclusively on functional resting-state data. Furthermore, we were able to analyze the results from several classifiers for two different types of chronic pain conditions: FM and chronic back pain. Sundermann et al. (2014) used only resting-state images to identify potential functional connectivity among patients with FM, rheumatoid arthritis (RA), and pain-free controls. The author tried different types of classifiers based

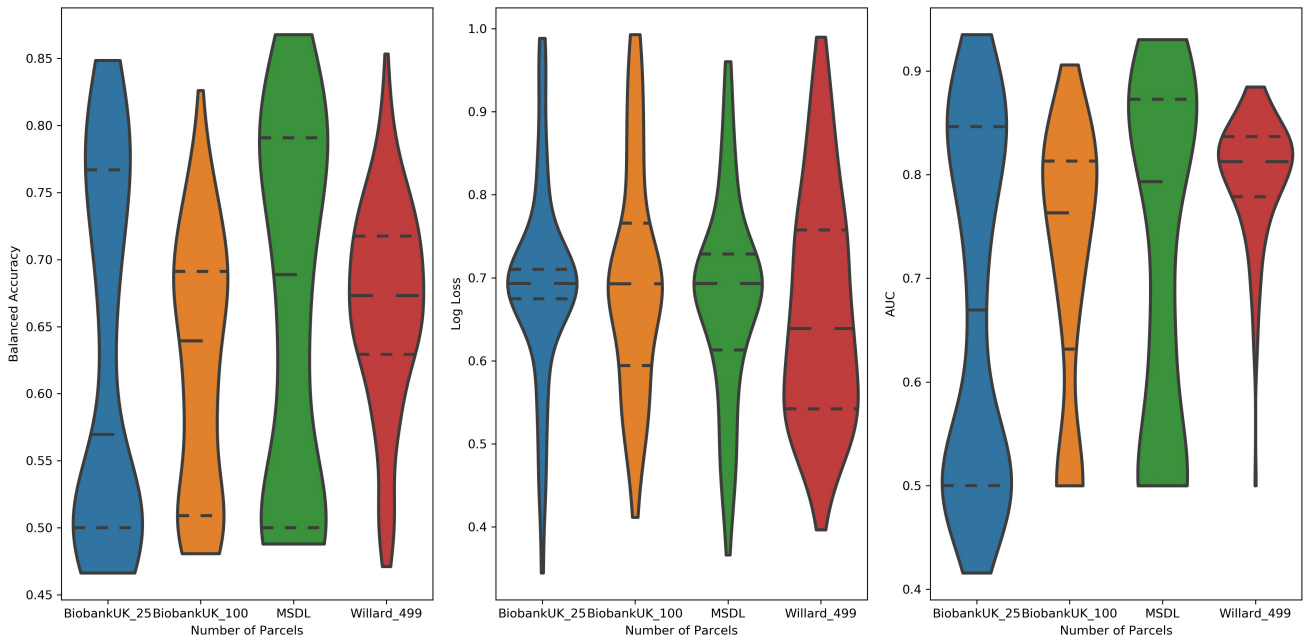


FIGURE 4 | Distribution of values of **(left)** Balanced Accuracy, **(center)** Log Loss, and **(right)** AUC grouped by parcelation size: Biobank_UK_25 (blue), Biobank_UK_100 (Orange), MSDL (green), and Willard (red). All tails were truncated in maximum and minimum values while the dash lines delimit the quartiles.

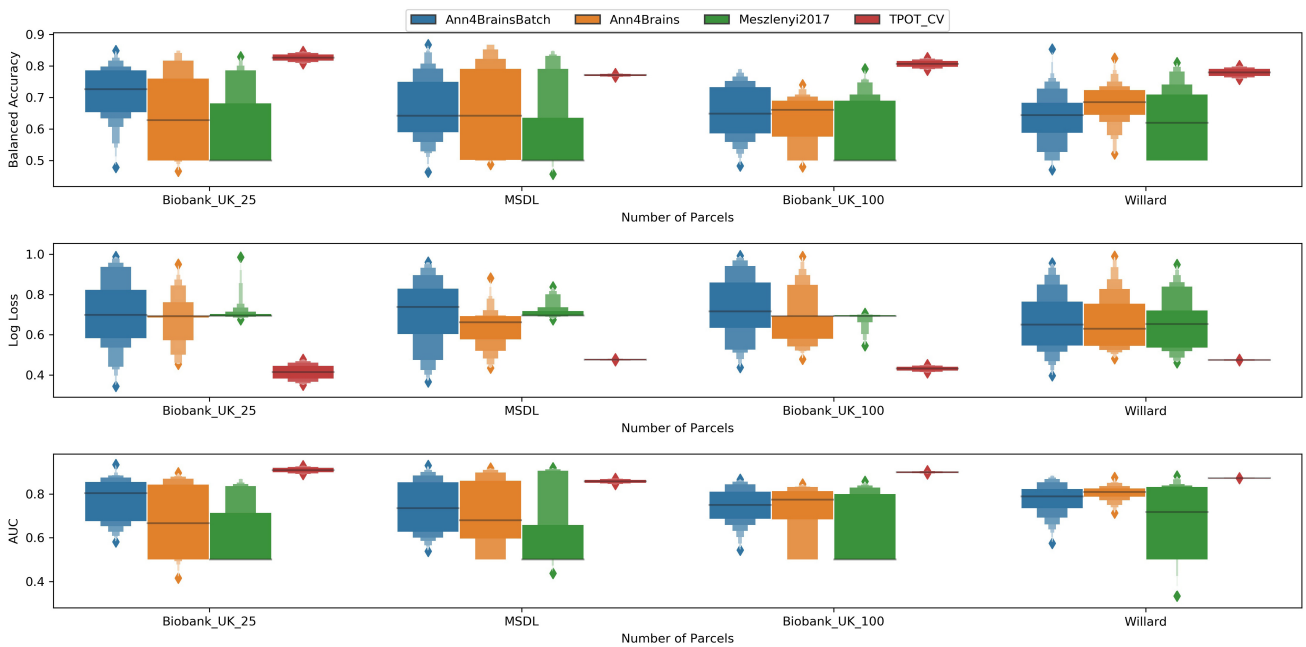


FIGURE 5 | Enhanced box plot of values of Balanced Accuracy **(top)**, Log Loss **(middle)**, and AUC **(bottom)** grouped by parcelation size (horizontal axis) and classifier architecture: Ann4BrainsBatch (blue), Ann4Brains (orange), Meszlenyi2017 (green), and TPOT (red).

on SVM, decision tree, naive Bayes classifiers, etc. Comparing FM with RA and using a k-NN based classifier the best accuracy of 79% was reached. Meanwhile, the best accuracy comparing FM and pain-free controls only archived 73.5% with an SVM based classifier. Comparing with our work, the main difference is the absence of DL algorithms in the previous study. But, even so, if compared with our results using TPOT, our work outperforms the results found by Sundermann et al. (2014). Similarly, studies that used only structural images to classify chronic pain had a lower performance (Baliki et al., 2011; Ung et al., 2014;

TABLE 2 | Best values of balanced accuracy, log loss, and AUC for each combination of classifier and parcelation using DTW.

	Biobank_UK_25	MSDL	Biobank_UK_100	Willard
Ann4Brains	BACC = 0.849	BACC = 0.868	BACC = 0.740	BACC = 0.825
	LLOSS = 0.453	LLOSS = 0.435	LLOSS = 0.479	LLOSS = 0.482
	AUC = 0.897	AUC = 0.918	AUC = 0.846	AUC = 0.875
Ann4BrainsBatch	BACC = 0.849	BACC = 0.868	BACC = 0.779	BACC = 0.853
	LLOSS = 0.345	LLOSS = 0.366	LLOSS = 0.438	LLOSS = 0.396
	AUC = 0.935	AUC = 0.930	AUC = 0.868	AUC = 0.885
Meszlenyi2017	BACC = 0.805	BACC = 0.849	BACC = 0.678	BACC = 0.810
	LLOSS = 0.676	LLOSS = 0.675	LLOSS = 0.546	LLOSS = 0.462
	AUC = 0.853	AUC = 0.916	AUC = 0.760	AUC = 0.882
TPOT	BACC = 0.846	BACC = 0.776	BACC = 0.826	BACC = 0.801
	LLOSS = 0.351	LLOSS = 0.472	LLOSS = 0.412	LLOSS = 0.471
	AUC = 0.929	AUC = 0.872	AUC = 0.906	AUC = 0.874

Robinson et al., 2015). The best result obtained by Baliki et al. (2011) was a balanced accuracy of 81.25% in a multi-class problem involving patients with CBP, complex regional pain syndrome (CPRS), Osteoarthritis (OA), and pain-free controls.

Another important finding of this study was that DTW outperforms Correlation when applied to measure the connectivity among brain areas. This result matches the results found by Meszlényi et al. (2016), where the authors demonstrated that DTW emphasizes group differences resulting in a better classification. Meszlényi et al. (2016) used a unique brain parcelation with 90 ROIs (Shirer et al., 2012), and two classifiers based on SVM and the LASSO. In this scenario, DTW and correlation were evaluated in two tasks: classify the correct gender and identify ADHD patients. Our analysis expands the work proposed by Meszlényi et al. (2016), showing that DTW outperforms Correlation in different scenarios composed of four types of parcelations. Moreover, with the aid of TPOT toolkit, we tested both connectivity measures for a variety of classifiers, including linear and non-linear approaches. The tests demonstrated that DTW had better results than Correlation.

The interpretability of neural networks is a challenge due to its "black-box" characteristics. Some efforts like the Lime tool tries to make an approximation of the neural network behavior and provide some interpretation about how the neural networks do its prediction. Unfortunately, tools like that are adapted to networks that use a typical convolutional filter with a 3×3 or 5×5 matrix convolving. These typical convolutional networks are specialized to identify edges, contrast, border, etc. In our case, our image is, in reality, a connectivity matrix and the neighborhood of a cell does not represent any relationship, requiring specialized convolutional filters. Because of that, we cannot make use of tools like Lime to interpret the prediction. In contrast, we can apply Lime to interpret the results from TPOT.

In **Figure 6**, we can see three cases of prediction: (6-a) HC correctly predicted, (6-b) CP correctly predicted, and (6-c) a wrong prediction. The features on top represent the elements that most contributed to this prediction. In the cases (6-a) and (6-b), we can notice that these top features have a connection involving the DMN or Insula, which suggests that the model is classifying the chronic pain mainly based on the connectivity of these two areas. This behavior is supported by other studies that relate these two areas with chronic pain syndromes (Baliki et al., 2008; Cauda et al., 2009). Analyzing the wrong prediction (6-c) we also can notice the presence of the DMN and Insula on the top features. But, in that case, the model predicted this HC subject as a CP subject with a lower probability of approximately 0.53. This number close to 0.5 means that the model was not confident about this prediction. Based on this interpretation, we can conclude that the model is making the predictions based on areas of the brain related to chronic pain.

A clear innovation of our work was the classification of chronic pain syndromes using neuroimaging and DL techniques. There are no records in the literature of any successful classification of chronic pain syndromes using a similar approach. This approach was already applied for other syndromes like attention-deficit and hyperactivity disorder, mild cognitive impairment, schizophrenia, Alzheimer's disease (AD), and others (Du et al., 2018). Also, some previous studies compared DL techniques with a kernel-based model like SVM classifiers (Vieira et al., 2017). The majority of the studies show that when the overfitting problem is controlled, DL techniques outperform SVM classifiers. This behavior was also observed in our experiments, where the DL classifier (Ann4BrainsBatch) had the highest accuracy and AUC. Due to the small number of samples, the major challenge was to control the overfitting. To facilitate the process, L2 regularization (0.01) and a dropout rate of 0.5 were used between the hidden layers. Despite that, based on our results, we can conclude that DL techniques can provide good results in the task of classifying chronic pain syndromes.

Two possible limitations of the present study were (1) the different number of participants in the groups (chronic pain patients vs. pain-free controls), and (2) the fact that only women composed the group of chronic pain patients. To address the first limitation, a metric was used that is not affected by the differences in the number of subjects that make up the groups. In the case of the second limitation, the same analyses were performed but using the sex variable as the target to classify the participants. The levels of balanced accuracy, log loss, and AUC obtained with this variable as classifier were lower than those obtained with the variable "presence of chronic pain," which reinforces the idea

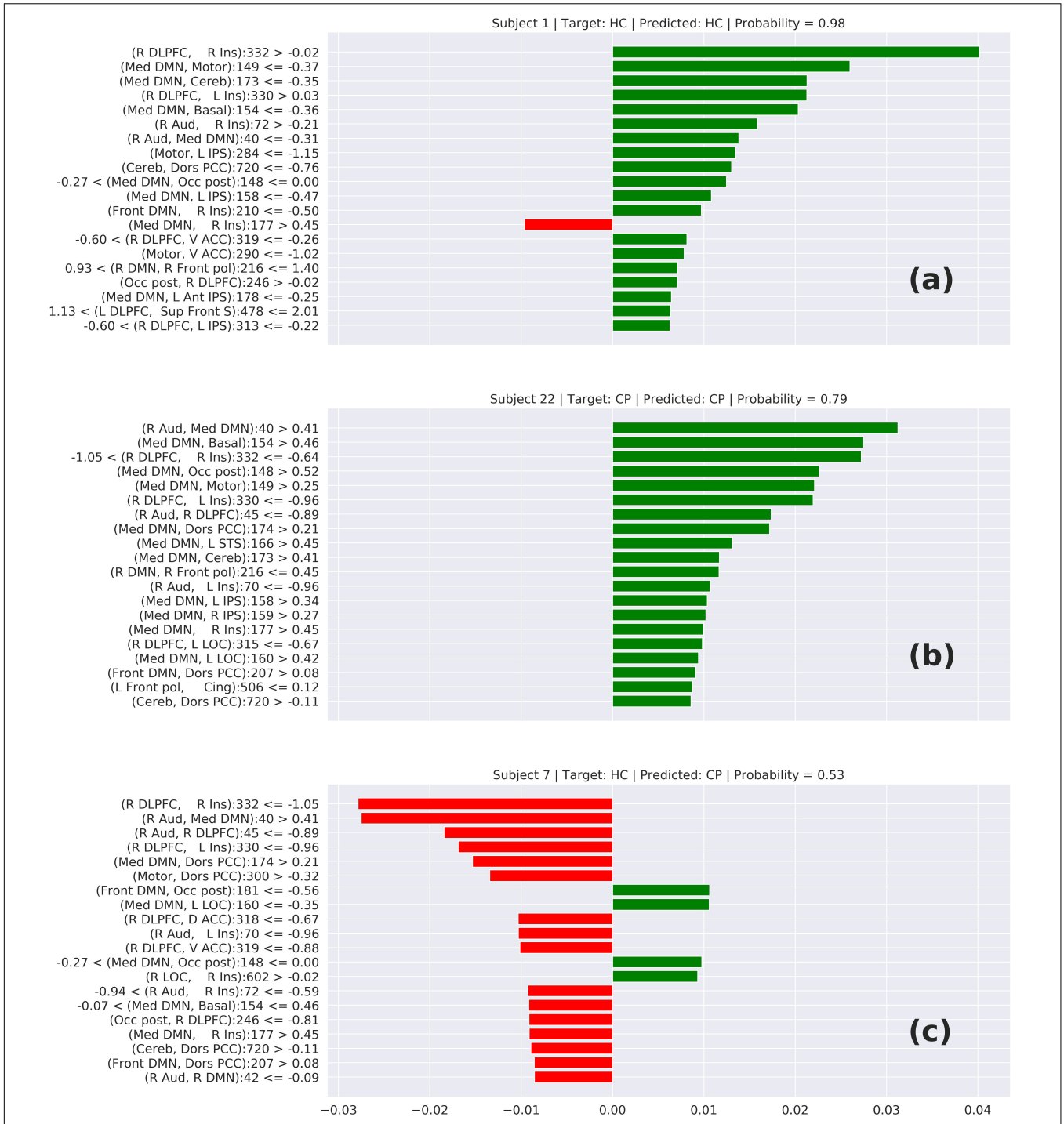


FIGURE 6 | Local interpretable model-agnostic explanations (LIME) for the best TOPT model listing the 20 features that most contribute to the prediction of **(a)** a correctly predicted HC, **(b)** a correctly predicted CP, and **(c)** a wrongly predicted HC. The bars represent how much each feature (axis *Y*) added to the prediction in terms of probability. Green bars indicate that the feature adds positively, in direction to the target class, while the red bars indicate that the features are in resistance to the target class.

that our method worked better to classify according to chronic pain instead of sex.

Nevertheless, it would be necessary to evaluate how these techniques perform in a multi-class problem differentiating not only between pain-free and chronic pain but identifying different chronic pain syndromes. Moreover, investigating the possibility to predict a syndrome, a longitudinal study of a population with the risk of developing chronic pain could permit the application

of an emerging and promising technique such as personalized machine learning (Rudovic et al., 2019).

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available because other studies have been conducted with the same data and they were not published yet. Requests to access the datasets should be directed to PM, pedro.montoya@uib.es.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee of the Balearic Islands. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AS would like to acknowledge the financial support of the CAPES Foundation, Brazil (proc. BEX 1703/2015-3). The research was also funded by several grants from ERDF/Spanish Ministry of Science, Innovation and Universities – State Agency of Research (Grant Nos: PSI2017-88388-C4-1-R and PSI2013-48260-C3-1-R). AS would also like to mention the support of the International Brain Research Organization (IBRO) in the form of its grant program with a short stay.

ACKNOWLEDGMENTS

The short stay was conducted at the King's College of London (United Kingdom) under the supervision of the professor Robert Leech.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01313/full#supplementary-material

REFERENCES

Ablin, J. N., and Buskila, D. (2015). Update on the genetics of the fibromyalgia syndrome. *Best Pract. Res. Clin. Rheumatol.* 29, 20–28. doi: 10.1016/j.berh.2015.04.018

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Muller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8:14. doi: 10.3389/fninf.2014.00014

Albe-Fessar, D., Berkley, K. J., Kruger, L., Ralston, H. J. III, and Willis, W. D. Jr. (1985). Diencephalic mechanisms of pain sensation. *Brain Res. Rev.* 9, 217–296. doi: 10.1016/0165-0173(85)90013-x

Apkarian, A. V., Sosa, Y., Krauss, B. R., Thomas, P. S., Fredrickson, B. E., Levy, R. E., et al. (2004). Chronic pain patients are impaired on an emotional decision-making task. *Pain* 108, 129–136. doi: 10.1016/j.pain.2003.12.015

Arnold, L. M., Bennett, R. M., Crofford, L. J., Dean, L. E., Clauw, D. J., Goldenberg, D. L., et al. (2019). AAPT diagnostic criteria for fibromyalgia. *J. Pain* 20, 611–628. doi: 10.1016/j.jpain.2018.10.008

Baliki, M. N., Geha, P. Y., Apkarian, A. V., and Chialvo, D. R. (2008). Beyond feeling: chronic pain hurts the brain, disrupting the default-mode network dynamics. *J. Neurosci.* 28, 1398–1403. doi: 10.1523/JNEUROSCI.4123-07.2008

Baliki, M. N., Schnitzer, T. J., Bauer, W. R., and Apkarian, A. V. (2011). Brain Morphological signatures for chronic pain. *PLoS One* 6:e26010. doi: 10.1371/journal.pone.0026010

Bevers, K., Watts, L., Kishino, N. D., and Gatchel, R. J. (2016). The biopsychosocial model of theassessment, prevention, and treatment of chronic pain. *Eur. Neurol. Rev.* 12:98. doi: 10.17925/USN.2016.12.02.98

Bishop, C. M. (2006). *Machine Learning and Pattern Recognition*. Berlin: Springer.

Brown, C. J., Miller, S. P., Booth, B. G., Poskitt, K. J., Chau, V., Synnes, A. R., et al. (2015). "Prediction of motor function in very preterm infants using connectome features and local synthetic instances," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, eds N. Navab, J. Hornegger, W. Wells, and A. Frangi (Cham: Springer).

Callan, D., Mills, L., Nott, C., England, R., and England, S. (2014). A tool for classifying individuals with chronic back pain: using multivariate pattern analysis with functional magnetic resonance imaging data. *PLoS One* 9:e98007. doi: 10.1371/journal.pone.0098007

Cámara, R. J. A., Merz, C., Wegmann, B., Stauber, S., Känel, R. V., and Egloff, N. (2016). Cost-Saving early diagnosis of functional pain in nonmalignant pain: a noninferiority study of diagnostic accuracy. *Pain Res. Treat.* 2016, 1–7. doi: 10.1155/2016/5964250

Cauda, F., Sacco, K., Duca, S., Cocito, D., D'Agata, F., Geminiani, G. C., et al. (2009). Altered resting state in diabetic neuropathic pain. *PLoS One* 4:e4542. doi: 10.1371/journal.pone.0004542

Cifre, I., Sitges, C., Fraiman, D., Muñoz, M. A., Balenzuela, P., González-Roldán, A., et al. (2012). Disrupted functional connectivity of the pain network in fibromyalgia. *Psychosom. Med.* 74, 55–62. doi: 10.1097/PSY.0b013e3182408f04

Cruz-Almeida, Y., and Fillingim, R. B. (2014). Can quantitative sensory testing move us closer to mechanism-based pain management? *Pain Med.* 15, 61–72. doi: 10.1111/pme.12230

Dampier, C., Palermo, T. M., Darbari, D. S., Hassell, K., Smith, W., and Zempsky, W. (2017). AAPT diagnostic criteria for chronic sickle cell disease pain. *J. Pain* 18, 490–498. doi: 10.1016/j.jpain.2016.12.016

Davis, K. D., and Moayedi, M. (2013). Central mechanisms of pain revealed through functional and structural MRI. *J. Neuroimmune Pharmacol.* 8, 518–534. doi: 10.1007/s11481-012-9386-9388

Davis, K. D., Racine, E., and Collett, B. (2012). Neuroethical issues related to the use of brain imaging: can we and should we use brain imaging as a biomarker to diagnose chronic pain? *Pain* 153, 1555–1559. doi: 10.1016/j.pain.2012.02.037

Diatchenko, L., Slade, G. D., Nackley, A. G., Bhalang, K., Sigurdsson, A., Belfer, I., et al. (2005). Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Hum. Mol. Genet.* 14, 135–143. doi: 10.1093/hmg/ddi013

Dodick, D. W., Loder, E. W., Adams, A. M., Buse, D. C., Fanning, K. M., Reed, M. L., et al. (2016). Assessing barriers to chronic migraine consultation, diagnosis, and treatment: results from the chronic migraine epidemiology and outcomes (CaMEO) study. *Headache* 56, 821–834. doi: 10.1111/head.12774

Du, Y., Fu, Z., and Calhoun, V. D. (2018). Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Front. Neurosci.* 12:525. doi: 10.3389/fnins.2018.00525

Dworkin, R. H., Bruehl, S., Fillingim, R. B., Loeser, J. D., Terman, G. W., and Turk, D. C. (2016). Multidimensional diagnostic criteria for chronic pain: introduction to the ACTTION–American Pain Society Pain Taxonomy (AAPT). *J. Pain* 17, T1–T9. doi: 10.1016/j.jpain.2016.02.010

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Freeman, R., Edwards, R., Baron, R., Bruehl, S., Cruccu, G., Dworkin, R. H., et al. (2019). AAPT diagnostic criteria for peripheral neuropathic pain: focal and segmental disorders. *J. Pain* 20, 369–393. doi: 10.1016/j.jpain.2018.10.002

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Map.* 2, 189–210. doi: 10.1002/hbm.460020402

Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., and Turk, D. C. (2010). The biopsychosocial approach to chronic pain: scientific advances and future directions. *Psychol. Bull.* 133, 581–624. doi: 10.1037/0033-2909.133.4.581

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Gorgolewski, K., Madison, C., Burns, C. D., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5:13. doi: 10.3389/fninf.2011.00013

Harper, D. E., Shah, Y., Peltier, S. J., Ichesco, E., and Gerstner, G. E. (2016). Multivariate classification of pain-evoked brain activity in temporomandibular disorder. *Pain Rep.* 1:e572. doi: 10.1097/pr9.0000000000000572

Hidalgo, B., Gilliaux, M., Poncin, W., and Detrembleur, C. (2012). Reliability and validity of a kinematic spine model during active trunk movement in healthy subjects and patients with chronic non-specific low back pain. *J. Rehabil. Med.* 44, 756–763. doi: 10.2340/16501977-16501015

Hidalgo, B., Nielens, H., Christine, P., Gilliaux, M., Hall, T., and Detrembleur, C. (2014). Use of kinematic algorithms to distinguish people with chronic non-specific low back pain from asymptomatic subjects: a validation study. *J. Rehabil. Med.* 46, 819–823. doi: 10.2340/16501977-16501836

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *J. Mol. Struct.* 1134, 63–66. doi: 10.1016/j.molstruc.2016.12.061

Johnson, K. A., Rana, M., Clauw, D. J., Farmer, M. A., Kirages, D. J., Cosand, L., et al. (2015). Altered resting state neuromotor connectivity in men with chronic prostatitis/chronic pelvic pain syndrome: a MAPP. *Neuroimage Clin.* 8, 493–502. doi: 10.1016/j.nicl.2015.05.013

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., et al. (2017). BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. *Neuroimage* 146, 1038–1049. doi: 10.1016/j.neuroimage.2016.09.046

Kress, H., Aldington, D., Alon, E., Coaccioli, S., Collett, B., Coluzzi, F., et al. (2015). A holistic approach to chronic pain management that involves all stakeholders: change is needed. *Curr. Med. Res. Opin.* 31, 1743–1754. doi: 10.1185/03007995.2015.1072088

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Proceeding NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe.

Lalonde, L., Leroux-Lapointe, V., Choinière, M., Martin, E., Lussier, D., Berbiche, D., et al. (2014). Knowledge, attitudes and beliefs about chronic noncancer pain in primary care: a canadian survey of physicians and pharmacists. *Pain Res. Manag.* 19, 241–250. doi: 10.1155/2014/760145

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lin, Y., Bai, Y., Liu, P., Yang, X., Qin, W., Gu, J., et al. (2017). Alterations in regional homogeneity of resting-state cerebral activity in patients with chronic prostatitis/chronic pelvic pain syndrome. *PLoS One* 12:e0184896. doi: 10.1371/journal.pone.0184896

Lindquist, M. A., Geuter, S., Wager, T. D., and Caffo, B. S. (2019). Modular preprocessing pipelines can reintroduce artifacts into FMRI Data. *Hum. Brain Mapp.* 40, 2358–2376. doi: 10.1002/hbm.24528

Lopez-Martinez, D., Peng, K., Lee, A., Borsook, D., and Picard, R. (2019). "Pain Detection with FNIRS-measured brain signals: a personalized machine learning approach using the wavelet transform and bayesian hierarchical modeling with dirichlet process priors," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII) Workshop on Recognition, Treatment and Management of Pain and Distress*, Cambridge.

Lopez-Martinez, D., Peng, K., Steele, S. C., Lee, A. J., Borsook, D., and Picard, R. (2018). "Multi-task multiple kernel machines for personalized pain recognition from functional near-infrared spectroscopy brain signals," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2320–2325. doi: 10.1109/ICPR.2018.8545823

López-Solà, M., Woo, C., Pujol, J., Deus, J., Harrison, B. J., Monfort, J., et al. (2017). Towards a neurophysiological signature for fibromyalgia. *Pain* 158, 34–47. doi: 10.1097/j.pain.0000000000000707

Lovati, C., Giani, L., Mele, F., Sinelli, A., Tien, T. T., Preziosa, G., et al. (2016). Brain plasticity and migraine transformation: FMRI evidences. *Expert Rev. Neurother.* 16, 1413–1425. doi: 10.1080/14737175.2016.1208565

Merskey, H., Addison, R. G., Beric, A., Blumberg, H., Bogduk, N., Boivie, J., et al. (1994). *Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms*. Washington, D.C: IASP PRESS.

Meszlényi, R., Peska, L., Gál, V., Vidnyánszky, Z., and Buza, K. (2016). "Classification of FMRI data using dynamic time warping based functional connectivity analysis," in *Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO)*, Budapest.

Meszlényi, R. J., Buza, K., and Vidnyánszky, Z. (2017). Resting state FMRI functional connectivity-based classification using a convolutional neural network architecture. *Front. Neuroinform.* 11:61. doi: 10.3389/fninf.2017.00061

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536. doi: 10.1038/nn.4393

Montoya, P., Larbig, W., Braun, C., Preissl, H., and Birbaumer, N. (2004). Influence of social support and emotional context on pain processing and magnetic brain responses in fibromyalgia. *Arthrit. Rheumat.* 50, 4035–4044. doi: 10.1002/art.20660

Napadow, V., and Harris, R. E. (2014). What has functional connectivity and chemical neuroimaging in fibromyalgia taught us about the mechanisms and management of 'Centralized' pain? *Arthrit. Res. Ther.* 16, 1–8. doi: 10.1186/s13075-014-0425-420

Ohrbach, R., and Dworkin, S. F. (2019). AAPT diagnostic criteria for chronic painful Temporomandibular disorders. *J. Pain* 20, 1276–1292. doi: 10.1016/j.jpain.2019.04.003

Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. (2016a). "Evaluation of a tree-based pipeline optimization tool for automating data science," in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, Denver, CO.

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., and Moore, J. H. (2016b). "Automating biomedical data science through tree-based pipeline optimization," *Applications of Evolutionary Computation. EvoApplications 2016. Lecture Notes in Computer Science* eds G. Squillero, and P. Burelli (Cham: Springer).

O'Shea, A., Craggs, J. G., Gay, C. W., Robinson, M. E., Lai, S., Staud, R., et al. (2015). Abnormal resting-state functional connectivity in patients with chronic fatigue syndrome: results of seed and data-driven analyses. *Brain Connect.* 6, 48–56. doi: 10.1089/brain.2015.0366

Pagé, I., Abboud, J., Shaughnessy, J. O., Laurencelle, L., and Descarreaux, M. (2015). Chronic low back pain clinical outcomes present higher associations with the STarT back screening tool than with physiologic measures: a 12-month cohort study. *BMC Musculoskel. Disord.* 16:201. doi: 10.1186/s12891-015-0669-660

Paice, J. A., Mulvey, M., Bennett, M., Dougherty, P. M., Farrar, J. T., Mantyh, P. W., et al. (2017). AAPT diagnostic criteria for chronic cancer pain conditions. *J. Pain* 18, 233–246. doi: 10.1016/j.jpain.2016.10.020

Paraschiv-Ionescu, A., Perruchoud, C., Buchser, E., and Aminian, K. (2012). Barcoding human physical activity to assess chronic pain conditions. *PLoS One* 7:e32239. doi: 10.1371/journal.pone.0032239

Pinheiro, E. S. S., Queirós, F. C., Montoya, P., Santos, C. L., Nascimento, M. A., Ito, C. H., et al. (2016). Electroencephalographic patterns in chronic pain: a systematic review of the literature. *PLoS One* 11:e0149085. doi: 10.1371/journal.pone.0149085

Richiardi, J., Altmann, A., Milazzo, A. C., Chang, C., Chakravarty, M. M., Banaschewski, T., et al. (2015). Correlated gene expression supports synchronous activity in brain networks. *Science* 348, 1241–1244. doi: 10.1126/science.1255905

Robinson, M. E., O'Shea, A. M., Craggs, J. G., Price, D. D., Letzen, J. E., and Staud, R. (2015). Comparison of machine classification algorithms for fibromyalgia: neuroimages versus self-report. *J. Pain* 16, 472–477. doi: 10.1016/j.jpain.2015.02.002

Rudovic, O., Utsumi, Y., Guerrero, R., Peterson, K., Rueckert, D., and Picard, R. W. (2019). "Meta-weighted gaussian process experts for personalized forecasting of AD cognitive changes," in *Machine Learning for Healthcare Conference (ML4HC2019)*, 1–15.

Sakoe, H., and Chiba, S. (1978). "dynamic programming algorithm optimization for spoken word recognition," in *Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal Processing*, Piscataway, NJ.

Salvador, S., and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* 11, 561–580. doi: 10.3233/IDA-2007-11508

Schmidt-Wilcke, T. (2015). Neuroimaging of chronic pain. *Best Pract. Res. Clin. Rheumatol.* 29, 29–41. doi: 10.1016/j.berh.2015.04.030

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb. Cortex* 22, 158–165. doi: 10.1093/cercor/bhr099

Smith, S. M., Dworkin, R. H., Turk, D. C., Baron, R., Polydefkis, M., Tracey, I., et al. (2017). The potential role of sensory testing, skin biopsy, and functional brain imaging as biomarkers in chronic pain clinical trials: IMMPACT considerations. *J. Pain* 18, 757–777. doi: 10.1016/j.jpain.2017.02.429

Sundermann, B., Burgmer, M., Pogatzki-Zahn, E., Gaubitz, M., Stüber, C., Wessolleck, E., et al. (2014). Diagnostic classification based on functional connectivity inchronic pain: model optimization in fibromyalgia and rheumatoid arthritis. *Acad. Radiol.* 21, 369–377. doi: 10.1016/j.acra.2013.12.003

Traeger, A. C., Henschke, N., Hübscher, M., Williams, C. M., Kamper, S. J., Maher, C. G., et al. (2016). Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. *PLoS Med.* 13:e1002019. doi: 10.1371/journal.pmed.1002019

Truini, A., Tinelli, E., Gerardi, M. C., Calistri, V., Iannuccelli, C., La Cesa, S., et al. (2016). Abnormal resting state functional connectivity of the periaqueductal grey in patients with fibromyalgia. *Clin. Exp. Rheumatol.* 34(2 Suppl. 96), 129–133.

Ultsch, A., Kringel, D., Kalso, E., Mogil, J. S., and Lötsch, J. (2016). A data science approach to candidate gene selection of pain regarded as a process of learning and neural plasticity. *Pain* 157, 2747–2757. doi: 10.1097/j.pain.0000000000000694

Ung, H., Brown, J. E., Johnson, K. A., Younger, J., Hush, J., and Mackey, S. (2014). Multivariate classification of structural MRI data detects chronic low back pain. *Cereb. Cortex* 24, 1037–1044. doi: 10.1093/cercor/bhs378

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). "Information processing in medical imaging," in *Lecture Notes in Computer Science; Information Processing in Medical Imaging*, eds G. Székely, and H. K. Hahn (Berlin: Springer).

Vieira, S., Pinaya, W. H. L., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Min. Knowl. Discov.* 26, 275–309. doi: 10.1007/s10618-012-0250-255

Widerström-Noga, E., Loeser, J. D., Jensen, T. S., and Finnerup, N. B. (2017). AAPT diagnostic criteria for central neuropathic pain. *J. Pain* 18, 1417–1426. doi: 10.1016/j.jpain.2017.06.003

Wolfe, F., Clauw, D. J., Fitzcharles, M., Goldenberg, D. L., Häuser, W., Katz, R. L., et al. (2016). 2016 revisions to the 2010/2011 fibromyalgia diagnostic criteria. *Semin. Arthrit. Rheum.* 46, 319–329. doi: 10.1016/j.semarthrit.2016.08.012

Zamorano, A. M., Cifre, I., Montoya, P., Riquelme, I., and Kleber, B. (2017). Insula-based networks in professional musicians: evidence for increased functional connectivity during resting state FMRI. *Hum. Brain Mapp.* 38, 4834–4849. doi: 10.1002/hbm.23682

Zhou, Q., Wesselmann, U., Walker, L., Lee, L., Zeltzer, L., and Verne, G. N. (2018). AAPT diagnostic criteria for chronic abdominal, pelvic, and urogenital pain: irritable bowel syndrome. *J. Pain* 19, 257–263. doi: 10.1016/j.jpain.2017.10.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Santana, Cifre, de Santana and Montoya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 14 January 2020
doi: 10.3389/fnins.2019.01325

Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network

Zeinab Sherkatghanad[1†], Mohammadsadegh Akhondzadeh[2†], Soorena Salari[3†], Mariam Zomorodi-Moghadam[4†], Moloud Abdar[5†], U. Rajendra Acharya[6,7,8], Reza Khosrowabadi[9] and Vahid Salari[1,10]*

[1] Department of Physics, Isfahan University of Technology, Isfahan, Iran, [2] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, [3] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran, [4] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, [5] Departement of Computer Science, University of Quebec in Montreal, Montreal, QC, Canada, [6] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore, Singapore, [7] Department of Biomedical Engineering, School of Science and Technology, Singapore School of Social Sciences, Singapore, Singapore, [8] International Research Organization for Advanced Science and Technology (IROAST) Kumamoto University, Kumamoto, Japan, [9] Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran, [10] Department of Physical Chemistry, University of the Basque Country UPV/EHU, Bilbao, Spain

OPEN ACCESS

Edited by:
Reza Lashgari, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Changiz Eslahchi, Shahid Beheshti University, Iran
Hamid Reza Marateb, Universitat Politecnica de Catalunya, Spain
Seyed-Mahdi Khaligh-Razavi, Massachusetts Institute of Technology, United States

***Correspondence:**
Vahid Salari
vahidsalari@iut.ac.ir

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 July 2019*
Accepted: *26 November 2019*
Published: *14 January 2020*

Citation:
Sherkatghanad Z, Akhondzadeh M, Salari S, Zomorodi-Moghadam M, Abdar M, Acharya UR, Khosrowabadi R and Salari V (2020) Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. Front. Neurosci. 13:1325. doi: 10.3389/fnins.2019.01325

Background: Convolutional neural networks (CNN) have enabled significant progress in speech recognition, image classification, automotive software engineering, and neuroscience. This impressive progress is largely due to a combination of algorithmic breakthroughs, computation resource improvements, and access to a large amount of data.

Method: In this paper, we focus on the automated detection of autism spectrum disorder (ASD) using CNN with a brain imaging dataset. We detected ASD patients using most common resting-state functional magnetic resonance imaging (fMRI) data from a multi-site dataset named the Autism Brain Imaging Exchange (ABIDE). The proposed approach was able to classify ASD and control subjects based on the patterns of functional connectivity.

Results: Our experimental outcomes indicate that the proposed model is able to detect ASD correctly with an accuracy of 70.22% using the ABIDE I dataset and the CC400 functional parcellation atlas of the brain. Also, the CNN model developed used fewer parameters than the state-of-art techniques and is hence computationally less intensive. Our developed model is ready to be tested with more data and can be used to prescreen ASD patients.

Keywords: convolutional neural networks, autism spectrum disorder, ABIDE, fMRI, atlas

1. INTRODUCTION

Autism spectrum disorder (ASD), a type of neurological disorder, appears in children between 6 and 17 years of age and affects communication skills and social behavior. ASD affects social interactions and communication and causes repetitive behaviors in patients (Bhat et al., 2014a,b; Huang et al., 2019). According to the WHO, ASD affects one child in 160, and these children often present with other conditions like depression, anxiety, and attention deficit hyperactivity

disorder (ADHD)[1]. Early diagnosis during childhood is important and can improve the social skills and communication problems of children with ASD and enhance their quality of life. In order to control and treat this disease, an early diagnosis is crucial. One of the most important tasks for diagnosing neurological diseases such as epilepsy, Alzheimer, and autism is to develop a model based on functional or structural region relationships in the brain (Wing, 1997; American Psychiatric Association, 2011; Chen et al., 2011). Hence, functional magnetic resonance imaging (fMRI) is used to study the brain and its structures. It detects correlated fluctuations in the blood oxygen level-dependent (BOLD) signals from the brain regions. The most common data-driven method for autism diagnosis and the investigation of its biomarkers is the autism brain imaging data exchange (ABIDE), which is a collaborative effort involving neuroimaging and phenotypic data obtained from 1,112 individuals (Di Martino et al., 2014). The ABIDE is a worldwide multi-site database consisting of two phases. The first phase (ABIDE I) consists of 1,112 individuals, with 539 ASD patients and 573 others, from 17 sites. The second phase (ABIDE II) has 521 ASD patients and 593 healthy controls and was obtained from 19 sites. The ABIDE I dataset is obtained from 17 international imaging sites and is composed of structural, resting-state fMRI data and phenotypic information.

Recently, many efforts have been made to identify ASD based on deep learning with fMRI (Koyamada et al., 2015; Anirudh and Thiagarajan, 2017; Subbaraju et al., 2017). In Koyamada et al. (2015), a deep neural network (DNN) model was investigated in order to build a subject-transfer decoder. The authors used principal sensitivity analysis (PSA) to construct a decoder for visualizing different features of all individuals in the dataset. Their proposed neural network includes two hidden layers and a softmax output layer, in which the two hidden layers in the middle classify brain activities into seven human categories from 499 subjects.

It has been shown that ASD disrupts the functional connectivity between the multiple brain regions that affect global brain networks. Therefore, the main goal of many researchers in this area is to classify ASD and control subjects based on the neural patterns of functional connectivity (Bourgeron, 2009; Anderson et al., 2011; Mennes et al., 2011; Schipul et al., 2011; Nielsen et al., 2013; von dem Hagen et al., 2013; Plitt et al., 2015; Dvornek et al., 2017; Parisot et al., 2017, 2018; Aghdam et al., 2018; Xing et al., 2018; Kazeminejad and Sotero, 2019; Sharif and Khan, 2019) and improve the accuracy of classification. For example, Nielsen et al. (2013) achieved 60% classification accuracy, and Abraham et al. (2017) obtained 67% accuracy in classifying ASD and control subjects. Heinsfeld et al. (2018) applied deep learning algorithms to identify ASD patients and improved the accuracy, reaching 70%. They employed two stacked denoising autoencoders to extract a lower-dimensional version of the ABIDE I dataset and also identified the areas of the brain that played the most important role in differentiating ASD from typical controls (TC). The volumetric convolutional neural network (CNN) model, which is considered as the full-resolution 3D spatial structure of resting-state functional MRI data, is investigated in Khosla et al. (2018).

In recent years, the use of CNN has attracted a lot of attention in the field of classification and representation learning. CNNs are powerful classifiers with high accuracies in many applications with many free parameters. Also, CNN models have higher accuracy for feature extraction and can handle many free parameters. The CNN model includes different parts such as an activation function, convolutional layers, fully connected layers, normalization layers, and pooling layers.

The CNN technique has the ability to interpret brain biomarkers in ASD patients using fMRI. The ASD biomarkers play an important role in early diagnosis and treatment (Li et al., 2018b). Li et al. (2018a) proposed multi-channel convolutional neural networks based on a patch-level data-expanding method to diagnose early biomarkers of ASD. Choi (2017), multivariate and high dimensional data are reduced to two-dimensional features, and the functional connectivity pattern associated with ASD is investigated by using a variational autoencoder (VAE).

The stereotypical motor movements (SMM) in autism patients are body rocking and complex hand movements, which will affect learning and social skills. The CNN is used to learn different features from multi-sensor accelerometer signals of SMM (Rad et al., 2015). A fully automated brain tumor segmentation method using CNN was proposed in Havaei et al. (2017).

The purpose of the present study is to investigate the performance of a CNN in classifying ASD and control subjects. We used the fMRI data represented by a multi-site database known as ABIDE I. The ABIDE I data have been preprocessed by the Preprocessed Connectomes Project (PAC). We improved the previously reported results and obtained 70.2% accuracy in the distinction of ASD from control subjects. The performance of the developed model is evaluated using three supervised methods, namely SVM (support vector machine), KNN (K-nearest neighbors), and RF (random forest) classifiers on the preprocessed ABIDE I dataset. Our results show that the average accuracy values after optimization or hyperparameter tuning for SVM, KNN, and RF are 69, 62, and 60%, respectively. Therefore, the proposed CNN model outperformed these machine learning methods. It has been shown that having a CNN model with fewer parameters is very important and leads to less overhead for the new models (Iandola et al., 2016). Our developed model has obtained high accuracy and was also able to train with fewer parameters, which reduces the computation time. An autoencoder has been used to diagnose schizophrenia (Zeng et al., 2018). Functional connectivity MRI data from multiple sites have been used for classification. The authors obtained an accuracy of 85% for multi-site pooling classification and 81% for leave-site-out transfer classification. In this approach, each time, one site out of 17 sites was used as a test and the rest were used for training. The results show that the sites named the Kennedy Krieger Institute, Baltimore (KKI), San Diego State University (SDSU), and University of Utah School of Medicine (USM) achieved higher accuracies than other sites.

The rest of the paper is organized as follows. The details of the ABIDE I dataset, the data preprocessing, and the development of

[1]http://www.healthdata.org/gbd

TABLE 1 | The distribution of sex and average age at different sites for typical control (TC) and ASD classes.

Site	Abbreviation	TC			ASD		
		Average age	Sex		Average age	Sex	
			Male	Female		Male	Female
CALTECH	California Institute of Technology	28	14	4	27.4	15	4
CMU	Carnegie Mellon University	26.8	10	3	26.4	11	3
KKI	Kennedy Krieger Institute, Baltimore	10	20	8	10	16	4
LEUVEN	University of Leuven	18.2	29	5	17.8	26	3
MAX MUN	Ludwig Maximilians University, Munich	24.6	27	1	26.1	21	3
NYU	NYU Langone Medical Center, New York	15.7	74	26	14.7	65	10
OHSU	Oregon Health and Science University	10.1	14	0	11.4	12	0
OLIN	Olin, Institute of Living, Hartford Hospital	16.7	13	2	16.5	16	3
PITT	University of Pittsburgh School of Medicine	18.9	23	4	19	25	4
SBL	Social Brain Lab BCN NIC UMC Groningen and Netherlands Institute for Neurosciences	33.7	15	0	35	15	0
SDSU	San Diego State University	14.2	16	6	14.7	13	1
STANFORD	Stanford University	10	16	4	10	15	4
TRINITY	Trinity Center for Health Sciences	17.1	25	0	16.8	22	0
UCLA	University of California, Los Angeles	13	38	6	13	48	6
UM	University of Michigan	14.8	56	18	13.2	57	9
USM	University of Utah School of Medicine	21.3	25	0	23.5	46	0
YALE	Child Study Center, Yale University	12.7	20	8	12.7	20	8

TABLE 2 | Different parameters in structural MRI imaging for each site in ABIDE I.

	Voxel size (mm^3)	Flip angle (deg)	TR (ms)	TE (ms)	T1 (ms)
CALTECH	1	10	1,590	2.73	800
CMU	1	8	1,870	2.48	1,100
KKI	1	8	8	3.7	843
LEUVEN	0.98× 0.98×1.2	8	9.6	4.6	885.145
MAX MUN	1	9	1,800	3.06	900
NYU	1.3 × 1.3	7	2,530	3.25	1,100
OHSU	1	10	2,300	3.58	900
OLIN	1	8	2,500	2.74	900
PITT	1.1 × 1.1 × 1.1	7	2,100	3.93	1,000
SBL	1	8	9	3.5	1,000
SDSU	1	45	11.08	4.3	NA
STANFORD	0.86 × 1.5 ×0.86	15	8.4	1.8	NA
TRINITY	1	8	8.5	3.9	1060.17
UCLA	1 ×1× 1.2	9	2,300	2.84	853
UM	1.2 × 1 × 1	15	250	1.8	500
USM	1 ×1× 1.2	9	2,300	2.91	900
YALE	1	9	1,230	1.73	624

the new CNN model are provided in section 2. In section 3, visual representations of the most important brain areas are presented. Section 4 shows the detailed results of analysis, and finally, the results are discussed in section 5.

2. MATERIALS AND METHODS

2.1. Materials

In this work, we used the first phase of resting-state fMRI data from the multi-site ABIDE I. ABIDE I is a consortium of collected resting-state fMRIs from 17 international imaging sites and matched controls that is provided for scientific research. Each site in the ABIDE I dataset uses different parameters and protocols. The fMRI protocol has been used as the imaging protocol at all of the sites. In this work, brain volume is represented by small cubic elements named voxels. The inclusion criteria for sites was having at least 20 subjects meeting other criteria for inclusion like successful preprocessing with manual visual inspection of normalization to MNI space of MPRAGE. The autism diagnostic observation tool and autism diagnostic interview-revised were used for ASD diagnosis or typical control confirmation in the majority of the sites. These types of data increase understanding of the neural bases of ASD. Resting-state fMRI is based on neural measurements of functional connectivity between multiple brain regions. This functional connectivity is calculated by the correlation of the average time series from the regions of interest (ROI). Fluctuations in blood oxygenation lead to low-frequency fluctuation correlations in resting-state fMRI, which gives the connectivity matrix. In the present study, we used the datasets from 505 ASD patients and 530 typical controls. These datasets contain T1 structural brain images, fMRI images, and phenotypic information relating to different patients. The phenotypic information is classified based on sex, age, and

FIGURE 1 | Proposed CNN architecture for automated detection of ASD.

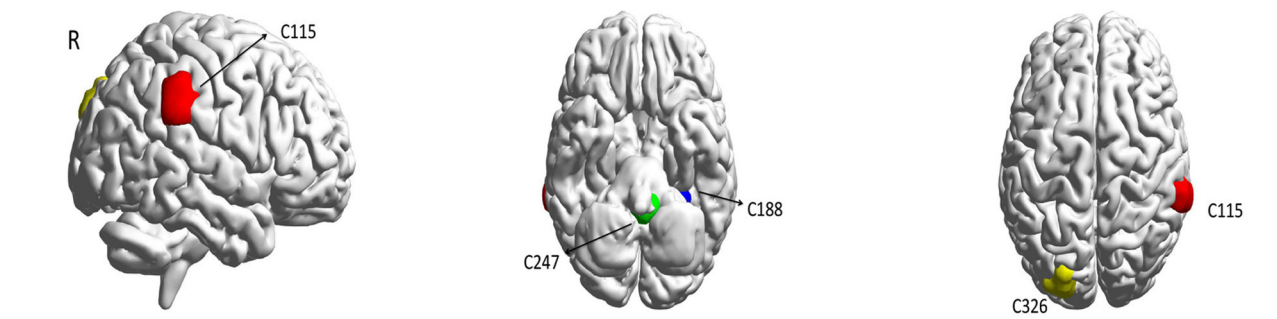


FIGURE 2 | The most important ROIs for ASD classification in the prediction model according to the saliency map. We consider Red, Blue, Green, and Yellow areas corresponding to (61.9; −36.3; 34.4), (−27.6; −40.2; −17.6), (−2.1; −43.0; −40.7), (−22.5; −85.5; 31.0), respectively.

autism diagnostic observation schedule (ADOS) score for ASD subjects and mean framewise displacement (FD) quality, which is a measure of subject head motion[2]. The distributions of sex and average age at different sites for typical control (TC) and ASD patients are summarized in **Table 1**.

[2]http://preprocessed-connectomes-project.org/abide/quality_assessment.html

TABLE 3 | Important areas for ASD classification in prediction mode.

Color	Red	Blue	Green	Yellow
ROI number	C115	C188	C247	C326
Center of mass	(61.9; −36.3; 34.4)	(−27.6; −40.2; −17.6)	(−2.1; −43.0; −40.7)	(−22.5; −85.5; 31.0)

2.2. Data Preprocessing of the ABIDE I Dataset

The Preprocessed Connectomes Project (PCP) is a publicly available preprocessed version of data from both the 1,000 Functional Connectomes Project (FCP) and the International Neuroimaging Data-Sharing Initiative (INDI)[3]. We used data from the FCP using the configurable pipeline, the Analysis of Connectomes (CPAC). After the preprocessing, we obtained 871 quality MRI images with phenotypic information. The preprocessing step included slice timing correction, correction for motion, and normalization of voxel intensity. Nuisance regression was employed to delete the signal fluctuations caused by head motion, respiration, cardiac pulsation, and scanner drift. The signal fluctuation was modeled using 24 motion parameters for head motion, a quadratic and linear term for scanner drift, and CompCor with five principal components for physiological noise (Friston et al., 1995; Fox et al., 2005; Lund et al., 2005; Behzadi et al., 2007). Bandpass filtering (0.01–10 Hz) was used in our analysis.

We used the CC400 functional parcellation atlas of the brain throughout our study. In this atlas, a brain connectivity matrix is constructed for the average time series of the ROI, partitioned into 400 regions. There are many different parameters in MRI imaging, including voxel size, flip angle, TR, TE, and T1. **Table 2** summarizes the different parameters in structural MRI imaging for each site in ABIDEI.

In the following, we will describe our proposed CNN architecture in detail.

2.3. Network Architecture

In this work, we obtained connectomes or functional connectivity matrices for the detection of ASD classes. This symmetric matrix shows the correlation between the mean values of the time series obtained from an ROI. Each cell in the matrix contains a Pearson correlation coefficient, and each row is the representation of the ROI.

The Pearson correlation coefficient (ranges from −1 to 1) is a correlation index between two areas of the brain regions, with 1 representing high correlation between the two areas of the brain and vice versa. Thus, a 392 × 392 matrix is found in the CC400 functional parcellation atlas for each subject, which represents the co-activation correlations of 392 brain areas. By considering each row as the representation of a brain region, we propose a CNN architecture for connectomic data. We used a CNN architecture with one convolutional layer, interspersed within max-pooling followed by densely connected layers (Please see **Figure 1**). The functional connectivity matrices between pairs of ROI are fed as input to convolutional layers. Our final CNN model is as follows: 1 fully connected hidden layer and each linear layer followed by a tanh activation function. The parallel filters with dimensions from 1 × 392 to 7 × 392 act on rows representing the brain regions. Thus, we take into account 400 filters of length 1 and width 392–400 filters of length 7 and width 392. In this condition, the sizes of the weights are equal to the representation matrix in the convolutional neural network. The hidden layer followed by max-pooling is used to reduce the number of features and avoid the overfitting problem. After the max-pooling layer, a dropout regularization keeps only 25% of the nodes for training. Finally, the output node is concated and fully connected to a dense layer, which is subsequently used for classification. Also, the model is trained for 300 epochs with a batch size of 32, and the learning rate is set to 0.005. The model as shown in **Figure 1** is developed using a 10-fold cross-validation strategy.

The proposed CNN model does not include feed-forward convolution. We employed concatenation of several convolution layers, and the whole result set obtained is passed to the multilayer perceptron (MLP) to complete the classification. In other words, each convolution layer has a specific meaning. For example, when the filter size is 1 × 392, the connection of each area with other areas will be considered, whereas when the filter size is increased to 7 × 392, the connection of 7 areas near each other with other areas will be seen. We combined these outputs to obtain the final output, which is ensemble learning from the convolution layers.

In the next section, we investigate the features that have the most contribution in ASD classification using a visualization method for our proposed CNN model.

3. VISUALIZATION OF IMPORTANT AREAS

Now, we are interested in visualizing the brain areas that are significant in the classification of ASD and control patients in the ABIDE I dataset. The field of computer vision has enabled vast progress for the visualization of CNN models. In neuroimaging, this technique provides the ability to gain more insights into biomarkers, which are important in early diagnosis and treatment. By using the visualization of image classification models learned via deep Convolutional Networks or ConvNet, we are able to reveal the important ROIs that play important roles in the classification (Simonyan et al., 2013). We obtained the important ROIs for ASD-detection using our model

TABLE 4 | Summary of performance values obtained for CNN with 10-fold cross-validation.

Fold	Accuracy	Confidence interval	Sensitivity	Specificity	F-score
1	0.6603	0.0901	0.6250	0.7000	0.6604
2	0.6699	0.0908	0.8889	0.4285	0.7384
3	0.7187	0.0899	0.8113	0.6046	0.7610
4	0.7582	0.0879	0.7755	0.7380	0.7755
5	0.7356	0.0926	0.7659	0.7000	0.7578
6	0.6395	0.1014	0.7826	0.4750	0.6990
7	0.7023	0.0978	0.7777	0.6153	0.7368
8	0.77901	0.0887	0.9318	0.6216	0.8283
9	0.6623	0.1056	0.7380	0.5714	0.7045
10	0.6849	0.1066	0.6500	0.7272	0.6933
Mean	0.7022	0.0855	0.7746	0.6182	0.7355

[3]http://preprocessed-connectomes-project.org

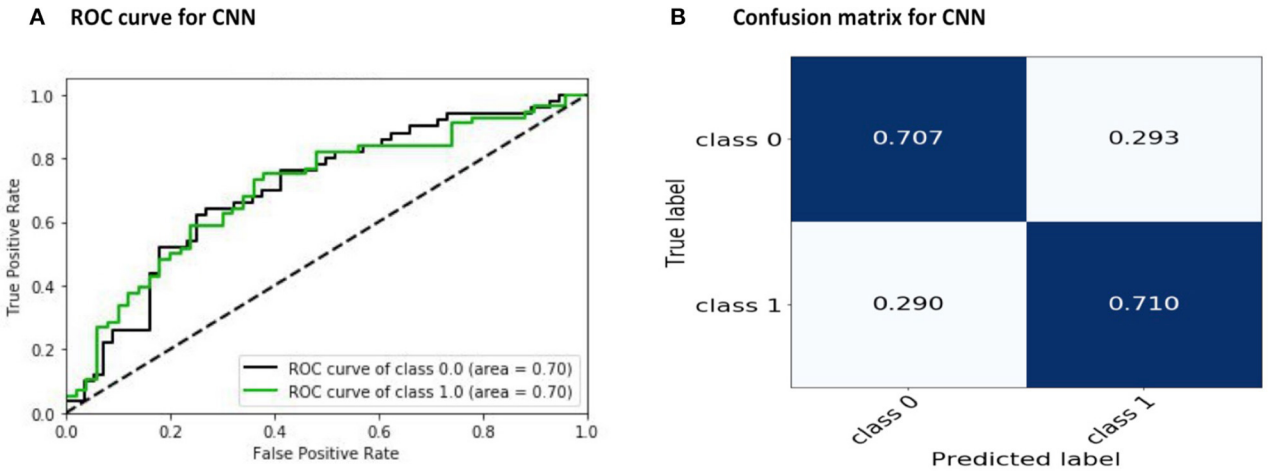


FIGURE 3 | Results of the proposed CNN model: **(A)** ROC and **(B)** convolution matrix. Here, Class 0 and Class 1 indicate the control subjects and ASD patients, respectively.

TABLE 5 | Summary of comparison table for automated detection of TC and ASD classes using the same database.

References	Protocol	Best method	Performance (%)		
			Specificity	Sensitivity	Accuracy
Nielsen et al. (2013)	–	Multiple bins and leave- one-out classifier	58.00	62.00	60.00
Parisot et al. (2017)	10-fold CV	Graph Convolutional Networks (GCN)	–	–	69.50
Dvornek et al. (2017)	10-fold CV	LSTM32	–	–	66.80
Parisot et al. (2018)	10-fold CV	Graph Convolutional Networks (GCN)	–	–	70.40
Aghdam et al. (2018)	10-fold CV	Deep belief Network (DBN)	32.96	84.00	65.56
Xing et al. (2018)	5-fold CV	CNN with element-wise filters (CNN-EW)	70.40	66.44	66.88
Kazeminejad and Sotero (2019)	Leave-one-site-out	Deep learning and PCA	65.00	67.00	66.00
Sharif and Khan (2019)	Leave-one-site-out	Multi-Layer Perceptron (MLP) and Feature Selection	–	–	56.26
Abraham et al. (2017)	10-fold CV	SVC-l1 and SVC-l2 Networks	–	–	67.00
Heinsfeld et al. (2018)	10-fold CV	SVM	62.00	68.00	65.00
Heinsfeld et al. (2018)	10-fold CV	Deep Neural Networks (DNN) and transfer learning	63.00	74.00	70.00
Present study	10-fold CV	CNN	61.00	77.00	70.20

TABLE 6 | Results of ROC for CNN, SVM, KNN, and RF classifiers before optimization (BO) and after optimization (AO).

	SVM		KNN		RF		CNN
	BO	AO	BO	AO	BO	AO	
Mean of accuracy	0.6890	0.6935	0.6142	0.6211	0.5983	0.5994	0.7022
Variance of accuracy	0.0022	0.0011	0.0011	0.0006	0.00062	0.00052	0.0020
Mean of sensitivity	0.7790	0.7459	0.7619	0.7452	0.7474	0.7595	0.7746
Variance of sensitivity	0.0028	0.0026	0.0045	0.004	0.0045	0.005	0.0078
Mean of specificity	0.5855	0.6325	0.4437	0.4784	0.4277	0.4149	0.6182
Variance of specificity	0.0057	0.0049	0.012	0.008	0.0065	0.0030	0.0098
Mean of AUC	0.7533	0.7553	0.6679	0.6724	0.6546	0.6635	0.7486
Variance of AUC	0.0018	0.0017	0.0021	0.0013	0.0005	0.0012	0.0006
Mean of F-score	0.6486	0.6719	0.5279	0.5516	0.5092	0.5015	0.7355

with saliency technique (**Figure 2**). This approach is based on computing the gradient of the class score with respect to the input image and calculating the class saliency map. In other words, we evaluated the gradient of the output category with respect to the input image:

$$\frac{\partial output}{\partial input} \tag{1}$$

Here, output indicates output category, and input is related to input image. The positive ratio indicates that a small change in the input image pixel leads to an increase in the output. Thus, we can obtain salient images of brain areas that play important roles in ASD detection.

We observed that four brain areas are significant in the diagnosis of ASD subjects for the CC400 functional parcellation atlas of the brain. These areas are named as C115, C188, C247, and C326, with the centers of mass equal to (61.9; −36.3; 34.4), (−27.6; −40.2; −17.6), (−2.1; −43.0; −40.7), and (−22.5; −85.5; 31.0), respectively (**Table 3**).

Our results show that the right supramarginal gyrus, which is considered to preserve self-other distinction during empathy in ASD patients (Hoffmann et al., 2015), seems to play a significant role in the diagnosis of autism. The fusiform gyrus, which is hypoactive in patients with autism (van Kooten et al., 2008), is also emphasized for ASD prediction. Also, the cerebellar vermis is indicated as an important area for the ASD classification, and this was reported to be smaller in autism cases (Kaufmann et al., 2003). In addition, these results support the idea of the disruption of anterior-posterior brain connectivity in ASD, which has been shown in Just (2004), Kana et al. (2009), and Cherkassky et al. (2006).

4. RESULTS

Nowadays, CNN is widely used for dataset classification. In this study, we designed a CNN model for automated detection of ASD using the ABIDE I dataset. The preprocessed neuroimaging data from the ABIDE I dataset is used in our experiment. There are 1,112 subjects (539 diagnosed with ASD, and 573 typical controls) in the ABIDE I dataset, reduced to 871 subjects after preprocessing. There is also a phenotype file for this dataset, which includes the automated metrics, specified with the prefix anat finc. Among them, we evaluated the functional metric called mean framewise displacement and removed the outliers where this parameter was over 0.2.

During training, the learning rate was set at 0.005 with batch sizes of 32 and 400 epochs. The input to the network is a 392 × 392 matrix, where each row represents one of the regions of

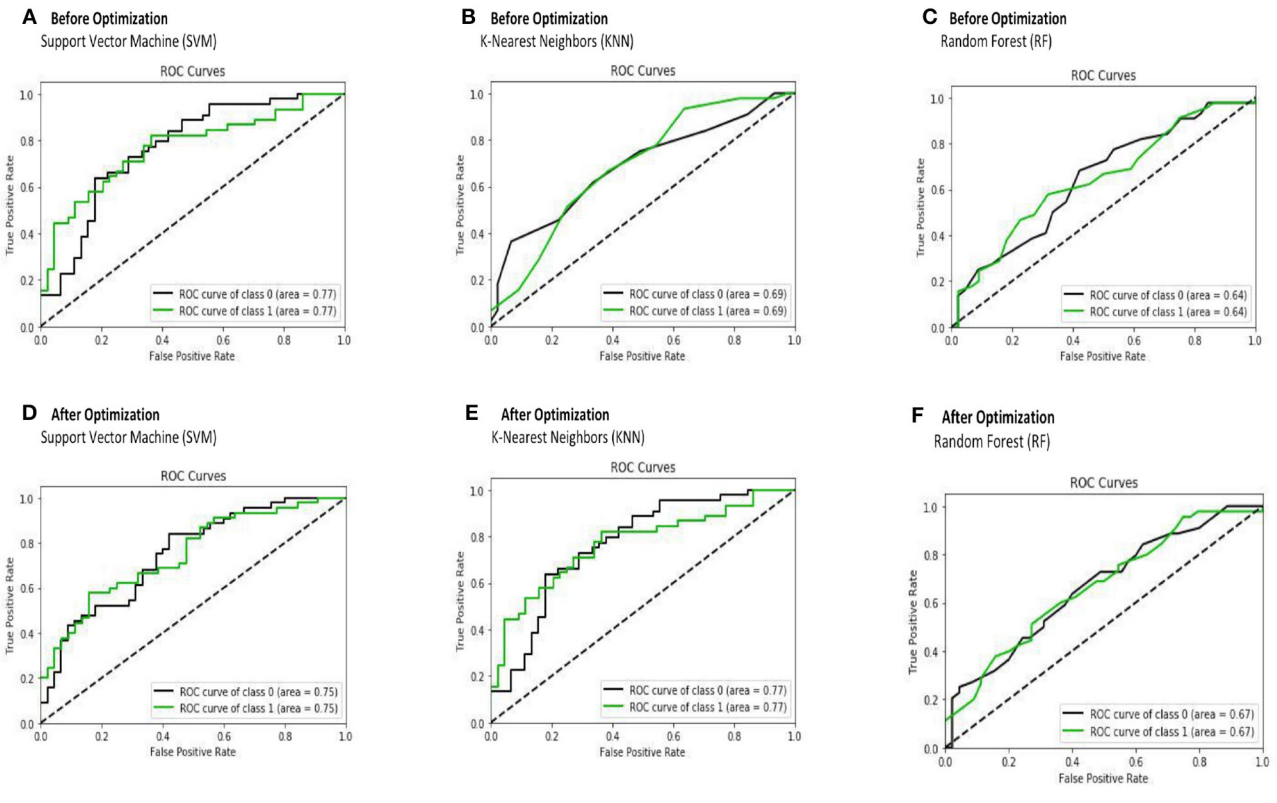


FIGURE 4 | The receiver operating characteristic curve (ROC) is depicted for SVM **(A,D)**, KNN **(B,E)**, and RF **(C,F)** classifiers before and after optimization.

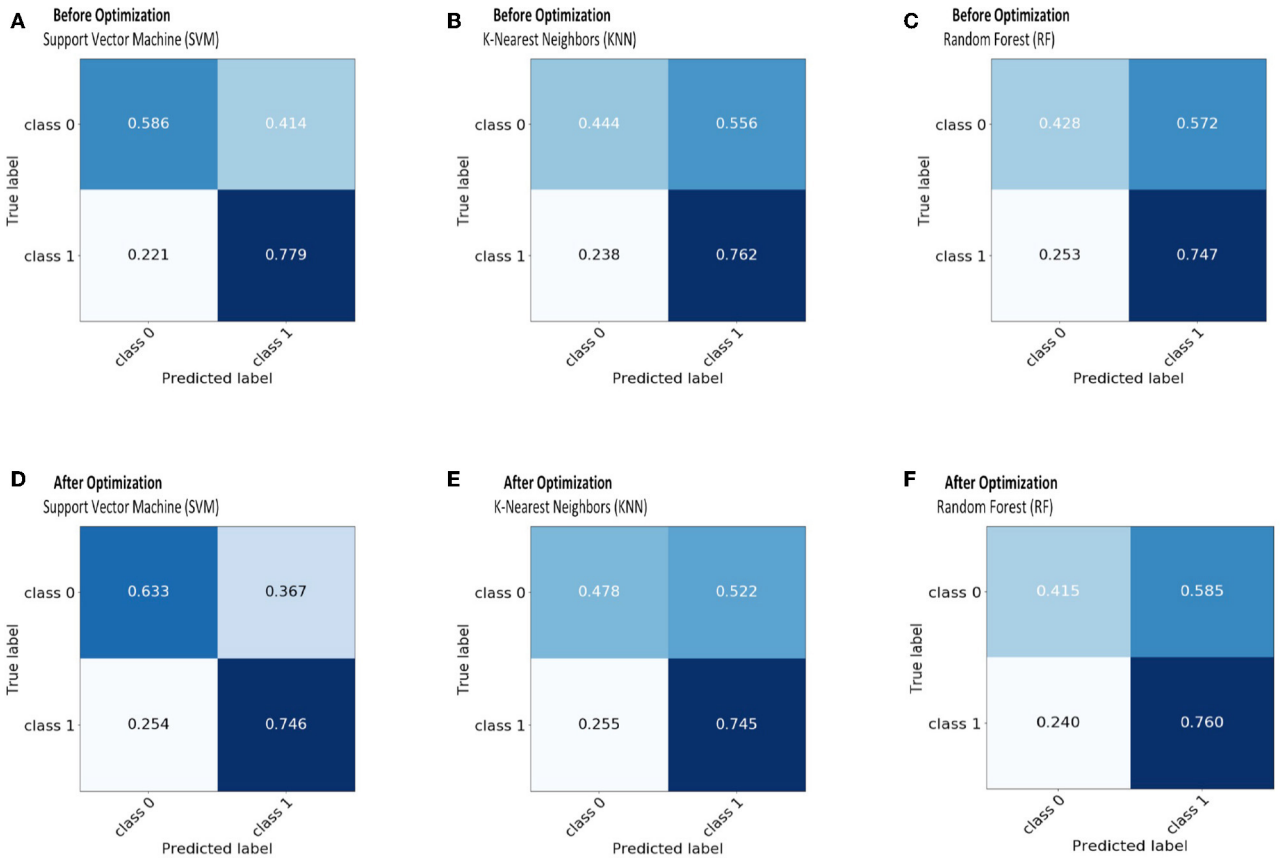


FIGURE 5 | The confusion matrix before and after optimization for classifiers: SVM-**(A,D)**; KNN-**(B,E)**; RF-**(C,F)**.

the brain. In our CNN architecture, we used the 400 filters with sizes from 1 × 392 to 7 × 392. Generally, the width of the filter can be of any size. Here, each row of the connectivity matrix represents the correlation between the corresponding region and the other regions of the brain. Therefore, we considered the width of the filter as the dimension of the corresponding region and equal to the size of each row of the connectivity matrix, which is equal to 392. The length of the filter is its number of rows. Choosing filters of larger sizes did not increase the accuracy of the result. The applied CNN model does not use common feed-forward convolution. In our proposed architecture, we concatenated several convolution layers, and the entire obtained result set was given to the MLP for classification. The filter size of 1 × 392 in the convolution layer means that the connection of each area with other areas will be seen, and the filter size of 4 × 392 means the connection of four areas near each other with other areas will be seen, and at the end, we combine these outputs to get the final output. The execution time for this work was about 12 h and 30 min using 10-fold cross-validation with the NVIDIA Tesla K80 model GPU. We achieved an accuracy of 70.22 %, which is better than the rest of the reported works (**Table 4**). The receiver operating characteristic curve (ROC) and the confusion matrix for our CNN model are shown in **Figure 3**.

Thus, to the best of our knowledge, the approach that has been proposed in this paper has obtained the best accuracy so far achieved using the ABIDE I dataset. **Table 5** compares the automated detection of TC and ASD classes achieved by different studies using the same database. It can be seen from the comparison table that we have obtained better results compared to the other state-of-art techniques.

We evaluated the performance of SVM (support vector machine), KNN (K-nearest neighbors), and RF (random forest) classifiers on the preprocessed ABIDE I dataset. After optimization (hyperparameter tuning), the average accuracy was found to be 0.69 for SVM, 0.62 for KNN, and 0.6 RF. The results of the three approaches after being trained with 10-fold cross-validation are presented in **Table 6**. It can be seen that the CNN-based architecture outperformed these ML classifiers in terms of accuracy, specificity, and sensitivity.

The receiver operating characteristic curve (ROC) and confusion matrix are used the evaluate the performance of the SVM, KNN, and RF classifiers before and after optimization (hyperparameter tuning), as shown in **Figures 4**, **5**. Before

optimization, we used a radial basis function (RBF) kernel with regularization parameter C = 8 for the SVM classifier to obtain the optimum performance. We chose k = 20 for the KNN classifier. We set the max number of features (max_{depth}) and max number of levels in each decision tree ($n_{estimators}$) as 300 and 100, respectively, for the RF classifier. After optimization, we selected kernals such as "linear," "rbf," "poly," and "sigmoid" for SVM. We employed the grid search method for the KNN classifier and chosen optimization parameters of 4, 8, 12, 16, 20, 24, 28, 32, 36, and 40. The tuning (max_depth and $n_estimators$) of the RF classifier was optimized using the grid search method. In this case, the max_{depth} values were varied from 120 to 600 with a step size of 60 and $n_{estimators}$ values were varied from 20 to 180 with a step size of 20. The results show that, by optimizing the tuning parameters, the area under the ROC curve (AUC) will increase and hence, the classification performance is improved. These results are summarized in **Table 6**.

TABLE 7 | Summary of performance values obtained for 17 sites using our proposed CNN model.

Site out	Size	Accuracy	Confidence interval	Specificity	Sensitivity	F-score
CALTECH	37	0.54	0.16	0.42	0.66	0.58
CMU	27	0.70	0.17	0.71	0.69	0.69
KKI	48	0.72	0.12	0.95	0.57	0.71
LEUVEN	63	0.65	0.12	0.37	0.88	0.73
MAX_MUN	52	0.46	0.13	0.45	0.46	0.48
NYU	175	0.65	0.07	0.41	0.84	0.73
OHSU	26	0.57	0.19	0.66	0.5	0.56
OLIN	34	0.58	0.16	0.57	0.6	0.56
PITT	56	0.69	0.12	0.51	0.88	0.73
SBL	30	0.56	0.18	0.4	0.73	0.62
SDSU	36	0.75	0.14	0.64	0.81	0.8
STANFORD	39	0.48	0.16	0.94	0.05	0.09
TRINITY	47	0.61	0.14	0.63	0.6	0.62
UCLA	98	0.69	0.09	0.72	0.65	0.65
UM	140	0.66	0.08	0.95	0.4	0.56
USM	71	0.77	0.09	0.8	0.72	0.69
YALE	56	0.69	0.12	0.82	0.57	0.65
Mean	61	0.63	0.13	0.64	0.62	0.61

In order to evaluate the classifier performance for different sites, we used a leave-site-out approach for our proposed CNN, SVM, KNN, and RF classifiers (Heinsfeld et al., 2018). In this method, each site is taken as one fold in the dataset, and we applied a cross-validation approach on the remaining sites instead of different folds. Therefore, each time, one site out of 17 is used to test, and the other sites are used for training. We observe that the sites KKI, SDSU, and USM achieved accuracies of more than 70% as compared to other sites when using our proposed CNN model. The accuracy, confidence interval 95%, specificity, sensitivity, and F-score values for various sites are presented in **Table 7** and the accuracy and the confidence interval 95% are depicted in **Figure 6**. Also, a summary of the performance values obtained for each site using the SVM, KNN, and RF classifiers after optimization are given in **Tables 8–10**, respectively.

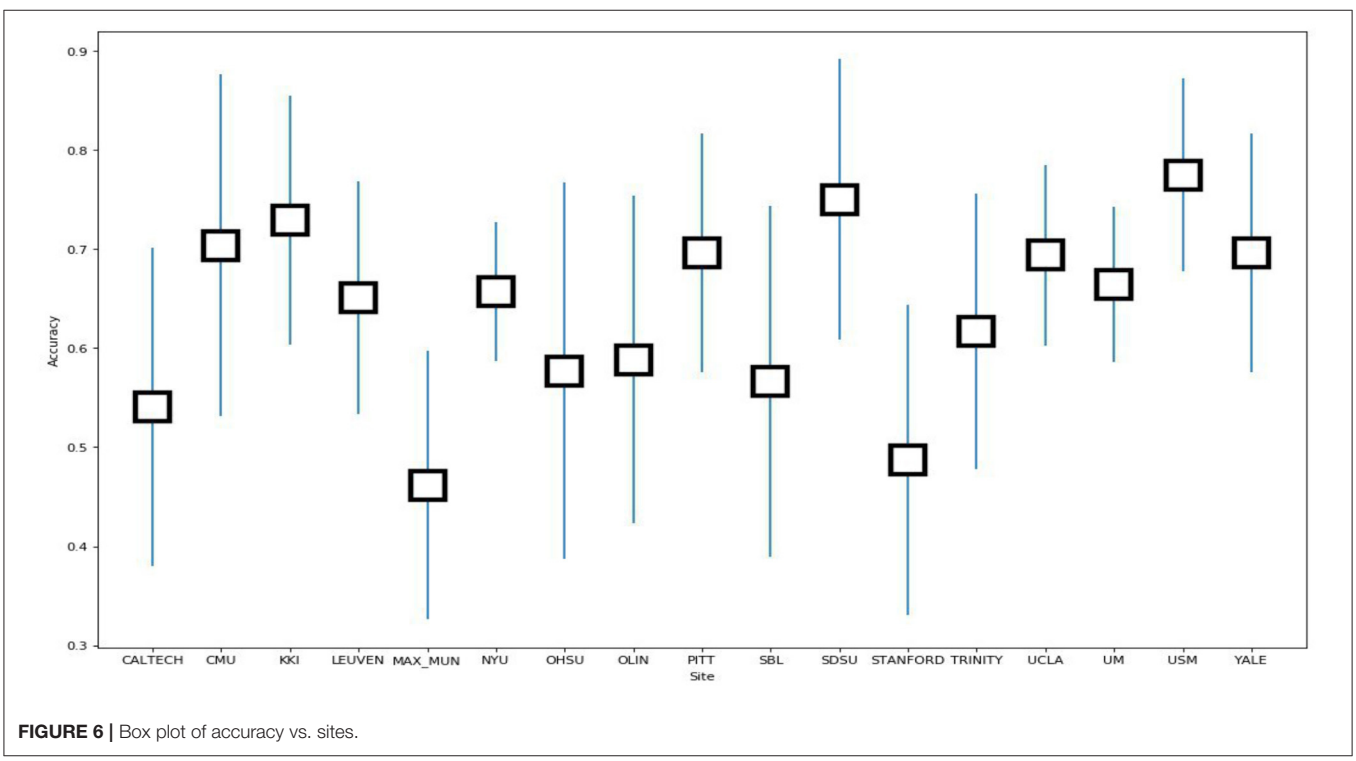


FIGURE 6 | Box plot of accuracy vs. sites.

TABLE 8 | Summary of performance values obtained for 17 sites using the SVM classifier after optimization.

Site out	Accuracy	Specificity	Sensitivity	F-score
CALTECH	0.70	0.89	0.50	0.62
CMU	0.74	0.71	0.77	0.74
KKI	0.75	0.95	0.61	0.74
LEUVEN	0.63	0.34	0.88	0.72
MAX_MUN	0.54	0.54	0.54	0.55
NYU	0.68	0.66	0.69	0.71
OHSU	0.73	0.58	0.86	0.77
OLIN	0.68	0.68	0.67	0.64
PITT	0.70	0.55	0.85	0.73
SBL	0.53	0.40	0.67	0.59
SDSU	0.72	0.64	0.77	0.77
STANFORD	0.61	0.94	0.30	0.44
TRINITY	0.57	0.72	0.44	0.52
UCLA	0.75	0.76	0.75	0.73
UM	0.76	0.80	0.72	0.76
USM	0.79	0.85	0.68	0.69
YALE	0.71	0.75	0.68	0.70
Mean	0.68	0.69	0.67	0.67

TABLE 9 | Summary of performance values obtained for 17 sites using the KNN classifier after optimization.

Site out	Accuracy	Specificity	Sensitivity	F-score
CALTECH	0.54	0.47	0.61	0.56
CMU	0.59	0.35	0.85	0.67
KKI	0.60	0.75	0.50	0.60
LEUVEN	0.62	0.28	0.91	0.72
MAX_MUN	0.50	0.50	0.50	0.52
NYU	0.59	0.40	0.74	0.68
OHSU	0.64	0.25	0.64	0.56
OLIN	0.68	0.68	0.66	0.64
PITT	0.58	0.41	0.78	0.65
SBL	0.57	0.33	0.80	0.65
SDSU	0.64	0.36	0.82	0.73
STANFORD	0.51	0.0.58	0.45	0.49
TRINITY	0.60	0.41	0.76	0.67
UCLA	0.63	0.48	0.82	0.67
UM	0.57	0.98	0.20	0.33
USM	0.46	0.20	0.96	0.56
YALE	0.64	0.43	0.86	0.70
Mean	0.58	0.46	0.70	0.61

TABLE 10 | Summary of performance values obtained for 17 sites using the RF classifier after optimization.

Site out	Accuracy	Specificity	Sensitivity	F-score
CALTECH	0.48	0.52	0.44	0.45
CMU	0.66	0.35	1.00	0.74
KKI	0.64	0.90	0.46	0.60
LEUVEN	0.63	0.27	0.94	0.73
MAX_MUN	0.56	0.42	0.67	0.62
NYU	0.68	0.50	0.82	0.75
OHSU	0.50	0.58	0.43	0.48
OLIN	0.64	0.63	0.67	0.62
PITT	0.67	0.48	0.89	0.73
SBL	0.60	0.33	0.87	0.68
SDSU	0.63	0.71	0.60	0.67
STANFORD	0.51	0.89	0.15	0.24
TRINITY	0.64	0.59	0.68	0.67
UCLA	0.61	0.46	0.79	0.65
UM	0.67	0.89	0.47	0.60
USM	0.70	0.65	0.80	0.65
YALE	0.66	0.53	0.79	0.70
Mean	0.62	0.57	0.67	0.62

5. DISCUSSION AND CONCLUSIONS

In the present study, we proposed a CNN architecture to identify and classify ASD patients and control subjects. Also, the performance of three supervised learning methods, SVM, KNN, and RF classifiers, on the preprocessed ABIDE I dataset was investigated. The results show that the average accuracy of our model using the test data is 70.2%, meaning that it outperformed the best accuracy obtained on this dataset so far. It has been observed that for the same accuracy, a CNN model with fewer parameters is more efficient and has less overhead for the new models (Iandola et al., 2016). Keeping this in mind, our model is able to train with fewer parameters and achieve an even better accuracy level than the best-performing models. The existing best-known method used a huge number of parameters (19, 961, 200) in its final stage, but our model used 4,398,802 parameters. The authors of Xing et al. (2018) used 1,268,160 parameters and obtained an accuracy of 66.88%, but we achieved an accuracy of 70.20%. Hence, our proposed CNN architecture is able to obtain higher classification performance with fewer parameters, which will reduce the training time. Therefore, our proposed model is less complex and faster as compared to other similar models. Also, we studied each row of the connectivity matrix as the representation of the correlation between the corresponding region and the other regions of the brain in our model.

Thus, we open up the possibility to illustrate the behavior of a region of the brain and corresponding biomarkers by performing a noise correction on each row of the connectivity matrix in future work.

The future recommendations for our proposed model are given below: 1. We have used few images in each class. There is a need to use more data to build a more robust model. 2. The time complexity of the model should be decreased when the whole dataset of all subjects are fed into it. 3. The impact of two features (sex and average age) need to be

considered in this study. 4. The performance may improve with balanced data.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

ZS, MAk, SS, MZ-M, and MAb have equal contributions in data preparation, data analysis, and preparing the first draft of the manuscript. UA improved the results and revised the text of the manuscript. RK and VS revised the results and prepared the final version.

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirio, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

Aghdam, M. A., Sharifi, A., and Pedram, M. M. (2018). Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network. *J. Digit. Imaging* 31, 895–903. doi: 10.1007/s10278-018-0093-8

American Psychiatric Association (2011). *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. American Psychiatric Pub.

Anderson, J. S., Nielsen, J. A., Froehlich, A. L., DuBray, M. B., Druzgal, T. J., Cariello, A. N., et al. (2011). Functional connectivity magnetic resonance imaging classification of autism. *Brain* 134(Pt 12), 3742–3754. doi: 10.1093/brain/awr263

Anirudh, R., and Thiagarajan, J. J. (2017). Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. *arXiv:1704.07487 [stat.ML]*.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. (2007). A component based noise correction method (CompCor) for bold and perfusion based fMRI. *Neuroimage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Bhat S., Acharya U. R., Adeli H., Bairy G. M., and Adeli, A. (2014b). Automated diagnosis of autism: in search of a mathematical marker. *Rev. Neurosci.* 25, 813–823. doi: 10.1515/revneuro-2014-003

Bhat S., Acharya U. R., Adeli H., Bairy, G. M., and Adeli, A. (2014a). Autism: cause factors, early diagnosis and therapies. *Rev. Neurosci*. 25, 841–850. doi: 10.1515/revneuro-2014-0056

Bourgeron T. (2009). A synaptic trek to autism. *Curr. Opin. Neurobiol.* 19, 231–234. doi: 10.1016/j.conb.2009.06.003

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., et al. (2011). Classification of alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology* 259, 213–221. doi: 10.1148/radiol.10100734

Cherkassky, V. L., Kana, R. K., Keller, T. A., and Just, M. A. (2006) Functional connectivity in a baseline resting-state network in autism. *NeuroReport* 17, 1687–1690. doi: 10.1097/01.wnr.0000239956.45448.4c

Choi, H. (2017). Functional connectivity patterns of autism spectrum disorder identified by deep feature learning. *arXiv:1707.07932 [cs.CV]*.

Di Martino, A., Yan, C. G., Li, Q., Castellanos, F. X., Alaerts, K., Anderson, J. S., et al. (2014). The autism brain imaging data exchange: towards a largescale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78

Dvornek, N. C., Ventola, P., Pelphrey, K. A., and Duncan, J. S. (2017). "Identifying autism from resting-state fMRI using long short-term memory networks," in *International Workshop on Machine Learning in Medical Imaging* (Cham: Springer), 362–370.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D., and Raichle, M. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Heinsfeld, A. S., Rosa Franco, A., Cameron Craddock, R., Buchweitz, A., and Meneguzzia, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hoffmann F., Koehne S., Steinbeis N., Dziobek I., and Singer, T. (2015). Preserved self-other distinction during empathy in autism is linked to network integrity of right supramarginal gyrus. *J. Autism Dev. Disord.* 46, 637–648. doi: 10.1007/s10803-015-2609-0

Huang B., Zhang K., Sanchez-Romero R., Ramsey J., Glymour M., and Glymour C. (2019). Diagnosis of autism spectrum disorder by causal influence strength learned from resting-state fMRI data. *arXiv:1902.10073 [q-bio.NC]*.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv:1602.07360*.

Just, M. A. (2004) Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain* 127, 1811–1821. doi: 10.1093/brain/awh199

Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., and Just, M. A. (2009). A typical fronto-posterior synchronization of theroy of mind regions in autism during mental state attribution. *Soc. Neurosci.* 4, 135–152. doi: 10.1080/17470910802198510

Kaufmann, W. E., Cooper, K. L., Mostofsky, S. H., Capone, G. T., Kates, W. R., Newschaffer, C. J., et al. (2003). Specificity of cerebellar vermian abnormalities in autism: a quantitative magnetic resonance imaging study. *J. Child Neurol.* 18, 463–470. doi: 10.1177/08830738030180070501

Kazeminejad, A., and Sotero, R. (2019). The importance of anti-correlations in graph theory based classification of autism spectrum disorder. *bioRxiv*. 557512. doi: 10.1101/557512

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2018). Ensemble learning with 3D convolutional neural networks for connectome-based prediction. *arXiv:1809.06219 [cs.CV]*. doi: 10.1016/j.neuroimage.2019.06.012

Koyamada, S., Shikauchi, Y., Nakae, K., Koyama, M., and Ishii, S. (2015). Deep learning of fMRI big data: a novel approach to subject-transfer decoding. *arXiv:1502.00093 [stat.ML]*.

Li, G., Liu, M., Sun, Q., Shen, D., and Wang, L. (2018a). Early diagnosis of autism disease by multi-channel CNNs. *Mach. Learn. Med. Imaging* 11046, 303–309. doi: 10.1007/978-3-030-00919-9_35

Li, X., Dvornek, N. C., Zhuang, J., Ventola, P., and Duncan, J. S. (2018b). Brain biomarker interpretation in ASD using deep learning and fMRI. *arXiv:1808.08296 [cs.CV]*. doi: 10.1007/978-3-030-00931-1_24

Lund, T. E., Nrgaard, M. D., Rostrup, E., Rowe, J. B., and Paulson, O. B. (2005). Motion or activity: their role in intra-and inter-subject variation in fMRI. *Neuroimage* 26, 960–964. doi: 10.1016/j.neuroimage.2005.02.021

Mennes, M., Vega Potler, N., Kelly, C., Di Martino, A., Castel-lanos, F. X., and Milham, M. P. (2011). Resting state functional connectivity correlates of inhibitory control in children with attention-deficit/hyperactivity disorder. *Front. Psychiatry*. 2:83. doi: 10.3389/fpsyt.2011.00083

Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., et al. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. *Front. Hum. Neurosci.* 7:599. doi: 10.3389/fnhum.2013.00599

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., et al. (2017). "Spectral graph convolutions for population-based disease prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 177–185.

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., et al. (2018). Disease prediction using graph convolutional networks: application

to Autism Spectrum Disorder and Alzheimer's disease. *Med. Image Anal.* 48, 117–130. doi: 10.1016/j.media.2018.06.001

Plitt, M., Barnes, K. A., and Martin, A. (2015). Functional connectivity classification of autism identifes highly predictive brain features but falls short of biomarker standards. *Neuroimage Clin.* 7, 359–366. doi: 10.1016/j.nicl.2014.12.013

Rad, M. N., Bizzego, A., Kia, S. M., Jurman, G., Venuti, P., and Furlanello, C. (2015). Convolutional neural network for stereotypical motor movement detection in autism. *arXiv:1511.01865 [cs.NE].* doi: 10.1109/ICDMW.2016.0076

Schipul, S. E., Keller, T. A., and Just, M. A. (2011). Inter-regional brain communication and its disturbance in autism. *Front. Syst. Neurosci.* 5:10. doi: 10.3389/fnsys.2011.00010

Sharif, H., and Khan, R. A. (2019). A novel framework for automatic detection of Autism: a study on Corpus Callosum and Intracranial Brain Volume. *arXiv:1903.11323.*

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv:1312.6034.*

Subbaraju, V., Suresh, M. B., Sundaram, S., and Narasimhan, S. (2017). Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging : a spatial filtering approach. *Med. Image Anal.* 35, 375–389. doi: 10.1016/j.media.2016.08.003

van Kooten, I. A., Palmen, S. J., von Cappeln, P., Steinbusch, H. W., Korr, H., Heinsen, H., et al. (2008). Neurons in the fusiform gyrus are fewer and smaller in autism. *Brain* 131, 987–999. doi: 10.1093/brain/awn033

von dem Hagen, E. A., Stoyanova, R. S., Baron-Cohen, S., and Calder, A. J. (2013). Reduced functional connectivity within and between 'social' resting state networks in autism spectrum conditions. *Soc. Cogn. Affect. Neurosci.* 8, 694–701. doi: 10.1093/scan/nss053

Wing, L. The autistic spectrum. *Lancet* (1997) 350, 1761–1766.

Xing, X., Ji, J., and Yao, Y. (2018). "Convolutional neural network with element-wise filters to extract hierarchical topological features for brain networks," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Madrid), 780–783.

Zeng, L.-L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., et al. (2018). Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 30, 74–85. doi: 10.1016/j.ebiom.2018.03.017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sherkatghanad, Akhondzadeh, Salari, Zomorodi-Moghadam, Abdar, Acharya, Khosrowabadi and Salari. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 17 January 2020
doi: 10.3389/fnins.2019.01435

Detecting the Information of Functional Connectivity Networks in Normal Aging Using Deep Learning From a Big Data Perspective

Xin Wen[1], Li Dong[2,3], Junjie Chen[1]*, Jie Xiang[1], Jie Yang[3], Hechun Li[3], Xiaobo Liu[3], Cheng Luo[2,3] and Dezhong Yao[2,3]*

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, China, [2] The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China, [3] School of Life Sciences and Technology, University of Electronic Science and Technology of China, Chengdu, China

A resting-state functional connectivity (rsFC)-constructed functional network (FN) derived from functional magnetic resonance imaging (fMRI) data can effectively mine alterations in brain function during aging due to the non-invasive and effective advantages of fMRI. With global health research focusing on aging, several open fMRI datasets have been made available that combine deep learning with big data and are a new, promising trend and open issue for brain information detection in fMRI studies of brain aging. In this study, we proposed a new method based on deep learning from the perspective of big data, named Deep neural network (DNN) with Autoencoder (AE) pretrained Functional connectivity Analysis (DAFA), to deeply mine the important functional connectivity changes in fMRI during brain aging. First, using resting-state fMRI data from 421 subjects from the CamCAN dataset, functional connectivities were calculated using sliding window method, and the complex functional patterns were mined by an AE. Then, to increase the statistical power and reliability of the results, we used an AE-pretrained DNN to relabel the functional connectivities of each subject to classify them as belonging to the attributes of young or old individuals. A method called search-back analysis was performed to find alterations in brain function during aging according to the relabeled functional connectivities. Finally, behavioral data regarding fluid intelligence and response time were used to verify the revealed functional changes. Compared to traditional methods, DAFA revealed additional, important aged-related changes in FC patterns [e.g., FC connections within the default mode (DMN) and the sensorimotor and cingulo-opercular networks, as well as connections between the frontoparietal and cingulo-opercular networks, between the DMN and the frontoparietal/cingulo-opercular/sensorimotor/occipital/cerebellum networks, and between the sensorimotor and frontoparietal/cingulo-opercular networks], which were correlated to behavioral data. These findings demonstrated that the proposed DAFA method was superior to traditional FC-determining methods in discovering changes in brain functional connectivity during aging. In addition, it may be a promising method for exploring important information in other fMRI studies.

Keywords: deep learning, big data, DNN, functional connectivity, aging

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Martin J. McKeown,
The University of British Columbia, Canada
Shijie Zhao,
Northwestern Polytechnical University, China

***Correspondence:**
Li Dong
Lidong@uestc.edu.cn
Junjie Chen
chenjj@tyut.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *13 August 2019*
Accepted: *19 December 2019*
Published: *17 January 2020*

Citation:
Wen X, Dong L, Chen J, Xiang J, Yang J, Li H, Liu X, Luo C and Yao D (2020) Detecting the Information of Functional Connectivity Networks in Normal Aging Using Deep Learning From a Big Data Perspective. Front. Neurosci. 13:1435. doi: 10.3389/fnins.2019.01435

INTRODUCTION

Functional networks (FNs) constructed by resting-state functional connectivity (rSFC) analysis using data from blood oxygen level dependence (BOLD)-based functional magnetic resonance imaging (fMRI) have greatly deepened our understanding of the functioning of the human brain. During the resting-state, time series from the same FN are temporally correlated with each other, and the FNs thus reflect functional communication between brain regions. Due to its non-invasiveness and effectiveness, resting-state FC analysis has become the most extensive and important tool for exploring changes in brain function (Biswal et al., 1995; De Luca et al., 2006). With an increasing global emphasis on human health, brain aging has become a research hotspot in brain science. Functional connectivity analysis based on rs-fMRI has been widely applied in brain aging studies (Ferreira and Busatto, 2013). These studies have indicated that the strength of multiple FNs, including the default mode network (DMN) and dorsal attentional network (DAN), decreases during the aging process. (Sheffield et al., 2015; Avelar-Pereira et al., 2017; Vij et al., 2018), the phenomenon of which indicates that attention, memory and executive control functions decline in cognitive aging. Meanwhile, researchers have also found that increased FC involving the frontal, parietal networks and motor, subcortical networks was related to aging, which may reflect the compensatory responses to the decreased strength of FC during aging (Biswal et al., 2010; Tomasi and Volkow, 2012; Ferreira and Busatto, 2013). These results have greatly enhanced our understanding of aging in the brain; functional connectivity analysis is expected to be a valuable tool for objectively evaluating the health status of the brain in the function and cognition of elderly individuals and thus may benefit the clinical diagnosis and intervention of brain aging-related diseases (Geerligs et al., 2017; de Vos et al., 2018). However, most of the present studies primarily rely on traditional, relatively small samples (from 10 to a few 100 samples) based on statistical analyses (e.g., two-sample *t*-test and multiple linear regression), which may result in insufficient analysis of resting-state FCs during brain aging, thus ignoring some important underlying information regarding FN alterations.

Machine learning is a set of algorithms used for the automatic development of models and learning of complex patterns from data. Over the past decade, machine learning has achieved great success in mining functional connectivity. For example, Calhoun et al. performed independent component analysis (ICA) to decompose fMRI data into reduced-dimensional time series and spatial patterns, which were further used for modal fusion and classification prediction (Calhoun and Sui, 2016). Du et al. (2017) proposed a new scheme of group information-oriented ICAs to mine dynamic functional connectivity from the fMRI data of patients with various mental disorders, detecting the differences of brain function between groups related to diseases, which were not found by traditional static methods. However, traditional linear machine learning algorithms rely on feature extraction, which cannot handle non-linear and complex relationships in data, and lack the ability to process FCs directly as well (Plis et al., 2014). Therefore, deep learning methods developed based on machine learning is gradually attracting the attention of neuroscience researchers. Accumulating evidence shows that deep learning is superior to traditional machine learning with regard to fMRI-related pattern recognition (Hu et al., 2018), dimensionality reduction (Suk et al., 2016; Kawahara et al., 2017), classification (Kim et al., 2016; Guo et al., 2017), and prediction (Chae et al., 2018; Khosla et al., 2019). For example, in terms of non-linear pattern recognition, Suk et al. (2016) applied auto-encoders to extract the dynamic changes in brain function from a time series of fMRI. These changes have significant sensitivity in distinguishing individuals with mild cognitive impairment from healthy people. For classification, Kim et al. (2016) used a DNN to mine the resting-state FNs with physiological significance, utilized mining features to classify schizophrenic patients and health controls, and obtained an accuracy of classification that was 22.3% higher than that of a support vector machine (SVM). However, the number of data samples used in the training models of these studies was still relatively small (<100), and so it is possible that these methods missed some important information due to inadequate learning (Xia and He, 2017).

With the continuous deepening of collaborative research on global brain aging, research based on open data sets [e.g., Cambridge Centre for Ageing and Neuroscience (CamCAN) dataset[1]] makes it possible to study brain aging from a big data perspective. The increase in open data set samples for analysis likely enhances the effectiveness of statistical analysis and facilitates new research designs (Peter and Jayati, 2018; Smith and Nichols, 2018). However, traditional statistics with big data may lead to possible confounding effects (Smith and Nichols, 2018), thus requiring the development of new data analysis methods in the context of big data. Considering the superiority of deep learning and big data analysis to traditional modeling methods, combining deep learning with neuroimaging big data is expected to deeply explore brain aging information by taking into account the increase in sample size and the complex relationships in the data.

Spontaneous brain activity during resting-state is an apparent variability of interaction between brain regions and may be dominated by traces of activity, which may involve different subregions in a network at different times (Liu and Duyn, 2013; Karahanoglu and Van De Ville, 2015). Furthermore, functional connectivity, usually measured by temporal correlations between brain regions, can reflect these functional interactions and yield details maps of complex neural systems (Biswal et al., 1995, 2010). Therefore, we proposed the hypothesis that most of the functional activities in young adults may reflect the "young FCs" (FCs which reflect intrinsic brain activities in young adults), while those in old adults may largely reflect the "old FCs" (FCs which reflect intrinsic brain activities in old adults). Under this hypothesis, traditional static FC method may miss some important information in comparing functional connectivity differences between young and old groups. Meanwhile, previous studies implied that deep learning methods (especially DNN) were superior to traditional machine

[1]http://www.cam-can.com

learning methods (Kim et al., 2016; Hu et al., 2018) both for classification and mining underlying information in fMRI data. Therefore, a new method based on the above-mentioned assumption is expected to detect the information of functional connectivity networks in normal aging. In this work, using deep learning from a big-data perspective, we proposed a new analysis method, named DNN with AE pretrained Functional connectivity Analysis (DAFA), to deeply mine the important functional connectivity changes in fMRI during brain aging. A sliding window method was used to increase the sample size of FCs for each subject by ∼100 times (∼hundred thousand) in order to meet the requirements of deep learning from a big data perspective to a certain extent. After calculating the resting state FCs of each subject at different time windows, the FCs are first fed to DAFA to train the AE-pretrained DNN model with a Softmax layer as a classifier to relabel all FC samples to "young FCs" or "old FCs." Then, according to the predicted labels, we compared the percentages of "young FCs" between young and old groups using a two-sample *t*-test and further investigated the differences between the mean "young FCs" in the young group and the mean "old FCs" in the old group.

MATERIALS AND METHODS

Participants

In this study, as shown in **Table 1**, fMRI data was obtained from a total of 412 individuals (after quality control, e.g., excluding excessive head motion, missing T1 images, etc.), which included 220 young adults whose ages ranged from 18 to 45 and 192 elderly adults whose ages ranged from 66 to 88. The binary young/old classes were based on the published paper of Cam-CAN open dataset (Shafto et al., 2017). In each group, the number of participants was approximately equal when divided into 10 year-wide bins. The neuroimaging data used were a subset of the Cambridge Centre for Ageing and Neuroscience (see text footnote 1) (Taylor et al., 2017). Participants performed cognitive tasks outside the MRI scanner. The tests used in this study were the Cattell Culture Fair test, used to assess fluid intelligence (Horn and Cattell, 1966), and the speed Choice Reaction Time (RT) task, used to assess speed of processing. For the RT tasks, the mean (M-RT) and variability (SD of RT values, SD-RT) were computed from individual trials. Ethical approval for the study was obtained from the Cambridgeshire 2 (now East of England – Cambridge Central) Research Ethics Committee. All participants gave written informed consent.

TABLE 1 | Participants' demographic information.

	Group	Number	Range	Mean	SD
Age*	Young	220	18–45	30.25	5.82
	Old	192	66–88	75.29	6.17
Cattell*	Young	213	22–44	36.59	4.35
	Old	188	11–39	26.08	6.04
M-RT*	Young	215	0.27–0.39	0.31	0.05
	Old	140	0.29–0.48	0.34	0.07
SD-RT*	Young	215	0.07–0.17	0.12	0.02
	Old	140	0.01–0.30	0.08	0.04

**indicates a significant difference between groups using two-sample t-test, $p < 0.001$.*

fMRI Acquisition and Preprocessing

Imaging data were collected by a 3 T MRI scanner (Siemens TIM Trio). During the scanning period, all participants were required to lie still and keep their eyes closed. The T1-weighted anatomical images were gathered using an MPRAGE sequence with the following parameters: TR/TE = 2250 ms/2.99 ms; flip angle = 9°; FOV = 256 × 240 × 192 mm^3; voxel size = 1 × 1 × 1 mm^3. For rs-fMRI measurements, 261 volumes of echo planar imaging (EPI) sequences were acquired with the following parameters: sequential descending order; slice thickness 3.7 mm with a slice gap of 20% for whole-brain coverage; TR/TE = 1970 ms/30 ms; flip angle = 78°; FOV = 192 × 192 mm^2; voxel size = 3 × 3 × 4.44 mm^3; number of slices = 32; duration time = 520 s.

To remove T1 saturation effects, the first five volumes were deleted from the resting-state fMRI data of each subject. Then, SPM12[2] and NIT http://www.neuro.uestc.edu.cn/NIT.html (Dong et al., 2018) were used to preprocess the resting-state fMRI data. The fMRI preprocessing contained the following steps: realignment, slice time correction, spatial normalization using T1-weighted MRI data (3 × 3 × 3 mm^3) and smoothing [6-mm full-width at half-maximum (FWHM)]. Nuisance noises such as linear trends, 12 head-motion parameters, global signals, and individual mean WM and CSF signals were removed using multiple linear regression analysis, and temporal bandpass filtering (pass band 0.01–0.08 Hz) was conducted on the fMRI data. The head motion of each participant was calculated using the mean framewise displacement (mean FD) (Power et al., 2012). Participants whose FD was two or more SD above the group mean FD were excluded from further analysis.

Method Overview

In this paper, a method named DAFA was used to locate significant differences in functional connectivity between the old and young groups. **Figure 1** shows the workflow of the proposed procedure.

The entire procedure consisted of three steps. (A), whole-brain functional connectivity patterns based on sliding windows (window length: 50 time points, window step: 1 time point) were calculated by the Pearson's correlation coefficient of the time courses from every pair of 160 ROIs (Dosenbach et al., 2010), followed by Fisher's r-to-z transformation. (B), an autoencoder was trained on the whole FC dataset using gradient descent learning with L_2 norm regularization. Then, the weights learned from the AE were set as an initialization of the parameters of the bottom layer to a DNN with a Softmax classifier as the top layer. DNN was used to relabel the FC samples to deeply mine the important functional connectivity changes between young and old groups (which may be not well captured with conventional method), under the condition of mixed labels. FCs in the dataset

[2]http://www.fil.ion.ucl.ac.uk/spm

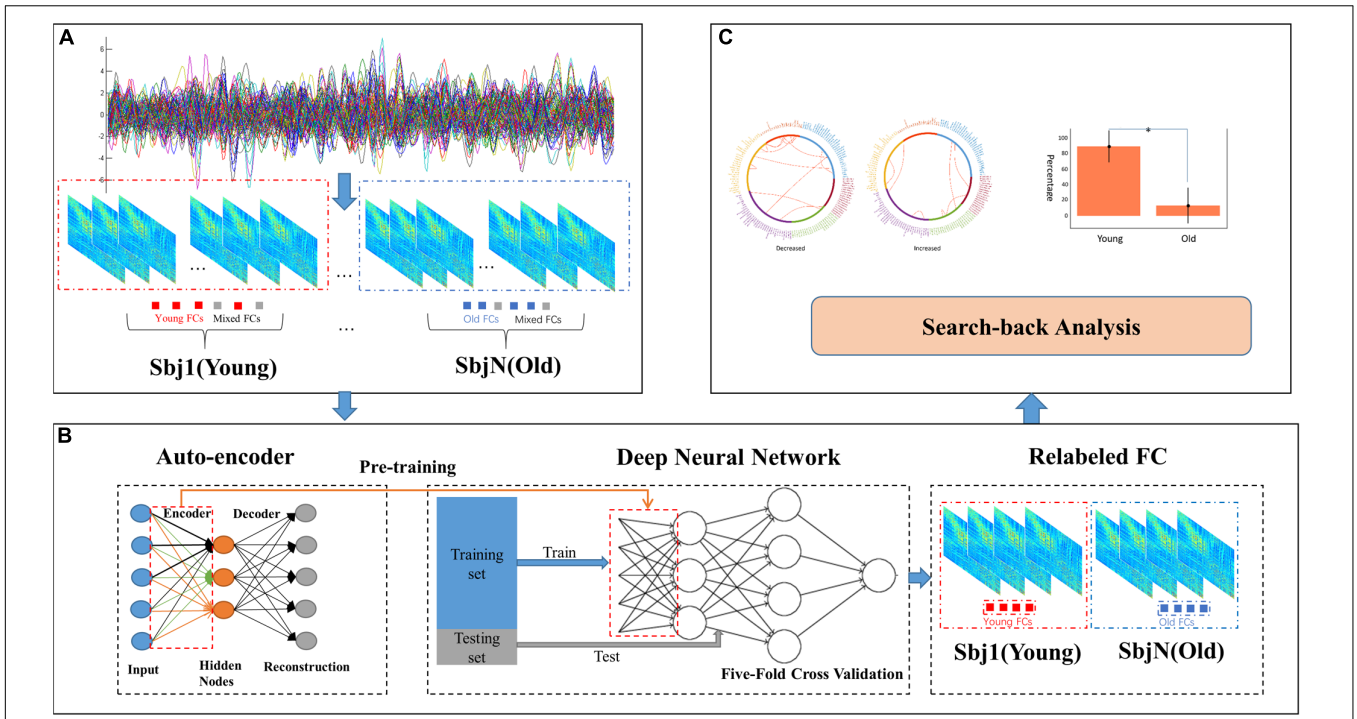


FIGURE 1 | Workflow of the DAFA method. **(A)** Constructing FCs using sliding windows and Pearson's correlation methods, all the FCs from a subject contains most of "young FCs"/"old FCs" and mix FCs. **(B)** Relabeling all FCs via a DNN based on AE pretraining, and then calculating the average FC of "young FCs" in young subject and "old FCs" in old subject as this subject's FC. **(C)** Performing a search-back analysis to compare the percentages of "young FCs" and functional alterations between the young and old groups.

were labeled as "young FCs" or "old FCs" by the DNN with five-fold cross validation (the cross-validation procedure was performed on subjects and each subject contained 212 FCs via sliding window method, FCs from a subject were either divided to training set or testing set). (C), a search-back analysis procedure was performed to reveal the functional alterations between the mean young FC in the young group and the mean old FC in the old group and the differences between the percentages of "young FCs" labeled in the young group and the old group.

Autoencoder

The architecture of an autoencoder is a three-layer feed-forward neuro network that contains an input layer, a hidden layer, and an output layer. All three layers are fully connected hierarchically through weighted connections that are updated by a back-propagation algorithm. An important property of the AE's structure is that the dimensions of the input layer and output layer are the same, while the dimension of the hidden layer is much smaller so that the output layer is forced to reconstruct the pattern of the input layer with the smaller size of the hidden nodes, the value in each hidden node could represent a low-dimension feature from the input data, and the whole AE could be interpreted as a dimension-reduction function.

The basic principle behind training the AE is the minimizing of the residual error between the values of the input and output layers. Let $X = [x_1, x_2, x_3, \ldots, x_d]$ denote all the FCs for the entire dataset, $y = [y_1, y_2, y_3, \ldots, y_p]$ denote the features that represent $\times$ after dimension reduction, and $z = [z_1, z_2, z_3, \ldots, z_d]$ denote the corresponding reconstructed data. The number of input and output layer nodes is indicated by d, whose size is equal to the number of observations, and the number of hidden nodes is denoted by p $(p < d)$; the weight and bias of the encoder and decoder can be denoted by $w^{(1,0)} \in R^{(p \times d)}$, $b^{(1,0)} \in R^p$ and $w^{(2,1)} \in R^{(d \times \rho)}$, $b^{(2,1)} \in R^d$. The features are computed by $y = f\left(w^{(1,0)}x + b^{(1,0)}\right)$ where $f(x)$ is a sigmoid function of the hidden nodes; the reconstructed data are computed by $z = g\left(W^{(2,1)}x + b^{(2,1)}\right)$, where $g(x)$ is the tanh function of the output nodes. In this paper, we chose autoencoders to perform dimension reduction, and the weights were transferred to the DNN model for pretraining. In every training iteration, a set of edges generated by the Pearson's correlation of time courses in a window from a sample were fed to the AE, following which the AE was trained using a back-propagation algorithm. After the AE was trained from the whole dataset, the cost function was converged, which ensures that the average reconstruction error through the training set was minimized, resulting in an AE that learned the essential features in a relatively low dimension (from 12700 to 100).

Relabeling FCs via Training the DNN Model

The DNN model consisted of two hidden layers and a Softmax layer as the output layer. The sliding window-generated FCs of

the young group were labeled with $[1, 0]^T$, and those of the old group were labeled with $[0, 1]^T$. The cost function of the DNN for the supervised fine-tuning step was defined using the cross-entropy loss function, L_2 norm regularization term as follows:

$$J(W) = \sum_{i=1}^{N} t^i \log y^i W + (1 - t^i) \log (1 - y^i) + \frac{1}{2}\lambda^{(J-1,\ J)} \sum_{j=0}^{l} \left|\left| W^{(J+1,\ J)} \right|\right|^2 \quad (1)$$

where $y^i W$ is a vector with elements of the output values from the Softmax layer for subject n in the training set, t^i is the target output value of the window of the i^{th} subject, $\lambda^{(J-1,\ J)}$ is the parameter of the L_2 norm regularization term between the $j^{th}$and $(j+1)^{th}$ layer, N is the total number of samples in the training set, and $l+1$ is the number of layers, including the input layer. The learning rate was initialized to 0.0015, and the parameter of the L_2 norm was set to 10^{-6} to prevent overfitting. A total of 50 epochs were used, and the number of hidden nodes in layers was set to [100:50:2]. A momentum term was added to the current weight update term to accelerate the learning procedure. Additionally, weights between the input layer and the first hidden layer of the DNN were initialized as weights obtained by the AE and were frozen during the fine-tuning process (pretrain). In this work, we used five-fold cross validation to prevent overfitting. Finally, the predicted labels from each window for each of the subjects in the dataset were obtained (relabeling).

RESULTS

Autoencoder Parameter Settings

The input size of our autoencoder was 12720, the number of edges in a subject's functional connectivity based on 160 ROIs. AE is an unsupervised learning model, and the hyper parameters are essential for the performance in an AE. The purpose of adjusting the value of the parameters is to minimize the reconstructed error. In addition, to avoid identical transformations and enhance the learning ability of the AE, the activation functions used by the encoder and the decoder were the sigmoid function and the tanh function, respectively. As mentioned in the see section "Materials and Methods," the learning rate and parameter of the L_2 norm were 0.0005 and 10^{-5}, respectively. As shown in **Figure 2**, the loss increases with a larger batch size; in this study, a batch size of 50 was chosen, as this was the location of the inflection point (**Figure 2A**); then, the number of epochs was set to 20, corresponding to the first minimum loss value (**Figure 2B**); in order to minimize the reconstructed error while avoiding redundant features, the number of hidden nodes was empirically set to 100. The initialization of these hyper parameters and learned parameters could best reduce the reconstructed error both in terms of the empirical setup and fine-tuning.

"Young FCs" in the Young and Old Groups

The first column of **Figure 3** illustrates the average functional connectivity patterns in both the young and old groups using three methods: static, the mean FC of the sliding window and DAFA.

Using the static, sliding window and DAFA methods (the second column of **Figure 2**), the distributions of "young FCs" were different in both the young and old groups. In the static case, one subject had one static FC, so the percentage of "young FCs" is meaningless. In the sliding window case, all the FCs of an individual subject were classified the same way as that subject's group (e.g., individuals in the young group only had "young FCs"). When switching to the proposed DAFA, the percentage of "young FCs" in each sample varied on an individual basis.

To evaluate our assumption, the FCs belonging to the young group ("young FCs") were relabeled using five-fold CV based on each subject. For DAFA, the average percentage of "young FCs" in the young group was 89.8% of the 212 FCs, and the average percentage in the old group was 13.2%. A two-sample t-test revealed a significant ($T = 27.1$, $p = 1.5 \times 10^{-92}$) difference between the percentage of "young FCs" in the young group and in old group after controlling for the covariant variables of gender, head motion parameter (mean FD) and intracranial volume.

Differences Between "Young and Old FCs"

As illustrated in **Figure 4**, a two-sample t-test was conducted on young and old groups to demonstrate the significant connections among all ROIs [$p < 0.01$, familywise error rate (FWE) corrected]. When comparing the results between the two groups, the FCs revealed by DAFA were similar to those from the static method. In detail, the connections within the DMN and the frontoparietal and cingulo-opercular networks were decreased, as well as connections between the frontoparietal and cingulo-opercular networks. Meanwhile, the connections within the sensorimotor network were increased, as well as the connections between the DMN and the other five networks and between the sensorimotor and frontoparietal/cingulo-opercular networks. In addition, both increased and decreased connections between the cingulo-opercular and cerebellum networks were found.

For the sliding window method, the group FC differences were similar to those of the static method, with the addition of altered connections between the cingulo-opercular and DMN and cerebellum networks. Further, compared with the static method, additional alterations in FC patterns ($p < 0.01$, FWE corrected) between the young and old groups were found with the DAFA method. In detail, decreased FCs were mainly within the cingulo-opercular networks, as well as in the connections between the cingulo-opercular and occipital/cerebellum networks. Increased FCs were mainly found in the connections between the DMN and the cingulo-opercular/occipital/cerebellum networks. Among these additional changed FCs, the main nodes were located in the postcingulate/precuneus/anterior cingulate cortex areas for the

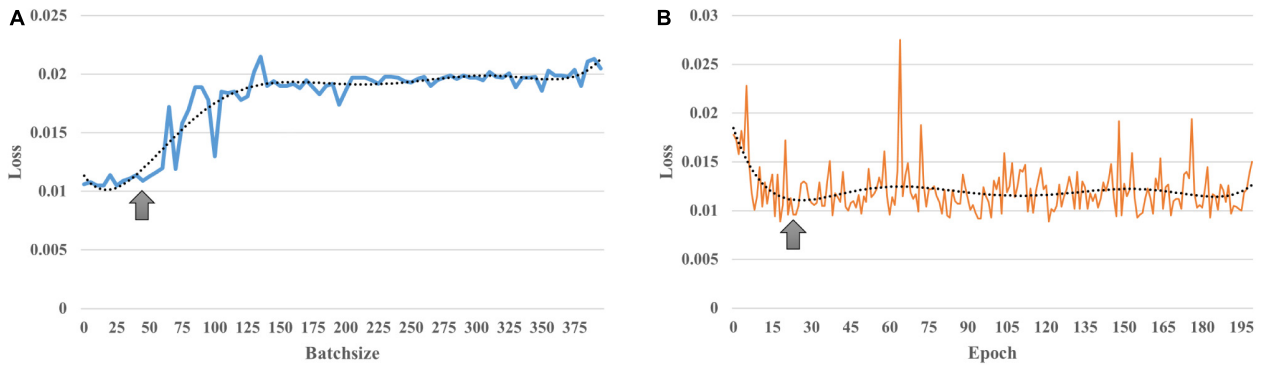


FIGURE 2 | Learnt parameters. **(A)** As the batch size increases, loss increases; **(B)** when the number of epochs = 20, loss reached its first minimum value and then changed periodically with increasing number of epochs.

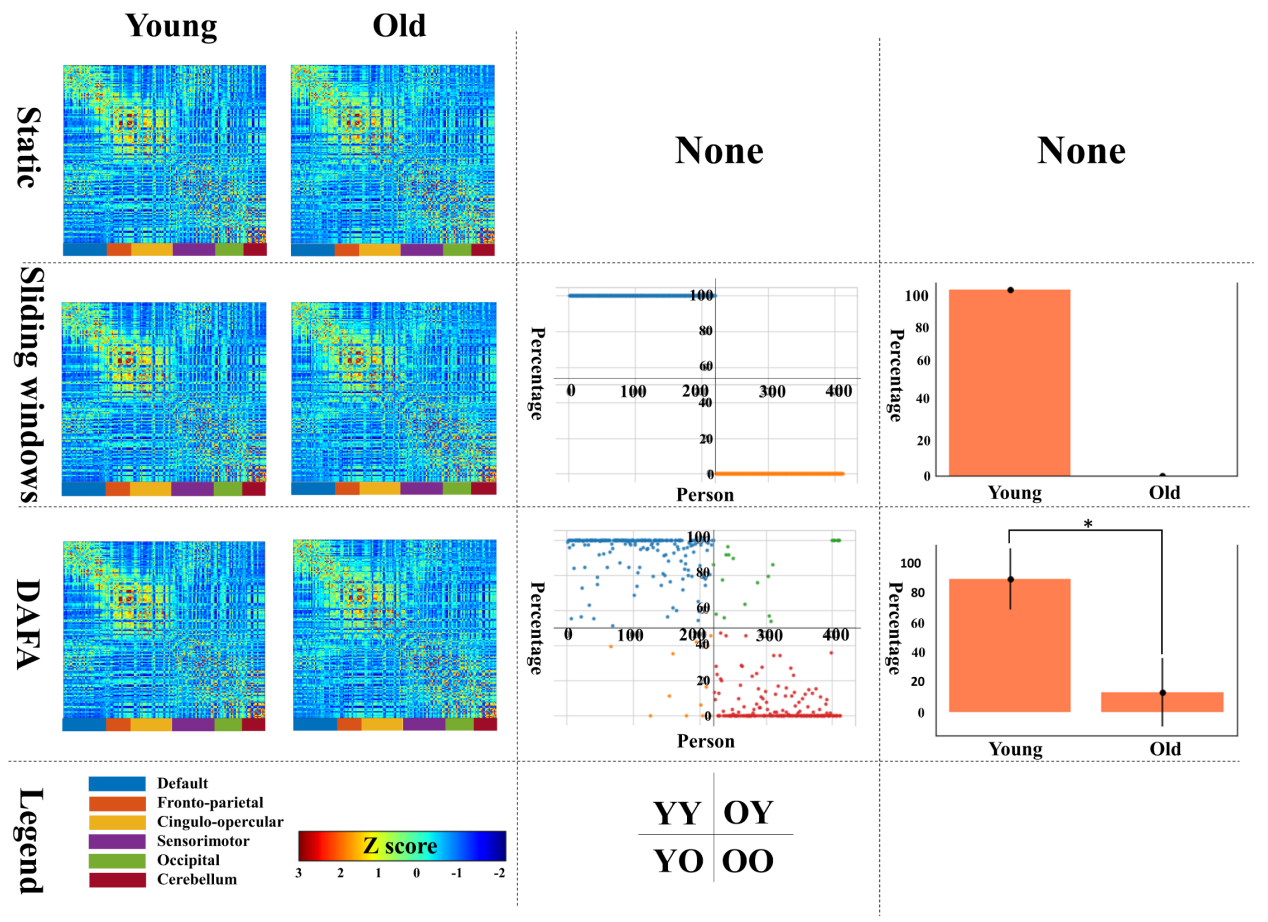


FIGURE 3 | The first column shows the average FC patterns of the young group and old group on the left and right, respectively; the second column shows the distribution of "young FCs" (samples in blue indicate individuals in the young group whose percentage of "young FCs" is greater than 50%; samples in orange indicate individuals in the young group whose percentage of "young FCs" is equal to or below 50%; samples in green indicate individuals in the old group whose percentage of "young FCs" is above 50%; and samples in yellow indicate individuals in the old group whose percentage of "young FCs" is equal to or less than 50%); the third column shows the mean and standard deviation of "young FCs" in the two groups. YY indicates samples in the young group for whom more than 50% of the FCs were relabeled to young; YO indicates samples in the young group for whom more than 50% of the FCs were relabeled to old; OO indicates samples in the old group for whom more than 50% of the FCs were relabeled to old; and OY indicates samples in the old group for whom more than 50% of the FCs were relabeled to young.

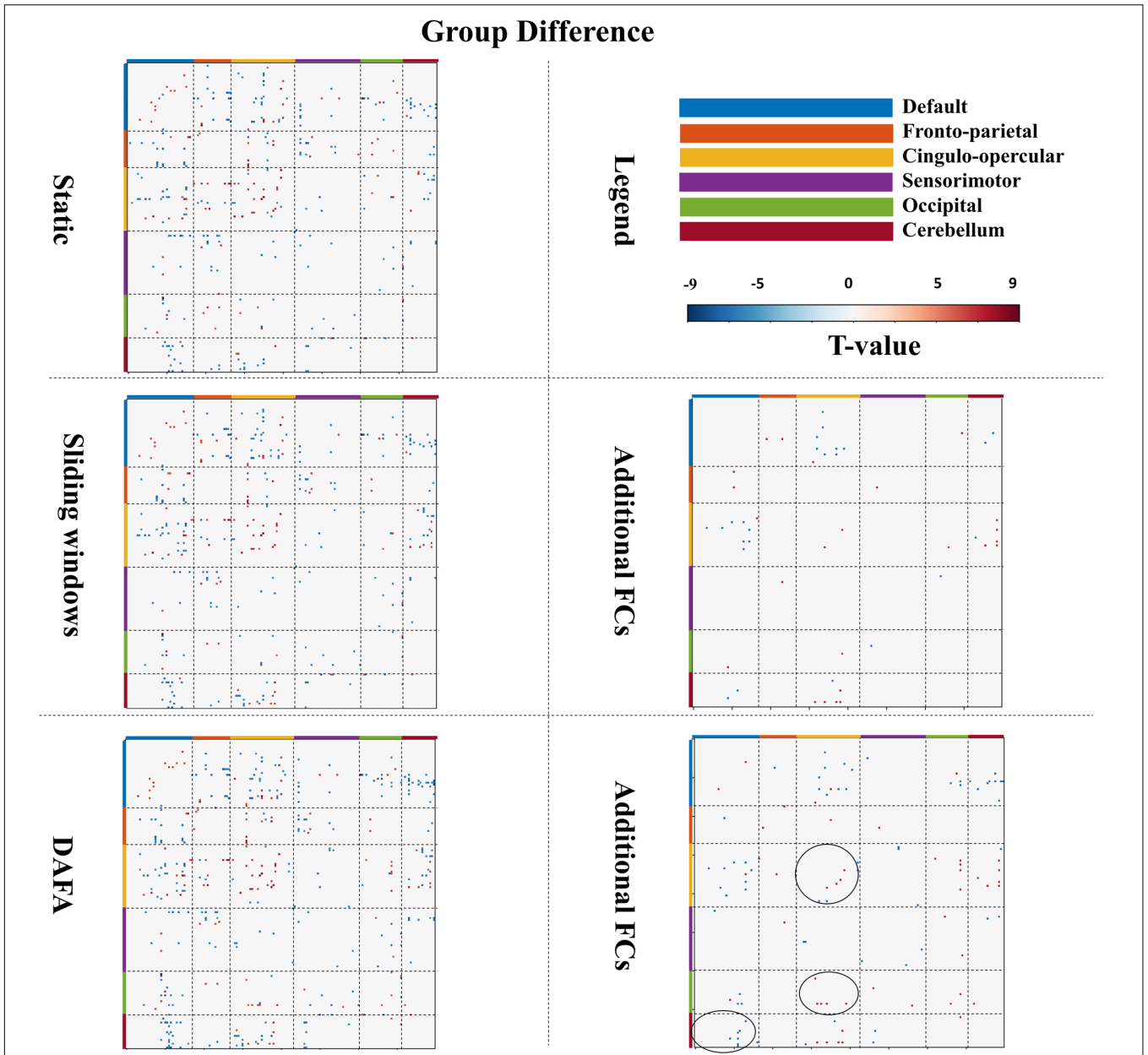


FIGURE 4 | Figures in the **left** column show significant changes in FCs (T-map, $p < 0.01$, FWE corrected) between the young and old groups via the static, sliding windows and DAFA methods. Blue indicates that the FC from the old group is stronger than that from the young group (i.e., old > young), and red indicates that the FC from the young group is stronger than that from old group (i.e., young > old). The figures in the **right** column show additional FCs revealed by the sliding windows and DAFA methods compared with the static method. Areas marked by circle indicated the additional altered FCs found by DAFA, compared with the sliding windows method.

DMN, dorsolateral frontal cortex for the frontoparietal network, ventral prefrontal cortex/basal ganglia/mid-insula/thalamus for the cingulo-opercular network, the supplementary motor area for the sensorimotor network, the postoccipital region in the occipital network, and infcerebellum/medcerebellum for the cerebellum network (**Figure 4**). Noting that, compared with the sliding window method, additional altered FC patterns were found by DAFA. These altered FCs were: (1) decreased connections within cingulo-opercular network, (2) increased connections between DMN and cerebellum networks, and (3) decreased connections between cingulo-opercular and occipital networks.

Relations Between Changed FCs and Behavioral Scores

To demonstrate the relationships between the altered FCs (revealed by the DAFA method) and behavioral performance,

partial correlations between the altered FCs and behavioral scores (fluid intelligence and speed Choice Reaction Time) were calculated while controlling for gender, head motion and intracranial volume. Here, the Cattell score was related to fluid intelligence, while M-RT and SD-RT were related to the speed Choice Reaction Time. The detected FC patterns were significantly correlated with fluid intelligence and the speed Choice Reaction Time ($p < 0.01$). In detail, the FCs within the DMN and the frontoparietal, cingulo-opercular and cerebellum networks, the connections between the DMN and the frontoparietal/sensorimotor networks and the connections between the cingulo-opercular network and the occipital/cerebellum/frontoparietal networks were positively correlated with the Cattell scores. The FCs within the sensorimotor and occipital networks were negatively correlated with the Cattell scores, as well as the connections between the cerebellum network and the DMN and sensorimotor networks and the connections between the cingulo-opercular and sensorimotor networks. In addition, the relationships between the FCs and the M-RT/SD-RT scores had the opposite trend from that of the results of the Cattell scores. The details are shown in **Figure 5**.

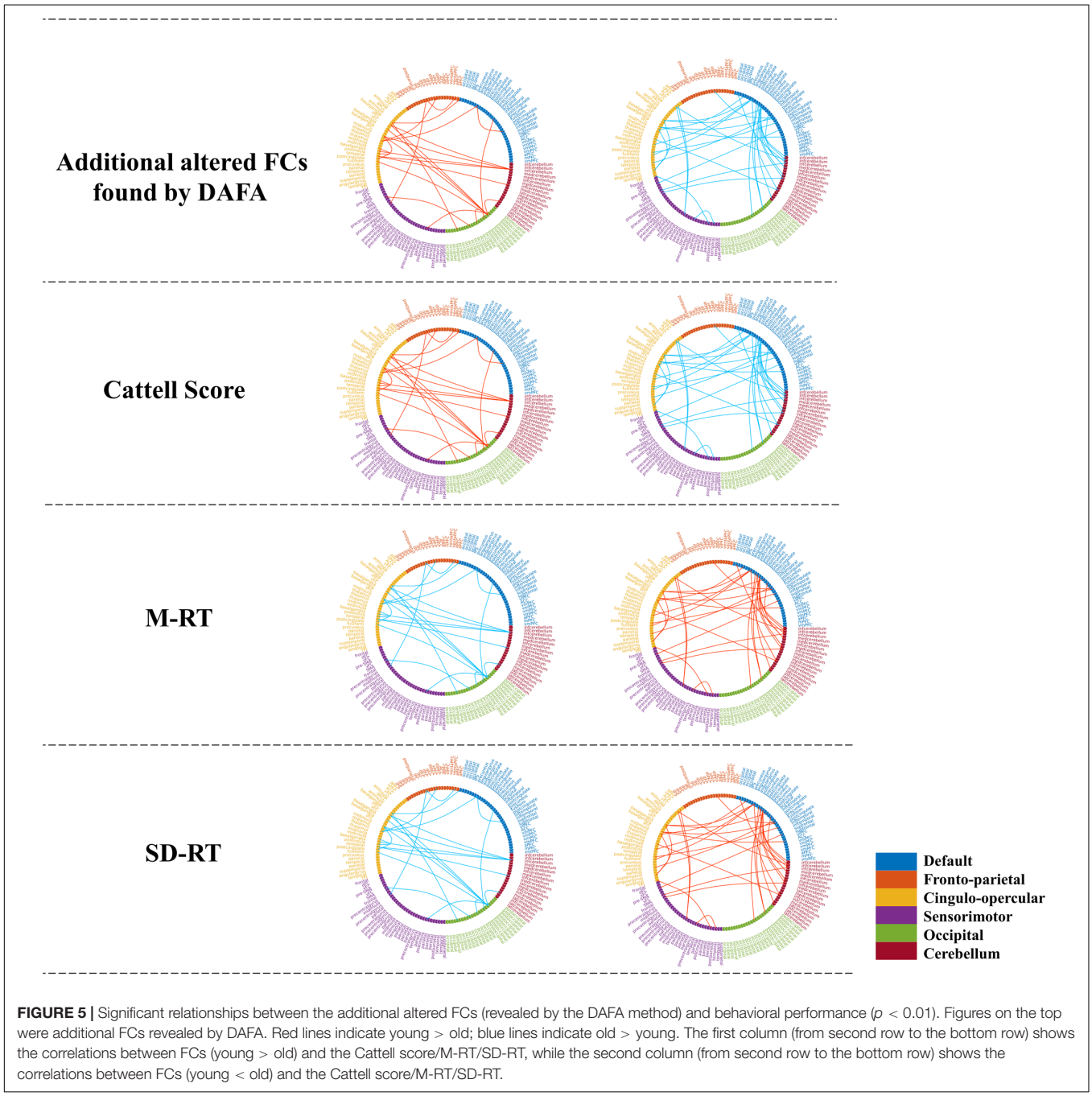


FIGURE 5 | Significant relationships between the additional altered FCs (revealed by the DAFA method) and behavioral performance ($p < 0.01$). Figures on the top were additional FCs revealed by DAFA. Red lines indicate young > old; blue lines indicate old > young. The first column (from second row to the bottom row) shows the correlations between FCs (young > old) and the Cattell score/M-RT/SD-RT, while the second column (from second row to the bottom row) shows the correlations between FCs (young < old) and the Cattell score/M-RT/SD-RT.

DISCUSSION

In this paper, based on the hypothesis that most of the functional activities in young adults may reflect the "young FCs", while those in old adults may largely reflect the "old FCs," we proposed a method named DAFA to analyze the changes in FCs during the process of aging from a big data perspective. DAFA revealed additional significant, altered FC patterns between the young and old groups, which were ignored by the static method, and these changes were correlated with behavior scores (Cattell score, M-RT, and SD-RT), which might be indicative of cognitive decline during aging.

DNN With AE Pretraining

For each training iteration, the sliding window method was used to calculate Pearson's correlation coefficient for 160 ROIs from a subject's brain, which enlarged the total sample size so that the autoencoder could examine the functional connectivity patterns more completely. The autoencoder is an unsupervised learning algorithm that can effectively mine hidden low-dimensional representations from data, which can then be used to reconstruct the original data (Tschannen et al., 2018). In this paper, we used a grid search to set and optimize the learning rate, batch size, epoch and other parameters of the autoencoder (**Figure 2**). The number of hidden nodes of AE is still open issue in deep learning. For example, if the number of the hidden nodes is based on the principle of minimum reconstruction error, the function of AE may present an identity transformation function and lose the ability of mining reasonable features. Therefore, referring to previous articles on the application of AE in fMRI (Suk et al., 2016; Guo et al., 2017) and empiricism setting of autoencoder, the hidden node of autoencoder in this work was set to 100. Finally, we obtained the model and corresponding parameters for the resting-state fMRI from CamCAN data. Unlike PCA or ICA, the autoencoder has no restrictions on whether the input data should be independent of each other and has no undetermined problem. It also revealed aging-related FCs from a comprehensive mining point of view with big data. The reconstruction error of 0.008 indicated that the extracted pattern was the representation of the FCs in low-dimensional space, which conveyed information for reconstructing the original signal to the greatest extent. The autoencoder could extract complex features; thus, the extracted features were combinations of linear and non-linear properties. The AE used in this work had a tradeoff between maximizing data representation and numerous restrictions (e.g., avoiding identity transformation and redundant feature extraction).

An AE-pretrained DNN can effectively reduce the training complexity and improve the training effect (Kim et al., 2016). In most cases, the DNN had been regarded as a black box. However, in deep learning-based fMRI studies, the weights between the input layer and the hidden layer of a neural network were interpreted as functional connectivity networks (Kim et al., 2016; Suk et al., 2016; Guo et al., 2017). In our case, these functional connectivity patterns can be interpreted as subnetworks of FNs during the resting-state or discriminable patterns comparing young and old groups. Noteworthy, there has been no fixed rule for hyper parameter settings in DNN models. Therefore, in our work, we used four layers for the network structure, and the number of nodes in each layer were set by experience for classification. This work was not concerned with the classification accuracy of the DNN model but instead was designed to see whether the relabeling results conform to our previous assumptions. Compared with SVM, the relabeling results obtained by the DNN rarely misclassified the FCs (e.g., the PCA + SVM results showed that more than 40% of the old subjects were relabel to the young group, as seen in **Supplementary Material**). Moreover, ICA + SVM requires that the samples be independent of each other, yet FCs from the same subject are not. Therefore, within the scope of this study, DNN was better than SVM. In addition, an unsupervised machine learning method, k-means clustering, was also applied to classify FCs in two classes based on Squalidean distance. And it had poor performance that the distribution of relabeled FCs by k-means clustering was scattered, and might be hard to detect underlying information of FCs (**Supplementary Figure S3**).

Big Data Perspective

In the data preprocessing section, the Pearson's correlation coefficient of time courses from every pair of ROIs was calculated by the sliding window method for one subject's rs-fMRI. Several previous studies suggested that the window length should exceed the slowest frequencies (in this work, the time series of calculating FCs are high-pass filtered at 0.01 Hz) which commonly assumed to comprise the BOLD signal (Andrew and Michael, 2015; Leonardi and Van De Ville, 2015). Thus, the ideal window length was 100 s. Considering the TR of rs-fMRI in Cam-CAN dataset was 1.97 s, the window length of 50 time points was used in this work (i.e., 100/1.97≈50). This procedure enlarged the training dataset for the AE and satisfied the requirements of big data analysis processing flow (Cobb et al., 2018). From a big data perspective, with the increase in data volume, more credible results can be obtained, and errors caused by individual differences can be eliminated as much as possible (Xia and He, 2017; Peter and Jayati, 2018). However, a large data sample size almost inevitably leads to confusion effects, such as an incorrect attribution of samples (Smith and Nichols, 2018). In the proposed method, supervised learning was used to relabel all sliding window FCs. At the same time, deep learning performed better with large data samples (Lecun et al., 2015). Thus, combined with deep learning, DAFA enlarged the datasets and satisfied deep learning applications under big data circumstances, which led to more reliable results (Peter and Jayati, 2018). On the other hand, when the sample size increased, there was a subset in the sample that caused the statistical method to always produce significant statistical results and lead to incorrect conclusions (Smith and Nichols, 2018). Therefore, we averaged identical samples for each subject and took the resulting mean FC as that subject's final FC, which made the samples used for statistical analysis independent and eliminated interference factors. The search-back step only considered the effect on the subject, not on the FCs derived from sliding window, which prevented confounding effects during the statistical analysis for big data (Smith and Nichols, 2018). Additionally, since we only analyzed young and old groups, each FC calculated from sliding window had a discriminatory

prior (label), but according to the hypothesis, a small number of samples may have been mislabeled, so it was necessary to determine the samples corresponding to these false priors. Due to the poor performance of the clustering algorithm with high-dimensional data, it was not effective or sufficiently accurate to classify the samples; thus, in this article, the DNN was pretrained with the AE, and a supervised learning method was implemented to relabel and filter out the mixed samples.

FC Changes in Aging

In our work, age-related FC changes were successfully revealed by the DAFA method. First, these changes included age-related decreased patterns, consisting of the FCs within the DMN and the frontoparietal and cingulo-opercular networks, as well as the connections between the frontoparietal and cingulo-opercular networks. Previous studies have suggested that decreased FC within the DMN and the cingulo-opercular network reflected the decline in cognitive function related to the attention, memory and executive functions in elderly individuals (Dijk et al., 2010; Ferreira and Busatto, 2013; Geerligs et al., 2015; Grady et al., 2016; Spreng et al., 2016), which is also supported by our findings on the correlation between behavioral score and the FCs in these networks. Regarding the decreased between-network FCs, the frontoparietal and cingulo-opercular networks have been hypothesized to support the top-down control of executive function (Dosenbach et al., 2008). The decreased FCs between the frontoparietal and cingulo-opercular networks might be caused by a reduction of harmonization of these networks and cognitive function loss in normal aging. Second, the connections within the sensorimotor network were increased, as were the connections between the sensorimotor and the frontoparietal/cingulo-opercular networks and between the DMN and the fronto-parietal/cingulo-opercular/sensorimotor/occipital/cerebellum networks. The sensorimotor function of elderly individuals shows a decline compared with young adults, and the increased FCs with the sensorimotor network might imply that a higher level of anticipated preparation was required for the decline of sensorimotor function (Mathys et al., 2014). Moreover, the increased connectivity between the sensorimotor and frontoparietal/cingulo-opercular networks might reflect a compensatory response to the dysfunction of other brain networks and neurotransmitter decline (He et al., 2016). Furthermore, the increased FCs between the DMN and the other FNs have also been found by previous studies (Zhang et al., 2015; Grady et al., 2016; Spreng et al., 2016; Damoiseaux, 2017), which might be due to the DMN playing an enhanced intermediary role in regulating primary and higher-order networks (Margulies et al., 2016; Kernbach et al., 2018). Third, both increased and decreased connections between the cingulo-opercular and cerebellum networks were also found, which might reflect other dysfunction in elderly individuals. In brief, the results from DAFA were consistent with the abovementioned studies to some extent.

Our DAFA method yielded additional changed FCs compared to the static method, which showed decreased FCs compared with the young group were mainly in the connections between the cingulo-opercular and occipital/cerebellum networks. Previous studies have implied that the cingulo-opercular network is important for set maintenance and other functions, including processing negative effects, pain and cognitive control (Dosenbach et al., 2008; Church et al., 2009; Sylvester et al., 2012); additionally, an alertness study found that the cingulo-opercular network showed a task-positive response in an event-related design with auditory and visual stimuli (Coste and Kleinschmidt, 2016). Therefore, the decreased FCs between the cingulo-opercular and occipital networks in our work likely implied the decline in high-order functions (e.g., set-maintenance and alertness functions). Furthermore, the cerebellum is associated with many cognitive functions involving emotion, executive function, language and working memory (Keren-Happuch et al., 2014). A decreased FC between the cingulo-opercular and cerebellum networks might suggest a loss in the harmonization of these networks in elderly individuals. Meanwhile, the increased additional FCs were mainly in the connections between the DMN and cingulo-opercular/occipital/cerebellum networks, which was consistent with the trend in the static results, thereby adding more detailed evidence to the FC changes in aging. In addition, the result of additional FCs detected via DAFA might imply that two kinds of neural circuits (decreased connections between the cingulo-opercular network and the occipital/cerebellum networks, and increased connections between DMN and cingulo-opercular/occipital/cerebellum networks) existed, which perhaps were corresponding to harmonization loss and compensation mechanism, respectively.

Limitations

The limitations of current work are following: first, the aim of this work was to detect the information between health young group and old group (aging), and the current age division was based on the published paper of Cam-CAN open dataset (Shafto et al., 2017). However, a finer bin-based age classification may be investigated in the future effort. Second, considering deep learning performed better with large data samples (Lecun et al., 2015) and it perhaps has satisfied performance on mixed labels (Guan et al., 2017), we proposed a new method combining deep learning with neuroimaging big data, to fully explore the potentials of data with new hypotheses, and some replicate existing findings were also obtained. However, our method may need to be further improved, e.g., heuristic/clinically inspired method could be developed to prevent the autoencoder on digesting irrelevant connections. Third, the deep learning models applied in this study were an autoencoder and a DNN. During the training process, these models required initialization of their hyper parameters, which generally relied on user experiences. For example, there is no gold standard for the number of hidden nodes in an auto-encoder. Another is that performing the proposed DAFA method requires GPUs to train the deep learning models; otherwise, it would consume more time than traditional methods such as ICA and SVM. The current version of DAFA codes was available on https://github.com/Xin-cqu/DAFA. At last, more efforts were needed to verify that DAFA might be a promising method for exploring important information in other fMRI studies.

CONCLUSION

In this work, we developed a new method named DAFA to deeply detect important information about age-related FC changes. The results demonstrated that DAFA had the following advantages: (1) the potential confused relationships in the data was taken into account and combined with deep learning methods, DAFA could detect the complex information of functional connectivity networks for normal aging in a more comprehensive way; (2) from big data perspective, DAFA has enlarged the dataset and reduced the cofounds effect of statistical analysis resulting in improving the reliability of the analysis; and (3) DAFA was constructed as a new method to detect underlying important brain information from fMRI analysis. Additionally, it may be a promising method for exploring important alteration information in fMRI studies on diseases such as Alzheimer's disease and schizophrenia.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Cambridgeshire 2 (now East of England – Cambridge Central) Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LD and JC conceived the project. XW and LD implemented the code of the method and wrote the manuscript. XL and JY preprocessed the MRI data. LD and CL supervised the analysis of FC patterns discovered by the method. HL, JX, DY, and CL provided critical suggestions for the manuscript.

FUNDING

This study was financially supported by the National Natural Science Foundation of China (Nos. 81861128001, 81701778, 61672374, and 61873178) and the Sichuan Science and Technology Program (Nos. 2018JZ0073 and 2019YJ0179).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01435/full#supplementary-material

REFERENCES

Andrew, Z., and Michael, B. (2015). Towards a statistical test for functional connectivity dynamics. *Neuroimage* 114, 466–470. doi: 10.1016/j.neuroimage.2015.03.047

Avelar-Pereira, B., Backman, L., Wahlin, A., Nyberg, L., and Salami, A. (2017). Age-related differences in dynamic interactions among default mode, frontoparietal control, and dorsal attention networks during resting-state and interference resolution. *Front. Aging Neurosci.* 9:152. doi: 10.3389/fnagi.2017.00152

Biswal, B.B., Mennes, M., Zuo, X., Gohel, S., Kelly, C., Smith, S.M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107

Biswal, B.B., Yetkin, F.Z., Haughton, V.M., and Hyde, J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Calhoun, V.D., and Sui, J. (2016). Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1, 230–244. doi: 10.1016/j.bpsc.2015.12.005

Chae, S., Kwon, S., and Lee, D. (2018). Predicting infectious disease using deep learning and big data. *Int. J. Environ. Res. Public Health* 15:E1596. doi: 10.3390/ijerph15081596

Church, J.A., Fair, D.A., Dosenbach, N.U., Cohen, A.L., Miezin, F.M., Petersen, S.E., et al. (2009). Control networks in paediatric Tourette syndrome show immature and anomalous patterns of functional connectivity. *Brain* 132(Pt 1), 225–238. doi: 10.1093/brain/awn223

Cobb, A.N., Benjamin, A.J., Huang, E.S., and Kuo, P.C. (2018). Big data: more than big data sets. *Surgery*.164, 640–642 doi: 10.1016/j.surg.2018.06.022

Coste, C.P., and Kleinschmidt, A. (2016). Cingulo-opercular network activity maintains alertness. *Neuroimage* 128, 264–272. doi: 10.1016/j.neuroimage.2016.01.026

Damoiseaux, J.S. (2017). Effects of aging on functional and structural brain connectivity. *Neuroimage* 160, 32–40. doi: 10.1016/j.neuroimage.2017.01.077

De Luca, M., Beckmann, C., De Stefano, N., Matthews, P., and Smith, S.M. (2006). fMRI resting state networks define distinct modes of long-distance interactions in the human brain. *Neuroimage* 29, 1359–1367. doi: 10.1016/j.neuroimage.2005.08.035

de Vos, F., Koini, M., Schouten, T.M., Seiler, S., van der Grond, J., Lechner, A., et al. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *Neuroimage* 167, 62–72. doi: 10.1016/j.neuroimage.2017.11.025

Dijk, K.R.A., Van Trey, H., Archana, V., Evans, K.C., Lazar, S.W., and Buckner, R.L. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* 103, 297–321. doi: 10.1152/jn.00783.2009

Dong, L., Luo, C., Liu, X., Jiang, S., Feng, H., Li, J., et al. (2018). Neuroscience information toolbox: an open source toolbox for EEG-fMRI multimodal fusion analysis. *Front. Neuroinform.* 12:56. doi: 10.3389/fninf.2018.00056

Dosenbach, N.U., Fair, D.A., Cohen, A.L., Schlaggar, B.L., and Petersen, S.E. (2008). A dual-networks architecture of top-down control. *Trends Cogn. Sci.* 12, 99–105. doi: 10.1016/j.tics.2008.01.001

Dosenbach, N.U.F., Binyam, N., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., et al. (2010). Prediction of individual brain maturity using fMRI. *Science* 329, 1358–1361. doi: 10.1126/science.1194144

Du, Y., Pearlson, G.D., Lin, D., Sui, J., Chen, J., Salman, M., et al. (2017). Identifying dynamic functional connectivity biomarkers using GIG-ICA: application to schizophrenia, schizoaffective disorder, and psychotic bipolar disorder. *Hum. Brain Mapp.* 38, 2683–2708. doi: 10.1002/hbm.23553

Ferreira, L.K., and Busatto, G.F. (2013). Resting-state functional connectivity in normal brain aging. *Neurosci. Biobehav. Rev.* 37, 384–400. doi: 10.1016/j.neubiorev.2013.01.017

Geerligs, L., Renken, R.J., Saliasi, E., Maurits, N.M., and Lorist, M.M. (2015). A brain-wide study of age-related changes in functional connectivity. *Cereb. Cortex* 25, 1987–1999. doi: 10.1093/cercor/bhu012

Geerligs, L., Tsvetanov, K.A., Cam, C., and Henson, R.N. (2017). Challenges in measuring individual differences in functional connectivity using fMRI: the case of healthy aging. *Hum. Brain Mapp.* 38, 4125–4156. doi: 10.1002/hbm.23653

Grady, C., Sarraf, S., Saverino, C., and Campbell, K. (2016). Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiol. Aging* 41, 159–172. doi: 10.1016/j.neurobiolaging.2016.02.020

Guan, M.Y., Gulshan, V., Dai, A.M., and Hinton, G.E. (2017). Who said what: modeling individual labelers improves classification. *arXiv* [Preprint]. Available: https://ui.adsabs.harvard.edu/abs/2017arXiv170308774G (accessed March 01, 2017) .

Guo, X., Dominick, K.C., Minai, A.A., Li, H., Erickson, C.A., and Lu, L.J. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front. Neurosci.* 11:460. doi: 10.3389/fnins.2017.00460

He, H., Luo, C., Chang, X., Shan, Y., Cao, W., Gong, J., et al. (2016). The functional integration in the sensory-motor system predicts aging in healthy older adults. *Front. Aging Neurosci.* 8:306. doi: 10.3389/fnagi.2016.00306

Horn, J.L., and Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *J. Educ. Psychol.* 57, 253–270. doi: 10.1037/h0023816

Hu, X., Huang, H., Peng, B., Han, J., Liu, N., Lv, J., et al. (2018). Latent source mining in FMRI via restricted Boltzmann machine. *Hum. Brain Mapp.* 39, 2368–2380. doi: 10.1002/hbm.24005

Karahanoglu, F.I., and Van De Ville, D. (2015). Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks. *Nat. Commun.* 6:7751. doi: 10.1038/ncomms8751

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., et al. (2017). BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. *Neuroimage* 146, 1038–1049. doi: 10.1016/j.neuroimage.2016.09.046

Keren-Happuch, E., Annabel Chen, S-H., Ho, M-H., and Desmond, J. E. (2014). A meta-analysis of cerebellar contributions to higher cognition from PET and fMRI studies. *Hum. Brain Mapp.* 35, 593–615. doi: 10.1002/hbm.22194

Kernbach, J.M., Yeo, B.T.T., Smallwood, J., Margulies, D.S., Thiebaut de Schotten, M., Walter, H., et al. (2018). Subspecialization within default mode nodes characterized in 10,000 UK Biobank participants. *Proc. Natl. Acad. Sci. U.S.A.* 115, 12295–12300. doi: 10.1073/pnas.1804876115

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M.R. (2019). Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *Neuroimage* 199, 651–662. doi: 10.1016/j.neuroimage.2019.06.012

Kim, J., Calhoun, V.D., Shim, E., and Lee, J.H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage* 124(Pt A), 127–146. doi: 10.1016/j.neuroimage.2015.05.018

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Leonardi, N., and Van De Ville, D. (2015). Erratum to "On spurious and real fluctuations of dynamic functional connectivity during rest". *NeuroImage* 104, 464–465. doi: 10.1016/j.neuroimage.2014.10.045

Liu, X., and Duyn, J.H. (2013). Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4392–4397. doi: 10.1073/pnas.1216856110

Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12574–12579. doi: 10.1073/pnas.1608282113

Mathys, C., Hoffstaedter, F., Caspers, J., Caspers, S., Sudmeyer, M., Grefkes, C., et al. (2014). An age-related shift of resting-state functional connectivity of the subthalamic nucleus: a potential mechanism for compensating motor performance decline in older adults. *Front. Aging Neurosci.* 6:178. doi: 10.3389/fnagi.2014.00178

Peter, S., and Jayati, D.-M. (2018). Big data: what it can and cannot achieve. *BJPsych. Adv.* 24, 237–244. doi: 10.1192/bja.2018.15

Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018

Shafto, M.A., James, L.E., Abrams, L., Tyler, L.K., and Cam, C.A.N. (2017). Age-related increases in verbal knowledge are not associated with word finding problems in the cam-can cohort: what you know won't hurt you. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 72, 100–106. doi: 10.1093/geronb/hgbw074

Sheffield, J.M., Repovs, G., Harms, M.P., Carter, C.S., Gold, J.M., MacDonald, A.W., 3rd, et al. (2015). Fronto-parietal and cingulo-opercular network integrity and cognition in health and schizophrenia. *Neuropsychologia* 73, 82–93. doi: 10.1016/j.neuropsychologia.2015.05.006

Smith, S.M., and Nichols, T.E. (2018). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron* 97, 263–268. doi: 10.1016/j.neuron.2017.12.018

Spreng, R.N., Stevens, W.D., Viviano, J.D., and Schacter, D.L. (2016). Attenuated anticorrelation between the default and dorsal attention networks with aging: evidence from task and rest. *Neurobiol. Aging* 45, 149–160. doi: 10.1016/j.neurobiolaging.2016.05.020

Suk, H.I., Wee, C.Y., Lee, S.W., and Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *Neuroimage* 129, 292–307. doi: 10.1016/j.neuroimage.2016.01.005

Sylvester, C.M., Corbetta, M., Raichle, M.E., Rodebaugh, T.L., Schlaggar, B.L., Sheline, Y.I., et al. (2012). Functional network dysfunction in anxiety and anxiety disorders. *Trends Neurosci.* 35, 527–535. doi: 10.1016/j.tins.2012.04.012

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., et al. (2017). The cambridge centre for ageing and neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* 144(Pt B), 262–269. doi: 10.1016/j.neuroimage.2015.09.018

Tomasi, D., and Volkow, N.D. (2012). Aging and functional brain networks. *Mol. Psychiatry* 17, 549–558. doi: 10.1038/mp.2011.81

Tschannen, M., Bachem, O., and Lucic, M. (2018). Recent advances in autoencoder-based representation learning. *arXiv* [Preprint]. doi: 10.1109/JBHI.2018.2871678

Vij, S.G., Nomi, J.S., Dajani, D.R., and Uddin, L.Q. (2018). Evolution of spatial and temporal features of functional brain networks across the lifespan. *Neuroimage* 173, 498–508. doi: 10.1016/j.neuroimage.2018.02.066

Xia, M., and He, Y. (2017). Functional connectomics from a "Big Data" perspective. *Neuroimage* 160, 152–167. doi: 10.1016/j.neuroimage.2017.02.031

Zhang, D., Liu, X., Chen, J., Liu, B., and Wang, J. (2015). Widespread increase of functional connectivity in Parkinson's disease with tremor: a resting-state FMRI study. *Front. Aging Neurosci.* 7:6. doi: 10.3389/fnagi.2015.00006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wen, Dong, Chen, Xiang, Yang, Li, Liu, Luo and Yao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 25 February 2020
doi: 10.3389/fnhum.2019.00472

Glioma Migration Through the Corpus Callosum and the Brainstem Detected by Diffusion and Magnetic Resonance Imaging: Initial Findings

Guive Sharifi[1,2†], Amir Mohammad Pajavand[1,2,3†], Saeedeh Nateghinia[2], Tohid Emami Meybodi[1,4] and Hossein Hasooni[2]*

[1]Skull Base Research Center, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [2]Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [3]Institute for Cognitive and Brain Sciences, Shahid Beheshti University GC, Tehran, Iran, [4]Cellular and Molecular Research Center, Iran University of Medical Sciences, Tehran, Iran

OPEN ACCESS

Edited by:
Reza Lashgari, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Paul Matthew Thompson, University of Southern California, United States
Hamidreza Saligheh Rad, Tehran University of Medical Sciences, Iran

***Correspondence:**
Amir Mohammad Pajavand
shahryarpajavand@outlook.com

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Brain Imaging and Stimulation, a section of the journal Frontiers in Human Neuroscience

Received: *20 August 2019*
Accepted: *24 December 2019*
Published: *25 February 2020*

Citation:
Sharifi G, Pajavand AM, Nateghinia S, Meybodi TE and Hasooni H (2020) Glioma Migration Through the Corpus Callosum and the Brainstem Detected by Diffusion and Magnetic Resonance Imaging: Initial Findings. Front. Hum. Neurosci. 13:472. doi: 10.3389/fnhum.2019.00472

Purpose: Glioma cell infiltration, in which the glioma tumor cells spread long distances from the primary location using white matter (WM) or blood vessels, is known as a significant challenge for surgery or localized chemotherapy and radiation therapy. Following the World Health Organization (WHO), the glioma grading system ranges from stages I to IV, in which lower-grade gliomas represent benign tumors, and higher grade gliomas are considered the most malignant.

Materials and Methods: We gathered magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) data for seven patients with right precentral gyrus-located tumors and six age- and sex-matched healthy subjects for analysis. Tract-Based Spatial Statistics (TBSS) was utilized to evaluate whole-brain WM implication due to probable tumor infiltration. Also, along-tract statistics were used in order to trace the implicated WM tracts. Finally, for cortical evaluation of probable tumor cell migration, voxel-based morphometry (VBM) was utilized, which allowed us to do whole-brain cortical estimation.

Results: The TBSS results revealed significantly higher fractional anisotropy (FA) and lower mean diffusivity (MD) in the left side superior corona radiata. Also, higher FA was observed in the right corticostriatal tract. Along-tract statistics were also compiled on the corpus callosum (CC), which is anatomically known as a hub between hemispheres. The body of the CC, which connected with the superior corona radiata anatomically, showed significantly higher FA values relative to healthy subjects, which are in line with the TBSS results. Consistent with these results, whole-brain gray matter changes were analyzed *via* VBM, which showed significant hypertrophy of both sides of the brainstem.

Conclusion: In future investigations, focusing on the genetic basis of the glioma patients in line with imaging studies on a larger sample size, which is known as genetics imaging, would be a suitable approach for tracing this process.

Keywords: tract-based spatial statistics, voxel-based morphometry, magnetic resonance imaging, corpus callosum, white matter

INTRODUCTION

Glioma invasion, as the major obstacle for curing patients, is hypothesized to be carried out *via* extracellular routes to the other brain structures, and causes significant complications for complete surgical resection and chemotherapy and radiation therapy.

Glioma cell migration, which can occur over long distances *via* white matter (WM) or blood vessels to injure WM or cortical structures, has been investigated previously in mammalian brains (Cayre et al., 2009). Scherer (1938) investigated 100 patients with glioma tumors and created criteria for glioma invasion through the brain parenchyma, preexisting blood vessels, subarachnoid space, and WM tracts.

Diffusion tensor imaging (DTI) is a modern technique that is sensitive to the diffusion of water flow along the WM tracts. Hence, it can detect tumor infiltration through the WM tracts when conventional magnetic resonance imaging (MRI) scans appear to be normal (Price et al., 2003).

The infiltrating glioma cells extend beyond a surgeon's reach, which can limit the effectiveness of localized therapy (Hochberg and Pruitt, 1980; Burger et al., 1988; Kreth et al., 1993; Shapiro, 1999).

Regarding the symptoms of right precentral gyrus glioma patients, sensory deficits in the contralateral side are usually represented in these patients, which could be due to atrophy or even hypertrophy of sensory cortices, such as the primary and secondary somatosensory cortex (SSC), the brainstem, and insula.

The pathway of glial cell migration from the left precentral gyrus to its contralateral side and its effects on cortical regions are the main focus of this study. Therefore, according to the aforementioned findings, we hypothesized that (see **Figure 1**) the glial cells of left precentral gyrus tumors may migrate along the corpus callosum (CC) WM tract, which acts as a hub between hemispheres. The specific region of the CC involved is not yet known and needs to be clarified in the study.

Next, if glioma migration is proved by the results, the probable cortical changes of the left hemisphere also need to be investigated. That is why, we used DTI techniques [deterministic tractography, tract-based spatial statistics (TBSS), and along-tract statistics] to clarify this probable mechanism.

MATERIALS AND METHODS

Participants

In this study, seven patients with different types of glioma tumors of the right motor cortex (see **Table 1**; **Figure 1**) and seven healthy subjects (age- and sex-matched) were included. Subjects with previous neurological problems, such as Parkinson's disease or previous cancer history, were excluded. In addition, in order to prevent statistical problems, we performed a two-tailed *t*-test, and no significant differences were seen in either the sex ($P = 0.2$) or age ($P = 0.32$) of the groups.

Image Acquisition

Structural MRI and DTI Data Acquisition

All structural MRI scans were acquired from 3T MRI scanners (Siemens Prisma). 3D T1 MPRAGE anatomic acquisitions was done (1 mm slices, 256 × 256 matrix, echo time [TE] = 3.74 ms, repetition time [TR] = 1810 ms, flip angle = 30°) and used to superimpose DTI images.

Also, diffusion-weighted imaging (DWI) brain scan was done *via* the same scanner with a 64-channel head coil. Other acquisition parameters were as follows: number of slices, 68; diffusion directions, 30; FOV, 256×256 mm^2; voxel size, $2 \times 2 \times 2$ mm^3; TR/TE, 9,000/90 ms.

Data Analysis

Structural Data Preprocessing and Voxel-Based Morphometry (VBM)

Structural data were analyzed using FSL-VBM protocol (Ashburner and Friston, 2000; Good et al., 2001). First, nonbrain tissues were extracted using the brain extracting tool (BET; Smith, 2002) and GM-segmented before MNI152 standard space registration using nonlinear registration.

The resulting GM partial volume images were averaged to create a study template, and then all native GM images were registered nonlinearly to this template. The modulated segmented images were then smoothed with an isotropic Gaussian kernel with a sigma of 3 mm. For the last step, the voxel-wise general linear model (GLM) was applied using permutation-based testing. Threshold-free cluster enhancing (TFCE) was used for thresholding (Smith and Nichols, 2009). Images were thresholded at $P < 0.05$ and corrected for multiple comparisons (see **Figure 2**).

DTI Data Preprocessing and TBSS

In order to preprocess the DTI data, we utilized the FMRIB Software Library (FSL 5)[1]. At first, all diffusion-weighted images were checked visually for any visible artifacts and then corrected for B_0 inhomogeneities and eddy-current distortion, with each subject's DWI registered to the corresponding $b = 0$ images *via* affine transformation.

Second, the data were brain-extracted *via* FSL BET to remove unwanted voxels, and finally, fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), and eigenvector L1, L2, and L3 maps were created using DTIFIT.

For analyzing WM changes between groups, TBSS was performed (Smith et al., 2006) using Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL) version 5.0 (Smith et al., 2004). For the first step, all FA images were nonlinearly aligned to a common space (FMRIB58FA_1 mm). In order to create a mean FA skeleton, mean FA images were created for each subject, and each FA image was projected onto the mean FA skeleton. TBSS was also performed for non-FA data (MD and RD). Nonlinear registration was obtained for FA data, and each MD image was projected to the mean FA skeleton.

[1] http://www.fmrib.ox.ac.uk/fsl

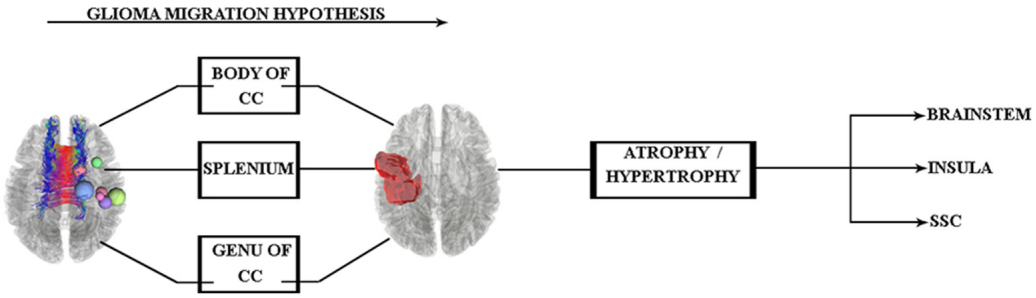


FIGURE 1 | Glioma tumor migration hypothesis between hemispheres and its probable effect on cortical regions of the contralateral side. Abbreviations: CC, corpus callosum; SSC, somatosensory cortex.

TABLE 1 | Demographic and clinical data of the study population.

	Patients 7	Controls 6	*T*-value -	*P*-value -
Age	45 ± 11	45 ± 12	0.30	0.76*
Sex (Male)	5 (2)	4 (2)	-	-
Pathology results				
1. Diffuse astrocytoma with foci of anaplastic transformation (grade II)				
2. Oligodendroglioma, WHO grade II				
3. Oligodendroglioma, WHO grade II				
4. Gemistocytic astrocytoma with anaplastic transformation, WHO grade III				
5. Ganglioglioma, WHO grade I				
6. Ganglioglioma, WHO grade I				
7. Oligoastrocytoma WHO grade II				

*Seven patients with no cancer background or metastatic tumor were included in the study. Values are expressed as mean ± SD. *P-value calculated independent two-sample t-test.*

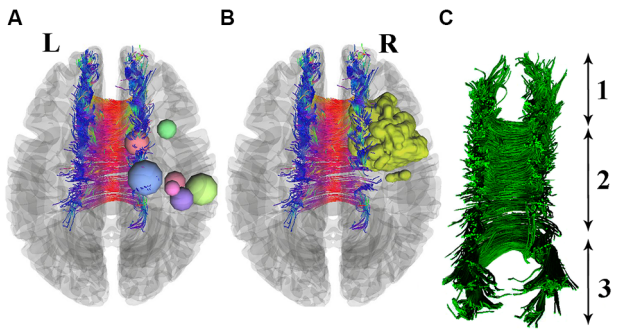


FIGURE 2 | Seven patients with right precentral gyrus glioma (nodes are used for better visualization) are shown in section **(A)**. One patient with a 3D-segmented tumor is visualized as an example in part **(B)**. Deterministic tractography of CC is presented in part **(C)**, which includes its whole neuroanatomy: (1) Genu of CC; (2) Body of CC; and (3) Splenium.

Statistical Analyses

Voxel and Tract-Based Statistics

A voxel-wise GLM was performed for structural T1 data after performing the VBM protocol using 5,000 permutations; finally, TFCE was used for finding significant clusters with thresholds set at $p < 0.05$, corrected for multiple comparisons (Smith and Nichols, 2009).

Statistical analyses for DTI data were performed using voxel-wise statistical analysis of FA, MD data using TBSS V1.2 part of FSL (Smith et al., 2006), and FA, MD changes were assessed using permutation-based nonparametric testing with 5,000 random permutations (Nichols and Holmes, 2002). Our statistical threshold was TFCE and the family-wise error-corrected *P*-value of 0.05 (Smith and Nichols, 2009).

After performing voxel-wise statistics, MNI coordinates of significant clusters in GM were calculated, and anatomical regions were identified using Harvard subcortical and Juelich atlases.

Along-Tract Statistics

VBM and TBSS allowed us to evaluate the cortical and WM alterations in glioma patients relative to healthy subjects, but in order to achieve the disclosure of glioma invasion patterns along WM tracts, we used along-tract statistical analyses (Colby et al., 2012) to show the probable hub of this invasion.

Based on this technique, which could be a suitable complement for TBSS, the FA values along the significant tracts that were visualized by TBSS were extracted again separately to determine the WM hypertrophy or demyelination probability in glioma patients vs. healthy controls.

According to the small sample size, we decided to use a linear mixed-effects model, in which the FA is the only dependent variable. Hence, using the along-tract statistics MATLAB

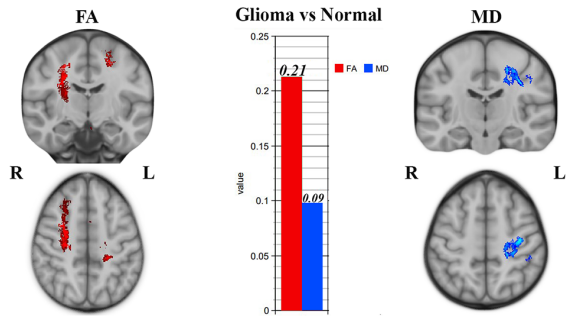


FIGURE 3 | Tract-based spatial statistics (TBSS) results showing significant superior corona radiata and corticostriatal tract alterations in patients vs. normal subjects. Significantly higher fractional anisotropy (FA) and lower mean diffusivity (MD) values were seen in CC, which could be a sign of white matter (WM) maturation.

toolbox[2] allowed us to extract FA values and streamlines, and the standard deviation of the WM tract was added to the other variables for application on WM tract data as a serial univariate approach. Then fixed-effect results containing the position factor, overall intercept, and group:position interaction were analyzed.

All the procedures of statistical analyses were performed *via* R version 3.5.0 (R Core Team, 2018) and MATLAB (MathWorks, Natick, MA, USA) software.

RESULTS

Tract-Based Spatial Statistics

Significant WM maturation occurred in patients vs. healthy controls, and the FA was significantly higher ($P < 0.001$) than the MD in patients relative to healthy subjects (see **Figure 3**); this result could be a reflection of probable WM maturation due to glioma tumor migration.

According to **Figure 3**, right corticostriatal pathways in the ipsilateral side of the tumor's location showed higher FA but no significant MD changes. Hence, we cannot judge whether there was WM maturation or demyelination.

According to **Figure 3**, the left superior corona radiata showed WM maturation as indicated by the DTI metrics. The FA was significantly higher than the MD values ($P < 0.001$), which could be a sign of glioma migration along the CC, which acts as a hub between both hemispheres.

In keeping with our hypothesis diagram (see **Figure 1**), glioma migration into the contralateral side of the tumors' location is observed, but consistent with TBSS limitations (Bach et al., 2014), specific disclosure of glioma migration along the WM route would not be possible. Hence, we performed along-tract statistics to clarify whether there was a matured WM hub between hemispheres.

Along-Tract Statistics

In order to verify our probable WM pattern prediction of glioma migration, FA values were analyzed along the tract of the CC, which is anatomically known as a hub between hemispheres.

As, we show in **Figure 4A**, streamlines between groups were analyzed, and lower streamlines but no signs were found in glioma patients vs. healthy subjects ($t = -1.87$, $p = 0.065$). According to the visualized results in **Figures 4B,C**, along-tract statistical analyses were carried out in order to reveal the overall FA vs. position curves between the groups.

This analysis showed significant group–position implication ($F = 33.3$, $p = 0.00018$) in the body of the CC and right corticostriatal WM, which are illustrated in **Figure 4C**.

Hence, all the statistical results can be visualized in **Figure 4C**, which is consistent with the TBSS results showing higher FA in the left superior corona radiata in glioma patients.

According to these results and our hypothesis diagram (see **Figure 5**), the body of the CC played a significant role in glioma migration from the precentral gyrus to the contralateral side (see **Figure 4**). This migration would induce cortical changes, such as atrophy or hypertrophy of regions, which need to be verified by cortical imaging techniques, as we have done in this study.

Voxel-Based Morphometry

Structural T1 analyses *via* VBM revealed brainstem hypertrophy on both sides (see **Figure 6**) in glioma patients in comparison to healthy controls (TFCE corrected, P-value < 0.05). Our analyses also showed atrophied Brodmann areas 4, 6, and 31, but this was not significant ($P > 0.05$).

As indicated by the TBSS and along-tract statistics results, glioma migration through the body of the CC involved the left side precentral gyrus and also increased the FA values of the corticostriatal tracts, which could be the reason for the right brainstem hypertrophy.

However, WM analyses of the corticospinal tracts or internal capsule gave no significant results to justify the probable reason for the left side hypertrophy of the brainstem.

Altogether, all the results investigated by WM and gray matter analyses would indicate that the brainstem hypertrophy occurred because of the glioma migration (see **Figure 7**), and in order to trace this transition *in vivo*, future longitudinal investigations are needed.

DISCUSSION

Present Study

In the present study, we had seven patients with right precentral gyrus glioma, and the pathological examinations verified the tumor type. Following our hypothesis diagram (see **Figure 1**) and in keeping with previous literature, we predicted glioma migration through the CC from the left side precentral gyrus into its contralateral side. Subsequently, this probable migration would change in cortical regions, which needs evaluation.

In order to evaluate cortical changes due to glioma migration, we used VBM analyses, which allowed us to determine the whole-brain cortical volume. Interestingly, both sides of the brainstem showed hypertrophy, which could be a reflection of glioma cell aggregation in this region.

[2]http://github.com/johncolby/along-tract-stats

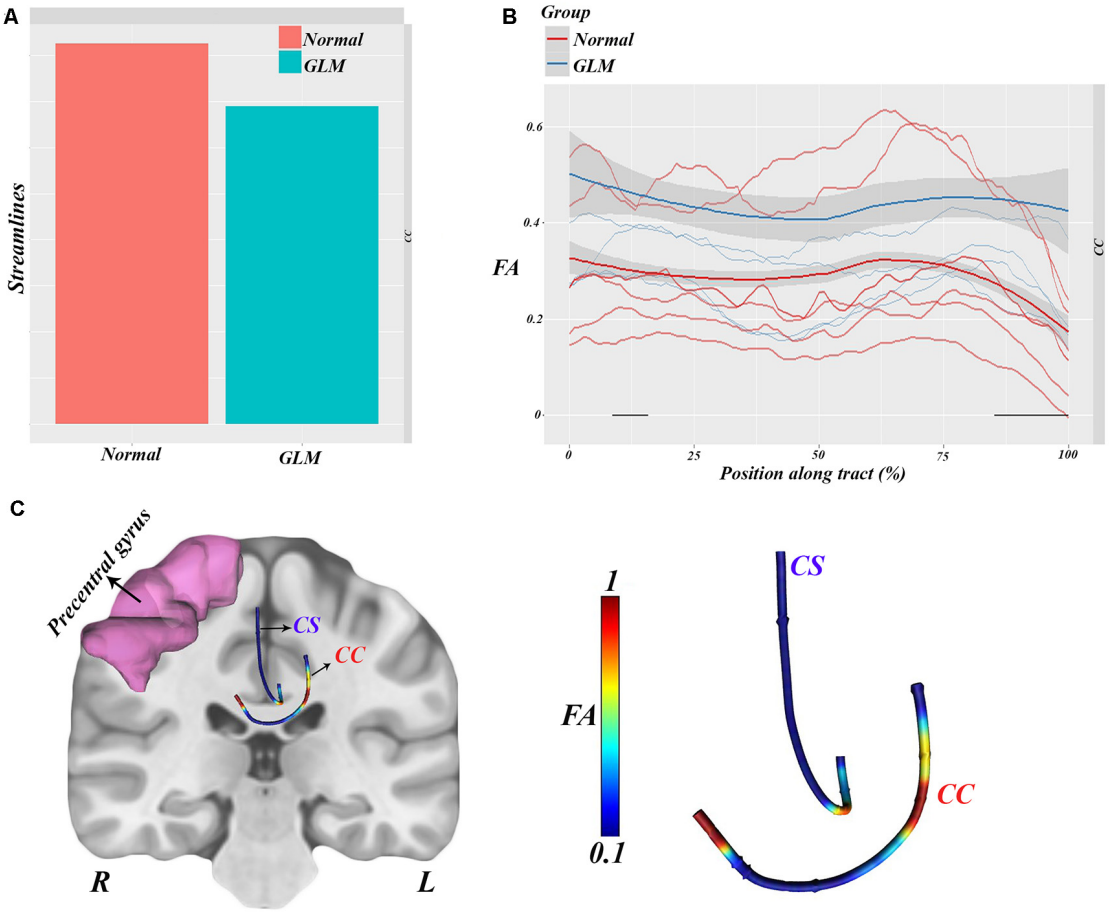


FIGURE 4 | Along-tract statistical analyses showing significantly lower streamlines in glioma patients relative to normal subjects (**A**). FA values along CC were analyzed between groups, which showed a remarkable difference (higher FA along streamlines) in patients vs. controls (**B**). All the statistical features visualized show the CC implication as a hub between hemispheres (**C**).

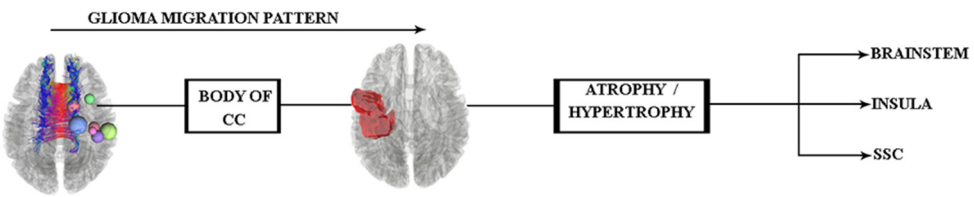


FIGURE 5 | Hypothesis of glioma migration from the right precentral gyrus to the contralateral side updated by the along-tract statistics results, which reveal this transition through the body of the CC.

TBSS analyses were performed for the prediction of glioma migration within whole-brain WM tracts, and the results supported our hypothesis of probable glioma migration along the CC to the right-side superior corona radiata. The involvement of WM tracts in glioma tumor cell migration has been shown to be a significant process (Pedersen et al., 1995; Bjerkvig et al., 1997; Soroceanu et al., 1999), which is in line with our TBSS results on superior corona radiata of the left side hemisphere with WM fiber maturation.

The involvement of the CC, which is known as a Scherer's structure in this process (Scherer, 1938), needs investigation *via* a complementary analysis to TBSS, which is carried out by along-tract statistics. The results showed maturation of the body of the CC, which is in line with TBSS results that showed no significant WM demyelination signs.

As stated in our study's hypothesis (see **Figures 1**, **5**), probable cortical changes would occur as a consequence of glioma migration from the right precentral gyrus to its contralateral side,

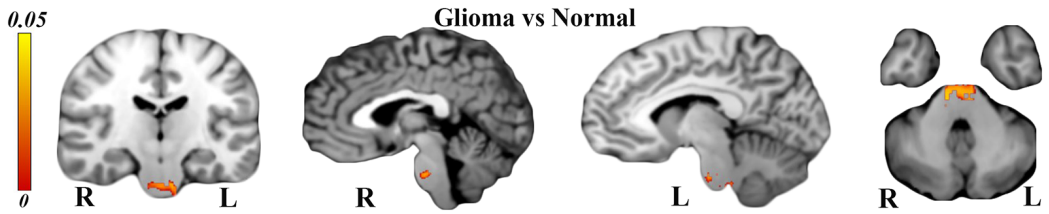


FIGURE 6 | Voxel-based morphometry (VBM) analyses showing significantly hypertrophied brainstem on both sides [$P < 0.05$ and threshold-free cluster enhancing (TFCE) corrected] in glioma patients relative to normal subjects.

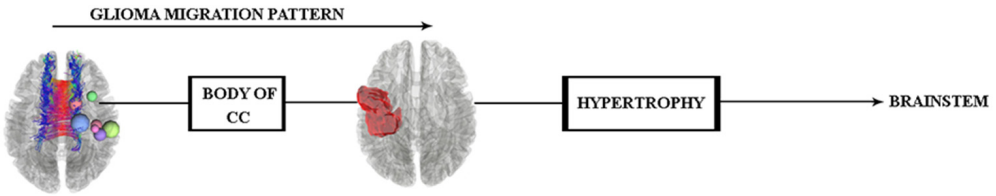


FIGURE 7 | Glioma migration hypothesis, which shows the transition pattern along the body of the CC and probable influence (hypertrophy) on both sides of the brainstem.

but to the best of our knowledge, no previous glioma cell invasion to the brainstem due to this isolated process has been reported.

White Matter Tract Maturation in the Glioma Migration Process

Consistent with the WM analysis results, WM maturation was seen in the right corticostriatal pathways, left superior corona radiata, and the body of the CC, which showed significantly increased FA and decreased MD values. The lack of demyelination signs in WM tracts would be a reflection of glioma cell migration along WM tracts to the contralateral side of the tumor's location, rather than degeneration.

However, we excluded the patients with previous metastatic tumor history or high-grade gliomas, and our WM analyses are in line with a study that reported higher MD and lowered FA in the peritumoral region of metastatic tumor patients (Holly et al., 2018). Interestingly, higher FA and lower MD values were observed in the ipsilateral peritumoral regions of glioma patients.

A few DTI studies on high-grade glioma patients have also claimed to show signs of glioma cell migration through the CC into the contralateral side (Price et al., 2004). In an investigation on 31 high-grade glioma patients utilizing DTI, the tumor-related area of the CC showed reduced FA and increased ADC values (Kallenberg et al., 2013).

These changes in anisotropy and diffusivity are in line with our results, which suggest tumor spread along the CC, but exclusion of high-grade glioma patients and analysis of the CC *via* along-tract statistics for increasing the accuracy are the novel aspects of our study.

In contrast to our study, in an investigation of glioma infiltrating from edema tissue, no significant difference was seen in FA and MD. However, the study population contained 18 glioblastomas and 22 metastatic tumors, which could be a limitation for this study (Hoefnagels et al., 2014). In addition, the DTI study was carried out between the tumoral and peritumoral site and the contralateral side, which would be another significantly different approach from what we claimed in this study.

Hypertrophy of Brainstem and Glioma Migration

The propagation of tumor cells throughout the whole brain and its overall invasion pattern is not yet clear, and more studies are needed for investigation in this context.

In accordance with the related symptoms of precentral gliomas, the involvement of the cortical homunculus (Penfield and Boldrey, 1937) is expected, and obviously, the main sensory projections from the brainstem into the cortical homunculus are also prominent. In keeping with these considerations, gray matter analyses in our study showed significant hypertrophy of the brainstem in glioma patients relative to healthy subjects.

This result may be a reflection of the tumor cell migration, which is in line with a histopathological investigation on mouse models that reported the accumulation of tumor cells in regions far from the transplant side (Mughal et al., 2018).

However, with the dispersal of glioma cells through the mouse brains, surprisingly, the hippocampus was remarkably free of glioma cell infiltration, which may be in line with our cortical analyses *via* VBM that showed no involvement of both sides of the hippocampus in this process.

Knowledge of tumor cell aggregation in located regions of the brain is also important in radiotherapy, so that the structures with accumulated tumor cells could receive higher dosage radiation, and the tumor-free structures could be relatively spared (Nourallah et al., 2017).

However, glioma invasion has been speculated to occur by invasion along vessels in the perivascular spaces (Cuddapah et al., 2014), but the present study and our results may not support this hypothesis. Nevertheless, more investigation with different approaches is needed to elucidate this process.

Precision Medicine

Precision medicine is currently known as a treatment paradigm considering the molecular and cellular features of a tumor along with its supplementary properties, such as genetics, in order to create a tailor-made treatment (Yates et al., 2018).

According to the World Health Organization (WHO), the glioma grading system ranges from stages I to IV, in which the lower grade (I–II) gliomas are considered to be benign tumors, and the higher-grade gliomas (III–IV) are considered malignant (Louis et al., 2007; Barchana et al., 2012; Ostrom et al., 2014).

However, there is no distinct genetic profile for benign or malignant gliomas to promote tailored therapies for these tumors, so producing new information about glioma cell properties such as migration, behavior, and genetics would be a crucial step toward the realization of precision medicine.

These considerations along with the modern developed multidisciplinary investigation methods, such as genetic imaging, would be a suitable approach to complement other cellular, molecular, and genetic approaches (Mackey et al., 2016).

Large-scale distributed analyses of MRI scans combined with a voxel-wise genome-wide association approach would be a useful tool in the future to prescribe tailored therapy for gliomas. However, these approaches are currently being adopted in the study of neurological disorders, and the results are quite promising (Thompson et al., 2010; Jahanshad et al., 2018).

Altogether, in gathering glioma genetic, cellular, and behavior imaging in order to reach a purpose-based criterion for future glioma precision therapy, taking advantage of previous investigations on neurological disorders using the same approach is obviously needed (Hampel et al., 2017, 2018; Titova and Chaudhuri, 2017; Ferretti et al., 2019).

CONCLUSION

In the present study, the glioma migration hypothesis was evaluated using DTI and structural MRI analyses. Based on the tumor location (precentral gyrus), the body of the CC showed significant alterations according to the DTI analyses. In addition, both sides of the brainstem also showed hypertrophy, which could be a reflection of glioma cell aggregation far from the primary tumor area.

It should be mentioned that the sample size of the present study is a major limitation, and future investigations need to overcome this shortcoming. Genome-wide association analyses in keeping with multimodal imaging analyses, such as genetic imaging, would be a practical approach to moving forward in precision tailored therapy.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to shahryarpajavand@outlook.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Skull Base Research Center, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

AP acquired, analyzed, and explained the data, drafted the manuscript and revised it. GS is the main neurosurgeon of the cases. AP, HH, SN, and TM revised the manuscript.

FUNDING

The current case report was supported by Skull Base Research Center, Department of Neurosurgery, Loghman-e Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran (Grant No. 133362-5445).

ACKNOWLEDGMENTS

We would like to acknowledge Iranian National Brain Mapping Laboratory (NBML), Tehran, Iran, for providing data acquisition service for this research. Also, we would like to thank the Clinical Research Development Unit (CRDU) of Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran for their support, cooperation, and assistance throughout the period of study (Grant No. 133362-5445).

REFERENCES

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Bach, M., Laun, F. B., Leemans, A., Tax, C. M., Biessels, G. J., Stieltjes, B., et al. (2014). Methodological considerations on tract-based spatial statistics (TBSS). *Neuroimage* 100, 358–369. doi: 10.1016/j.neuroimage.2014.06.021

Barchana, M., Margaliot, M., and Liphshitz, I. (2012). Changes in brain glioma incidence and laterality correlates with use of mobile phones—a nationwide population based study in Israel. *Asian Pac. J. Cancer Prev.* 13, 5857–5863. doi: 10.7314/apjcp.2012.13.11.5857

Bjerkvig, R., Lund-Johansen, M., and Edvardsen, K. (1997). Tumor cell invasion and angiogenesis in the central nervous system. *Curr. Opin. Oncol.* 9, 223–229. doi: 10.1097/00001622-199709030-00002

Burger, P. C., Heinz, E. R., Shibata, T., and Kleihues, P. (1988). Topographic anatomy and CT correlations in the untreated glioblastoma multiforme. *J. Neurosurg.* 68, 698–704. doi: 10.3171/jns.1988.68.5.0698

Cayre, M., Canoll, P., and Goldman, J. E. (2009). Cell migration in the normal and pathological postnatal mammalian brain. *Prog. Neurobiol.* 88, 41–63. doi: 10.1016/j.pneurobio.2009.02.001

Colby, J. B., Soderberg, L., Lebel, C., Dinov, I. D., Thompson, P. M., and Sowell, E. R. (2012). Along-tract statistics allow for enhanced tractography

analysis. *NeuroImage* 59, 3227–3242. doi: 10.1016/j.neuroimage.2011.11.004

Cuddapah, V. A., Robel, S., Watkins, S., and Sontheimer, H. (2014). A neurocentric perspective on glioma invasion. *Nat. Rev. Neurosci.* 15, 455–465. doi: 10.1038/nrn3765

Ferretti, M. T., Santuccione-Chadha, A., and Hampel, H. (2019). Account for sex in brain research for precision medicine. *Nature* 569:40. doi: 10.1038/d41586-019-01366-5

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., and Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* 14, 21–36. doi: 10.1006/nimg.2001.0786

Hampel, H., O'Bryant, S. E., Durrleman, S., Younesi, E., Rojkova, K., Escott-Price, V., et al. (2017). A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. *Climacteric* 20, 107–118. doi: 10.1080/13697137.2017.1287866

Hampel, H., Vergallo, A., Giorgi, F. S., Kim, S. H., Depypere, H., Graziani, M., et al. (2018). Precision medicine and drug development in Alzheimer's disease: the importance of sexual dimorphism and patient stratification. *Front. Neuroendocrinol.* 50, 31–51. doi: 10.1016/j.yfrne.2018.06.001

Hochberg, F. H., and Pruitt, A. (1980). Assumptions in the radiotherapy of glioblastoma. *Neurology* 30, 907–911. doi: 10.1212/wnl.30.9.907

Hoefnagels, F. W., De Witt Hamer, P., Sanz-Arigita, E., Idema, S., Kuijer, J. P., Pouwels, P. J., et al. (2014). Differentiation of edema and glioma infiltration: proposal of a DTI-based probability map. *J. Neurooncol.* 120, 187–198. doi: 10.1007/s11060-014-1544-9

Holly, K. S., Fitz-Gerald, J. S., Barker, B. J., Murcia, D., Daggett, R., Ledbetter, C., et al. (2018). Differentiation of high-grade glioma and intracranial metastasis using volumetric diffusion tensor imaging tractography. *World Neurosurg.* 120, e131–e141. doi: 10.1016/j.wneu.2018.07.230

Jahanshad, N., Roshchupkin, G., Faskowitz, J., Hibar, D. P., Gutman, B. A., Adams, H. H. H., et al. (2018). "Chapter one—multisite metaanalysis of image-wide genome-wide associations with morphometry," in *Imaging Genetics*, eds A. V. Dalca, N. K. Batmanghelich, L. Shen and M. R. Sabuncu (Academic Press), 1–23.

Kallenberg, K., Goldmann, T., Menke, J., Strik, H., Bock, H. C., Stockhammer, F., et al. (2013). Glioma infiltration of the corpus callosum: early signs detected by DTI. *J. Neurooncol.* 112, 217–222. doi: 10.1007/s11060-013-1049-y

Kreth, F. W., Warnke, P. C., Scheremet, R., and Ostertag, C. B. (1993). Surgical resection and radiation therapy versus biopsy and radiation therapy in the treatment of glioblastoma multiforme. *J. Neurosurg.* 78, 762–766. doi: 10.3171/jns.1993.78.5.0762

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., et al. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109. doi: 10.1007/s00401-007-0243-4

Mackey, S., Kan, K. J., Chaarani, B., Alia-Klein, N., Batalla, A., Brooks, S., et al. (2016). Genetic imaging consortium for addiction medicine: from neuroimaging to genes. *Prog. Brain Res.* 224, 203–223. doi: 10.1016/bs.pbr.2015.07.026

Mughal, A. A., Zhang, L., Fayzullin, A., Server, A., Li, Y., Wu, Y., et al. (2018). Patterns of invasive growth in malignant gliomas-the hippocampus emerges as an invasion-spared brain region. *Neoplasia* 20, 643–656. doi: 10.1016/j.neo.2018.04.001

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Nourallah, B., Digpal, R., Jena, R., and Watts, C. (2017). Irradiating the subventricular zone in glioblastoma patients: is there a case for a clinical trial? *Clin. Oncol.* 29, 26–33. doi: 10.1016/j.clon.2016.09.005

Ostrom, Q. T., Gittleman, H., Liao, P., Rouse, C., Chen, Y., Dowling, J., et al. (2014). CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007–2011. *Neuro Oncol.* 16, iv1–iv63. doi: 10.1093/neuonc/nou223

Pedersen, P. H., Edvardsen, K., Garcia-Cabrera, I., Mahesparan, R., Thorsen, J., Mathisen, B., et al. (1995). Migratory patterns of lac-z transfected human glioma cells in the rat brain. *Int. J. Cancer* 62, 767–771. doi: 10.1002/ijc.2910620620

Penfield, W., and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60, 389–440. doi: 10.1093/brain/60.4.389

Price, S. J., Burnet, N. G., Donovan, T., Green, H. A., Peña, A., Antoun, N. M., et al. (2003). Diffusion tensor imaging of brain tumours at 3T: a potential tool for assessing white matter tract invasion? *Clin. Radiol.* 58, 455–462. doi: 10.1016/s0009-9260(03)00115-6

Price, S. J., Peña, A., Burnet, N. G., Pickard, J. D., and Gillard, J. H. (2004). Detecting glioma invasion of the corpus callosum using diffusion tensor imaging. *Br. J. Neurosurg.* 18, 391–395. doi: 10.1080/02688690400005255

Scherer, H. J. (1938). Structural development in gliomas. *Am. J. Cancer* 34, 333–351.

Shapiro, W. R. (1999). Current therapy for brain tumors: back to the future. *JAMA Neurol.* 56, 429–432. doi: 10.1001/archneur.56.4.429

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31, 1487–1505. doi: 10.1016/j.neuroimage.2006.02.024

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Smith, S. M., and Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44, 83–98. doi: 10.1016/j.neuroimage.2008.03.061

Soroceanu, L., Manning, T. J. Jr., and Sontheimer, H. (1999). Modulation of glioma cell migration and invasion using Cl^- and K^+ ion channel blockers. *J. Neurosci.* 19, 5942–5954. doi: 10.1523/JNEUROSCI.19-14-05942.1999

Thompson, P. M., Martin, N. G., and Wright, M. J. (2010). Imaging genomics. *Curr. Opin. Neurol.* 23, 368–373. doi: 10.1097/WCO.0b013e32833b764c

Titova, N., and Chaudhuri, K. R. (2017). Personalized medicine in Parkinson's disease: time to be precise. *Mov. Disord.* 32, 1147–1154. doi: 10.1002/mds.27027

Yates, L. R., Seoane, J., Le Tourneau, C., Siu, L. L., Marais, R., Michiels, S., et al. (2018). The european society for medical oncology (ESMO) precision medicine glossary. *Ann. Oncol.* 29, 30–35. doi: 10.1093/annonc/mdx707

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sharifi, Pajavand, Nateghinia, Meybodi and Hasooni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 27 March 2020
doi: 10.3389/fnins.2020.00214

Deep Temporal Organization of fMRI Phase Synchrony Modes Promotes Large-Scale Disconnection in Schizophrenia

Tahereh S. Zarghami[1,2], Gholam-Ali Hossein-Zadeh[1] and Fariba Bahrami[2]*

[1] Bio-Electric Department, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran, [2] Human Motor Control and Computational Neuroscience Laboratory, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Robyn Miller,
Georgia State University, United States
Chuanjun Zhuo,
Tianjin Anding Hospital, China

****Correspondence:***
Gholam-Ali Hossein-Zadeh
ghzadeh@ut.ac.ir

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *05 December 2019*
Accepted: *27 February 2020*
Published: *27 March 2020*

Citation:
Zarghami TS, Hossein-Zadeh G-A and Bahrami F (2020) Deep Temporal Organization of fMRI Phase Synchrony Modes Promotes Large-Scale Disconnection in Schizophrenia. Front. Neurosci. 14:214. doi: 10.3389/fnins.2020.00214

Itinerant dynamics of the brain generates transient and recurrent spatiotemporal patterns in neuroimaging data. Characterizing metastable functional connectivity (FC) – particularly at rest and using functional magnetic resonance imaging (fMRI) – has shaped the field of dynamic functional connectivity (DFC). Mainstream DFC research relies on (sliding window) correlations to identify recurrent FC patterns. Recently, functional relevance of the *instantaneous phase synchrony* (IPS) of fMRI signals has been revealed using imaging studies and computational models. In the present paper, we identify the repertoire of whole-brain inter-network IPS states at rest. Moreover, we uncover a hierarchy in the temporal organization of IPS modes. We hypothesize that connectivity disorder in schizophrenia (SZ) is related to the (deep) temporal arrangement of large-scale IPS modes. Hence, we analyze resting-state fMRI data from 68 healthy controls (HC) and 51 SZ patients. Seven resting-state networks (and their sub-components) are identified using spatial independent component analysis. IPS is computed between subject-specific network time courses, using analytic signals. The resultant phase coupling patterns, across time and subjects, are clustered into eight IPS states. Statistical tests show that the relative expression and mean lifetime of certain IPS states have been altered in SZ. Namely, patients spend (45%) less time in a globally coherent state and a subcortical-centered state, and (40%) more time in states reflecting anticoupling within the cognitive control network, compared to the HC. Moreover, the transition profile (between states) reveals a deep temporal structure, shaping two metastates with distinct phase synchrony profiles. A metastate is a collection of states such that within-metastate transitions are more probable than across. Remarkably, metastate occupation balance is altered in SZ, in favor of the less synchronous metastate that promotes disconnection across networks. Furthermore, the trajectory of IPS patterns is less efficient, less smooth, and more restricted in SZ subjects, compared to the HC. Finally, a regression analysis confirms the diagnostic

value of the defined IPS measures for SZ identification, highlighting the distinctive role of metastate proportion. Our results suggest that the proposed IPS features may be used for classification studies and for characterizing phase synchrony modes in other (clinical) populations.

Keywords: resting-state fMRI, instantaneous phase synchrony, functional networks, temporal hierarchy, metastate, trajectory, schizophrenia

INTRODUCTION

Growing evidence suggests that large-scale functional connectivity (FC) is inherently *transient*. That is, functional networks reconfigure not only in response to task demands (Gonzalez-castillo and Bandettini, 2019) but also during resting state (Chang and Glover, 2010; Deco et al., 2013, 2017; Hutchison et al., 2013; Preti et al., 2017). Importantly, transient FC at rest has neuronal underpinnings, as evidenced by concurrent imaging and electrophysiological recordings (Tagliazucchi et al., 2012, 2015; Keilholz et al., 2013; Keilholz, 2014; Thompson et al., 2014) and emergent dynamics from computational models (Deco et al., 2011, 2017; Rabinovich and Varona, 2011; Cabral et al., 2014). Moreover, recent findings show that spontaneous FC patterns are quasi-periodically *recurrent* (Thompson et al., 2014), and their temporal organization is non-random (Deco et al., 2011; Vidaurre et al., 2017). Hence, dynamical organization of resting-state connectivity has been conceptualized as nonstop (and non-random) excursion through a bounded repertoire of metastable connectivity modes, called *states* (Ghosh et al., 2008; Deco et al., 2011; Cabral et al., 2017).

Neuroimaging furnishes non-invasive investigation of metastable brain phenomena, at multiple temporal and spatial scales. Specifically, functional magnetic resonance imaging (fMRI) is popular for its high spatial resolution and inherent circumvention of the inverse problem (i.e., source localization). As such, various techniques have been developed to capture regularities in the evolving spatiotemporal patterns of fMRI signals (Hutchison et al., 2013; Calhoun et al., 2014; Preti et al., 2017). Conventional dynamic FC (DFC) techniques track the statistical dependencies between neuroimaging series in successive epochs of data, which is known as the *sliding window* approach. Notably, window size has been a controversial parameter, posing a trade-off between the temporal resolution and statistical significance of the observed FC fluctuations (Hindriks et al., 2015; Leonardi and Ville, 2015; Zalesky and Breakspear, 2015). Meanwhile, a recent addition to the set of DFC measures is the instantaneous phase synchrony (IPS) of fMRI series, which eschews the window complications (Glerean et al., 2012) and has shown functional relevance in empirical and modeling studies.

Instantaneous phase synchrony is relatively new in fMRI (Laird et al., 2002; Deshmukh et al., 2004; Kitzbichler et al., 2009) despite being an established measure for analyzing electroencephalography (EEG), magnetoencephalography (MEG), and electrophysiological recordings (Tass et al., 1998; Lachaux et al., 1999, 2000). Once applied to fMRI signals, IPS reflects momentary phase alignment in the slow fluctuations of hemodynamic responses. Adopting IPS as a DFC measure can improve the temporal resolution of connectivity variations from the size of the window (typically 30–120 s) to the sampling time of the imaging modality (usually TR = 0.7–3 s).

Following evidence regarding the temporal variability of fMRI phase relationships within and between functional networks (Laird et al., 2002; Deshmukh et al., 2004; Kitzbichler et al., 2009; Chang and Glover, 2010; Glerean et al., 2012), verifying and characterizing the *spontaneous recurrence* of these patterns at rest and their functional relevance came into focus. The challenge is that, the instantaneity of IPS measures (between regions or networks) comes at the expense of noisy and irregular functional patterns, with spatiotemporal order that is not easy to capture. Early attempts relied on thresholding instantaneous phase couplings, converting them to binary patterns before further analysis (Kitzbichler et al., 2009; Ponce-Alvarez et al., 2015). In particular, Ponce-Alvarez et al. (2015) used non-negative matrix factorization (NNMF) to decompose a series of thresholded IPS matrices into the weighted sum of simpler (recurrent) components. This joint modeling and empirical study showed that phase interactions over a structurally plausible network give rise to recurrent synchronization patterns that resemble empirical arrangements – namely, the resting-state networks (Ponce-Alvarez et al., 2015). However, the threshold level ($\Delta\phi = \pi/6$) that defines phase synchrony had been chosen arbitrarily in this research. More recent studies (Cabral et al., 2017; Lord et al., 2019) avoid thresholding, but represent each bivariate IPS matrix by its first eigenvector, and apply (k-means) clustering on these summary vectors to identify recurring arrangements. This is also a practical approach that inevitably discards some of the phase coupling information, and still struggles with cluster validation (Lord et al., 2019). In short, finding natural order in IPS data has not been straightforward, despite certain simplifications.

Recently, conventional DFC[1] studies have disclosed more sophisticated spatiotemporal order in resting-state data, which remains to be verified in the phase coupling realm. That is because, any connectomic order arising from correlation-based DFC analyses does not necessarily generalize to IPS dynamics, due to their inherently different mathematical properties and timescale, moderate association between their temporal variations and non-linear relationship on a topological level (in terms of time averages) (Pedersen et al., 2018). In the following, we review some prominent DFC findings, which have not been

[1]Such as sliding window correlation and wavelet coherence analyses, which are not based solely on instantaneous phase relationships.

investigated using phase relationships and motivated the present analysis:

- Recent studies have revealed a *hierarchical organization* in the (non-random) temporal arrangement of functional states at rest (Vidaurre et al., 2017); i.e., temporal order beyond sample-to-sample transitions. Importantly, this slower temporal organization is heritable and related to personality traits (Vidaurre et al., 2017, 2019). Currently, we do not know whether the spontaneous recurrence of IPS patterns is similarly governed by a deep temporal arrangement.
- It was lately shown that the *trajectory* of connectivity evolution holds functionally relevant information, which becomes obscured in the usual cluster analyses (Miller et al., 2016). The IPS trajectory may also bear functionally relevant information, especially given its high (single-TR) temporal resolution.
- Functional integration is inherently multi-scale; that is, it coordinates neurons, micro-columns, and macroscopic regions (which form functional networks) and also organizes large-scale network interactions. Ongoing network interactions reflect higher-order functional organization in the brain, which is related to one's character and cognitive performance (Vidaurre et al., 2019) and is modulated in various disorders (Du et al., 2018; Díez-cirarda et al., 2018; Gilbert et al., 2018; Rashid et al., 2018; de Lacy and Calhoun, 2019; Espinoza et al., 2019; Fu et al., 2019). So far, a good number of IPS studies have relied on (anatomical) atlases (Ponce-Alvarez et al., 2015; Cabral et al., 2017; Lord et al., 2019) or have investigated few functional networks (Chang and Glover, 2010); some have not been concerned with the recurrence of IPS modes (Pedersen et al., 2018; Zhang et al., 2019) or their recurrence characterizations have not been purely phase-based[2] (Yaesoubi et al., 2015, 2017). Hence, current literature has not portrayed the repertoire of whole-brain inter-network IPS states based strictly on phase relationships – which could furnish new insights into the large-scale dynamic functional connectome. Notably, if large-scale phase couplings hold functional information – as speculated – they may also be modulated in neurological and psychiatric disorders.
- Schizophrenia[3] (SZ) is a severe psychiatric disorder, which has affected some 23.6 million people worldwide by 2013, with a lifetime prevalence of about 1% (Murray et al., 2016). SZ is commonly known as a connectivity disorder[4]. In 1998, Friston (1998) proposed a *disconnection hypothesis* for SZ that stemmed from positron emission tomography (PET) findings. Ever since, fMRI has expedited the examination of disrupted functional organization of the brain in SZ (Liang et al., 2006; Fornito et al., 2012; Damaraju et al., 2014; Ma et al., 2014; Van Den Heuvel and Fornito, 2014). Dysfunctional connections within and across brain regions have been frequently reported in SZ studies[5]. Importantly, SZ disconnection affects between-network interactions as well, based on prior (correlation-based) DFC analyses (Damaraju et al., 2014; Rashid et al., 2014). It remains to be investigated whether instantaneous phase coupling of functional networks reflects disconnection in SZ, and whether whole-brain network-level IPS measures hold diagnostic value in this context.

In the present work, we try to address the above questions in an empirical study. Specifically, we analyze a publicly available SZ dataset while extending IPS characterization to higher functional and temporal levels. We hypothesize that disconnection in SZ relates to the spatial and (deep) temporal organization of network level IPS modes, and to the trajectory of large-scale phase coupling patterns, in this disorder. The results on SZ show the effectiveness of the proposed approach, which may be used in other applications as well.

MATERIALS AND METHODS

Dataset and Pre-processing

We used the SZ dataset of the Center for Biomedical Research Excellence (COBRE) (Çetin et al., 2014). The dataset comprises of 72 SZ patients and 75 healthy subjects (18–65 years old). The patients had been diagnosed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) (First et al., 2002). Any subject with history of neurological disorder, mental retardation, severe head trauma, active substance abuse or dependence within the past year had been excluded from the study. Each participant was scanned at rest for 5 min on a 3-T Siemens Tim Trio scanner, and instructed to fixate on a central cross. A total of 150 (T2*-weighted) functional volumes were acquired using a gradient-echo EPI sequence (TR = 2 s, TE = 29 ms, flip angle = 75°, 33 axial slices, ascending acquisition, matrix size = 64 × 64, voxel size = 3.75 × 3.75 × 4.55 mm, field of view = 240 mm). A high-resolution T1-weighted structural image was also collected for each subject.

Functional images were pre-processed in SPM12 software[6] as follows: the first five volumes were discarded to allow for T1 equilibration; the remaining images were realigned (i.e., head motion corrected), slice-timing corrected, co-registered to the anatomical image of the subject, warped to the standard Montreal Neurological Institute (MNI) template (Collins et al., 1998), resampled to 3 mm^3 isotropic voxels and smoothed with a Gaussian kernel (FWHM = 6 mm). Global signal

[2]Yaesoubi et al. (2015, 2017) have accounted for IPS in identification of complex-valued recurrent functional modes using wavelet coherence, which contains both amplitude and phase coupling information. For more details, please refer to the "Discussion" section.

[3]Schizophrenia symptoms include delusions, hallucinations, disorganized thinking and speech, cognitive dysfunction, flat expressions, and asociality (American Psychogeriatric Association, 2000).

[4]The very term coined by the Swiss psychiatrist Eugen Bleuler (1857–1939) tends to signify a sort of "split" (schizen) in the normally integrated processes of the "mind" (phrenia) (Van Den Heuvel and Fornito, 2014).

[5]The modulated connection are mostly between the prefrontal, temporal, and parietal cortices, as well as the striatum, thalamus, amygdala, and cerebellum (Nejad et al., 2012).

[6]https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

was not removed (Murphy and Fox, 2017). Based on the motion realignment parameters, framewise displacement (FD) was computed as the total absolute head motion between consecutive time points, assuming 50-mm head radius for converting rotations to translations (Power et al., 2012). Nine subjects with maximum head translation exceeding 3 mm were removed; 12 patients with mean framewise displacement (MFD) above 0.7 mm were left out; and 5 normal females were excluded to match gender proportion. Hence, 119 subjects (68 healthy controls [HC]/51 SZ) were retained, whose demographics have been included in **Table 1**. Two-sample *t* tests on the maximum head translation, MFD, and age, plus chi-squared test on the gender proportion did not show significant difference between the two groups (uncorrected *p*-values = 0.78, 0.20, 0.82, and 0.16, respectively).

TABLE 1 | Demographics and clinical information of the participants.

	Number	Age	Female/male
Healthy controls	68	35.4 ± 11.8	18/50
Schizophrenia patients	51	35.9 ± 13.4	8/43

Spatially Constrained Spatial Independent Component Analysis

In order to achieve a refined functional parcellation of the brain, we used aggregate functional networks from Allen et al. (2014) as *spatial priors* to run (spatially) constrained spatial ICA (csICA) (Lin et al., 2010) on each subject. This approach has a number of advantages: (1) csICA respects spatial variability of the networks across subjects – unlike fixed network templates; (2) csICA optimizes independence of the ICs (i.e., the networks) at the subject level and simultaneously maintains their correspondence across the group; (3) csICA has superior performance[7] compared to a number of other established (principal component/regression based) back-reconstruction techniques (Calhoun et al., 2001; Beckmann et al., 2009; Erhardt et al., 2011); (4) csICA allows for the propagation of aggregate networks to individuals who were not included in the original group analysis. This follows because the spatial priors of csICA need to be only "partially correct" in this method (Lin et al., 2010).

Hence, we adopted the 50 aggregate networks (estimated from a 100-component[8] group spatial ICA analysis on 405 subjects in Allen et al. (2014), and implemented constrained spatial ICA (Lin et al., 2010) on the preprocessed data of every participant, using the Group ICA of fMRI Toolbox (GIFT[9]). Notably, the artifactual components (i.e., the physiological, head movement, and imaging artifact components) had already been identified and excluded in Allen et al. (2014), such that the remaining 50 functional parcels comprise sub-components of reproducible large-scale resting-state networks (Kiviniemi et al., 2009; Abou-Elseoud et al., 2010; Allen et al., 2011; Du et al., 2019). These networks include the subcortical (SC), auditory (AUD), somatomotor (SM), visual (VIS), cognitive control (CC), default mode network (DMN), and cerebellum (CB) (**Figure 2C**). For more details about the spatial maps of the networks and their peak coordinates, please see Supplementary Figure S2 and Supplementary Table S1 in Allen et al. (2014). For a schematic illustration of csICA please refer to the flowchart in **Figure 1A**.

Post-processing

The time courses associated with subject-specific ICs underwent additional post-processing to remove residual motion and artifactual sources of variation. That is, the time series were detrended to remove low-frequency scanner drift, orthogonalized with respect to the subject's estimated motion parameters and their derivatives, and despiked to replace outlier points. Despiking was performed using AFNI's 3dDespike algorithm which detects outlier time points (based on the median absolute deviation) and replaces them with interpolated values from a third-order spline fitted to the adjacent time points (Allen et al., 2014).

Static Functional Network Connectivity

Average connectivity among the networks over the whole session was computed (for each subject) as the sample covariance matrix of the network time courses. Since resting-state FC is primarily shaped by low-frequency fluctuations of fMRI signals (Cordes et al., 2001), the network time series were bandpass filtered between (0.01 and 0.15) Hz (using a 5th-order Butterworth filter) before computing the static functional network connectivity (sFNC) matrices. Furthermore, sFNC was calculated for each group by averaging over their respective subjects. To highlight group differences, two-sample *t* tests were performed on each entry of the (Fisher *z*-transformed) sFNC matrices, and *p*-values were adjusted for multiple comparison using the false discovery rate (FDR) approach (Benjamini and Hochberg, 1995). Following convention, $q_{\mathrm{FDR}} < 0.05$ was considered statistical significance.

Instantaneous Phase Synchrony

The analytic representation of a real valued signal $x(t)$ is a complex signal $x_a(t)$, with no negative frequency components. This complex signal can be constructed from the real signal using the Hilbert transform as follows (Boashash, 1992):

$$\mathrm{x_a}\,(\mathrm{t}) = \mathrm{x}\,(\mathrm{t}) + \mathrm{j}\mathcal{H}\,\{\mathrm{x}\,(\mathrm{t})\} \tag{1}$$

where $\mathcal{H}\{.\}$ denotes the Hilbert transform and j is the imaginary unit. The main property of analytic signal, $\mathrm{x_a}\,(\mathrm{t})$, is that its Fourier transform is the same as that of the original real-valued signal, but only covers the positive frequencies. As a result, this complex (analytic) signal converts the original time series into two separate time series (which are the real and imaginary parts), from which useful aspects of the signal can be studied – such

[7] csICA improves spatial and temporal accuracy, subject specificity, statistical independence, and group correspondence of the recovered networks (Lin et al., 2010; Du and Fan, 2013; Salman et al., 2019).

[8] Spatial ICA with high model order can achieve a detailed functional parcellation of cortical and subcortical structures with known anatomical and functional identity (Kiviniemi et al., 2009; Smith et al., 2009; Abou-Elseoud et al., 2010; Allen et al., 2014). However, model orders >100 have been shown to decrease ICA repeatability, as evaluated by ICASSO's cluster quality index (Himberg et al., 2004; Abou-Elseoud et al., 2010).

[9] http://trendscenter.org/software/gift/

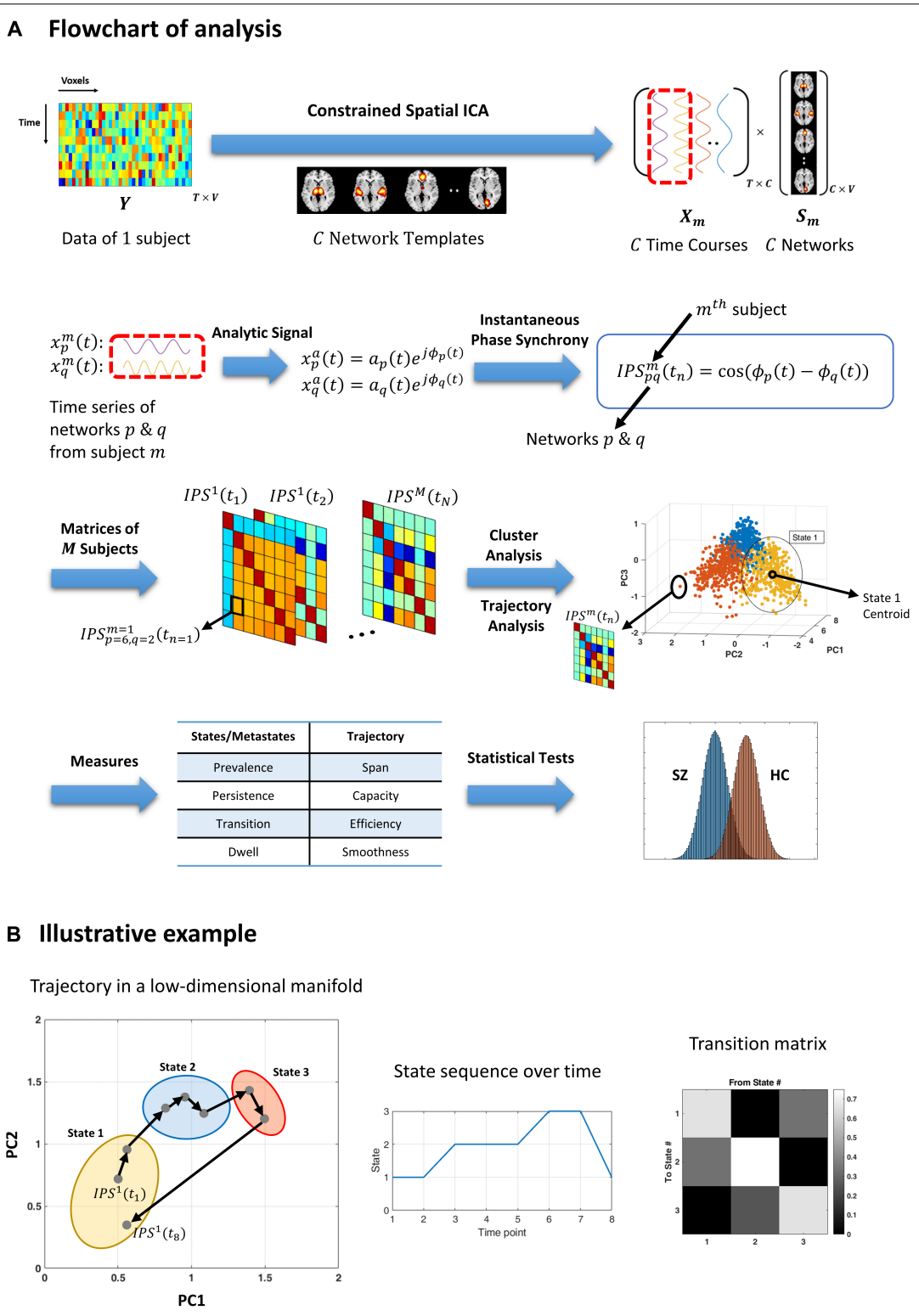


FIGURE 1 | Flowchart of analysis and an illustrative example. **(A)** Resting-state fMRI data from each subject is decomposed using (spatially) constrained spatial independent component analysis (csICA) into C sub-networks (C = 50), based on the aggregate network templates from Allen et al. (2014). Time series corresponding to subject-specific networks are converted to complex analytic signals, from which instantaneous phases are extracted. Instantaneous phase synchrony (IPS) is computed as cosine of the difference between momentary phases, for each pair of networks and per subject. Time indexed IPS matrices of all subjects are clustered using spectral clustering. The three-dimensional plot (third row, right panel) shows an exemplar stack of IPS matrices projected on a

(Continued)

FIGURE 1 | Continued
low-dimensional manifold, which is spanned by the first three principal components (PC). Cluster analysis assigns each data point (which is an IPS matrix) to one cluster centroid (i.e., a state). The color of each filled circle reflects the state label assigned to that IPS pattern, where only three states where assumed in this example. Trajectory analysis is conducted on the same IPS matrices (per subject), independently from cluster analysis. A number of measures are defined (in **Table 2**) to quantify the state and trajectory characteristics. These IPS measures are quantified and statistically compared between healthy controls (HC) and schizophrenia (SZ) patients. **(B)** An illustrative example highlighting the difference between cluster and trajectory analyses. Gray dots represent eight IPS matrices projected on their first two principal components. Arrows indicate temporal progression of the patterns, for a hypothetical session. The colorful ellipses correspond to three identified clusters. Hence, all the data points encompassed by one ellipse bear the same state label. The output of cluster analysis is a state sequence and a corresponding transition matrix (middle and right panels). Conversely, trajectory analysis focuses on sample-to-sample variations in the IPS patterns. For instance, the sum of the (L1) norms of all the black arrows constitutes the *trajectory length*. Also, the (L1) distance between the two farthest points is defined as the *span* of the trajectory. For the full list of IPS measures, please refer to **Table 2**.

as instantaneous phase. In particular, for a *narrowband* signal expressed as $x(t) = a(t)\cos(\phi(t))$, the corresponding analytical representation is (Bedrosian, 1962):

$$x_a(t) = a(t)\exp\left(j\phi(t)\right) \tag{2}$$

where $a(t)$ and $\phi(t)$ are instantaneous envelope and instantaneous phase, respectively. Hence, for two real narrowband signals ($x_p(t)$ and $x_q(t)$), the difference between their instantaneous phases reflects their phase synchrony. To get a normalized measure between -1 and $+1$, the cosine of the phase difference is considered (Glerean et al., 2012; Cabral et al., 2017):

$$IPS_{pq}(t) = \cos(\phi_p(t) - \phi_q(t)) \tag{3}$$

So, when the instantaneous phases of two signals fully align at time t, IPS(t) reaches its maximum value of $+1$; conversely, for 180° phase difference, this measure falls to -1, reminiscent of the familiar correlation and anti-correlation notions (herein called coupling and anticoupling).

The post-processed network time courses in our study were bandpass filtered in the 0.01–0.08 Hz range (Cabral et al., 2017) using Parks-McClellan linear-phase finite impulse response filter (Mcclellan et al., 1973; Glerean et al., 2012). The corresponding analytic signals were used to compute instantaneous phase series (Eqs. 1 and 2). Subsequently, IPS was calculated for each pair of networks (p and q) at each time point (using Eq. 3) to construct time-dependent IPS matrices (**Figure 1A**). These phase coupling matrices reflect the momentary phase similarities of $C = 50$ resting-state (sub)networks, for each subject. The outstanding task was to characterize the recurrence of these IPS patterns.

Identifying Recurrent IPS Patterns

In order to identify the main modes of phase coupling, across time and subjects, we used cluster analysis. Clustering relies on some measure of (dis)similarity between the samples, which are symmetric IPS matrices in this context. A common measure of similarity is *cosine similarity*:

$$\text{cosine similarity}\left(\overrightarrow{IPS^m}(t_i), \overrightarrow{IPS^n}(t_j)\right) = \frac{\overrightarrow{IPS^m}(t_i) \cdot \overrightarrow{IPS^n}(t_j)}{\left\|\overrightarrow{IPS^m}(t_i)\right\| \left\|\overrightarrow{IPS^n}(t_j)\right\|} \tag{4}$$

where $\overrightarrow{IPS^m}(t_i)$ is a vector holding the unique (upper or lower triangular) entries in the symmetric IPS matrix of the m^{th} subject, at time t_i; the dot in the numerator stands for dot product and $||\cdot||$ is the magnitude (i.e., Euclidean norm). We work with a related measure called *angular similarity,* defined as follows:

$$\text{angular similarity} = 1 - \text{angular distance} \tag{5}$$

$$\text{angular distance} = \frac{\cos^{-1}\left\{\text{cosine similarity}\right\}}{\pi} \tag{6}$$

Angular similarity varies between 0 and 1, and this positivity is useful for constructing an unsigned graph[10] (to perform spectral clustering, in what follows). Moreover, angular distance satisfies the triangle inequality (unlike cosine distance) and qualifies as a proper distance metric. Note that we have not performed data reduction on the phase coupling matrices prior to (angular) similarity computation, in order to retain maximal information about the phase synchrony patterns.

As such, for M subject sessions, each comprising N time samples, an $MN \times MN$ (angular) similarity matrix was computed. A representative portion of this matrix has been illustrated in **Figure 3A**. From this similarity matrix, a (weighted undirected unsigned) graph was constructed, which has been schematically depicted in **Figure 3B**. Graph representation furnishes the use of concepts and tools from graph theory. In particular, *spectral graph theory* is the mathematical field of studying graph properties from the eigenvectors and eigenvalues of their associated adjacency and Laplacian[11] matrices. For instance, optimally partitioning a graph into a number of subgraphs is an NP-hard problem for which *spectral clustering* has offered computationally tractable (approximate) solutions (Shi and Malik, 2000; Luxburg, 2007).

We used the spectral clustering algorithm of Ng et al. (2002). In this method, a low-dimensional representation of the pairwise similarity matrix is constructed from the k smallest eigenvectors[12] of the normalized graph Laplacian matrix. These vectors constitute the columns of a new matrix, whose (normalized) rows serve as the feature vectors for a clustering algorithm. We used Gaussian mixture model (GMM) at this stage. To establish the optimal (k) number of clusters, the same procedure was

[10] An unsigned graph is more straightforward to manipulate and interpret algebraically than a signed graph (Gallier, 2016).

[11] Laplacian matrix is defined as L = D − A, where A is the adjacency matrix and D is a diagonal matrix bearing the node degrees (Chung and Graham, 1997).

[12] Shorthand for the eigenvectors corresponding to the k smallest eigenvalues.

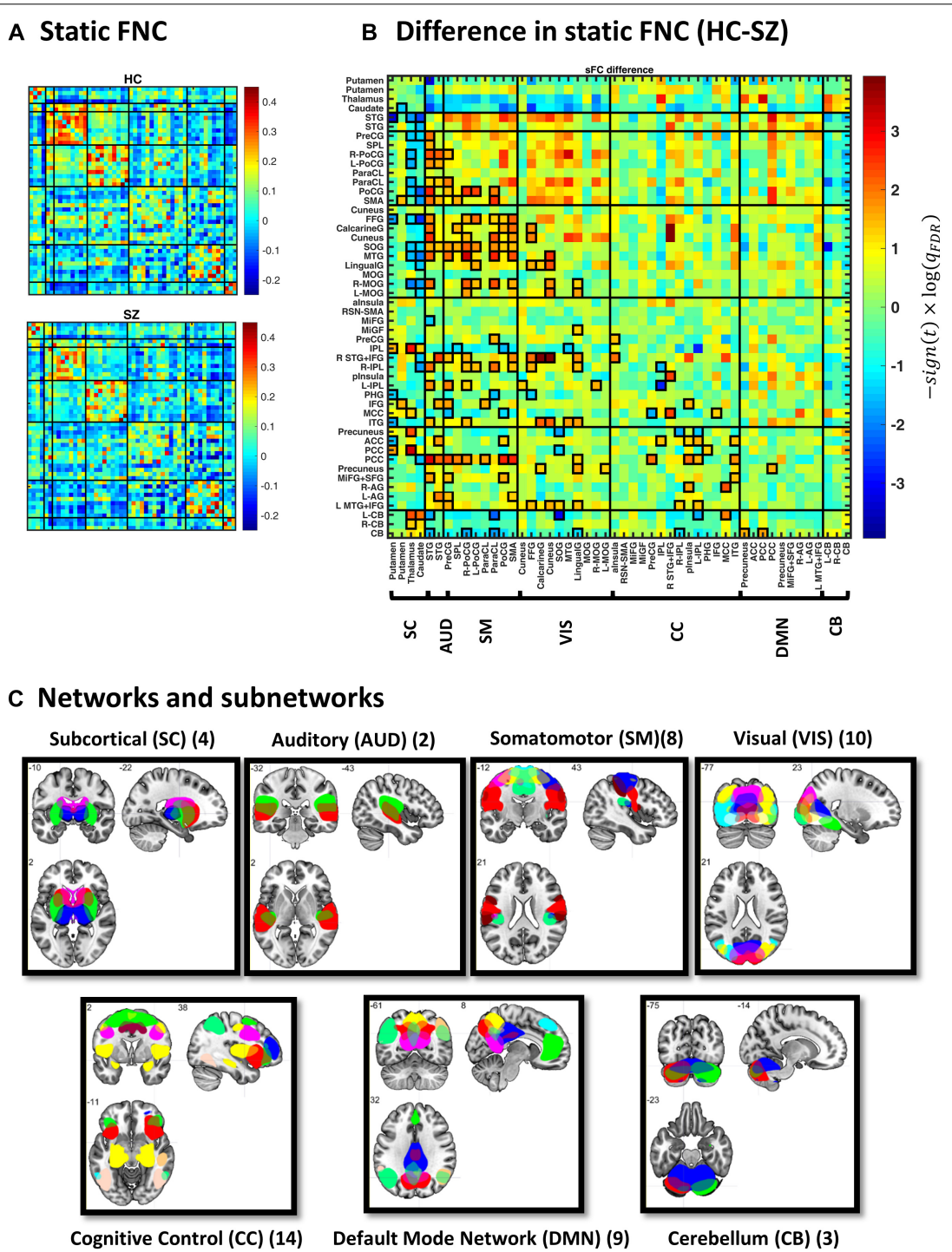


FIGURE 2 | Static functional network connectivity (sFNC) analysis. **(A)** Average sFNC pattern for the healthy control (HC, upper panel) and schizophrenic (SZ, lower panel) group. Diagonal (unity) entries have been removed for better visualization. **(B)** Difference in sFNC (HC–SZ). The color bar denotes $-10\log(q_{FDR}) \times sign(t - statistic)$, following two-sample *t* test on Fisher *z*-transformed correlation values ($z = \text{atan}h(r)$). The 50 sub-networks and their organization into seven major networks have been reflected in the labels. Sub-network labels reflect the brain region with peak amplitude and refer to bilateral activations unless specified as left (L) or right (R). See Supplementary Figure S2 and Supplementary Table S1 in Allen et al. (2014) for the peak coordinates. **(C)** Spatial maps of sub-networks grouped into seven networks based on their anatomical and functional properties. Abbreviations: STG, superior temporal gyrus; PreCG, precentral gyrus; PoCG, postcentral gyrus; SMA, supplementary motor area; ParaCL, paracentral lobule; SPL, superior parietal lobule; MTG, middle temporal gyrus; FFG, fusiform gyrus; MOG, middle occipital gyrus; SOG, superior occipital gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; MCC, middle cingulate cortex; pInsula, posterior insula; MiFG, middle frontal gyrus; IFG, inferior frontal gyrus; aInsula, anterior insula; PHG, parahippocampal gyrus; PCC, posterior cingulate cortex; AG, angular gyrus; ACC, anterior cingulate cortex; SFG, superior frontal gyrus; CB, cerebellum.

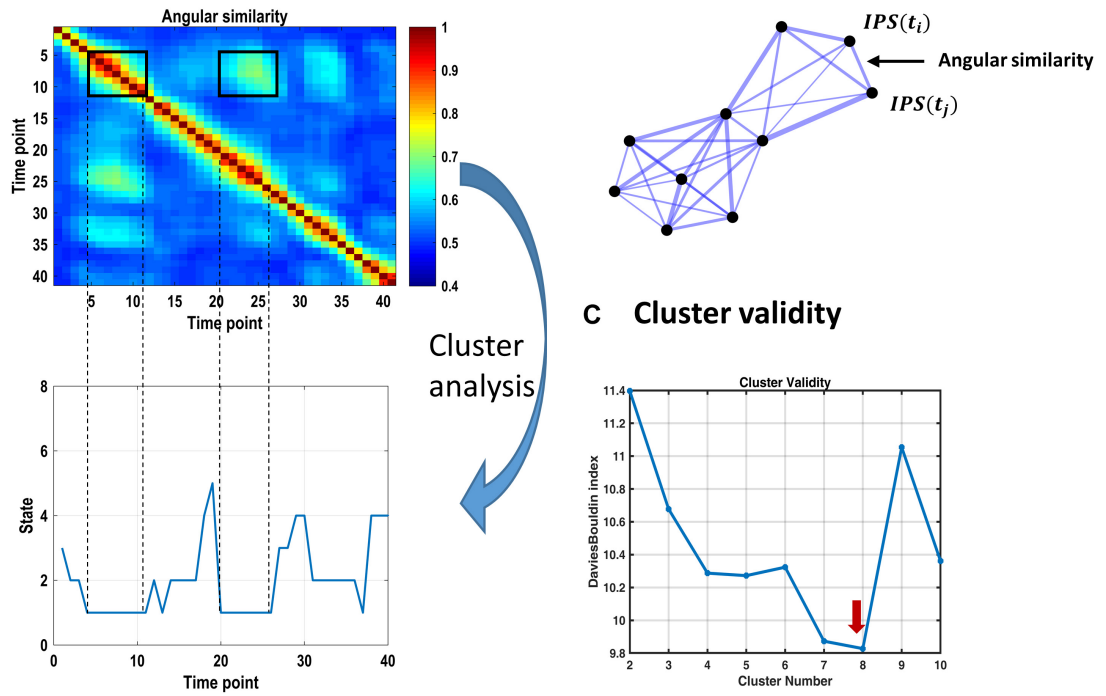


FIGURE 3 | Identifying recurrent IPS patterns through cluster analysis. **(A)** Top plot shows a representative portion of the (symmetric) angular similarity matrix for 40 consecutive IPS patterns. That is, each entry denotes the angular similarity between two time-indexed IPS matrices (Eq. 5). Recurrent transient patterns are automatically placed in the same cluster, following spectral clustering. For instance, IPS patterns realized between time points 5 and 11 (enclosed by a black square on the diagonal) are highly similar to each other and to the patterns emerging over time samples 20–26. Hence, IPS matrices from both periods identify with the same cluster number (i.e., state 1), as illustrated in the lower panel. **(B)** Schematic of an (undirected weighted) graph associated with an angular similarity matrix. Each node corresponds to a time-indexed IPS matrix, and edge thickness reflects angular similarity. Using spectral clustering, this graph can be optimally divided into well-connected groups of nodes that are plausibly separated from each other, comprising the clusters of interest. **(C)** Cluster validity is established using the Davies–Bouldin index (which is the ratio of within-cluster scatter to between-cluster distance) quantified for a range of 2–10 clusters. The lowest value of this index reflects the most plausible clustering solution, which is eight for our data. The associated cluster centroids are illustrated in **Figure 4**.

repeated for k = 2–10 components and Davies–Bouldin index[13] (Davies and Bouldin, 1979) was used to score cluster validity (**Figure 3C**). The optimal number of components (8 in this case) was selected and the resultant cluster centroids comprised the dominant *states* of inter-network phase coupling, across time and subjects (**Figure 4A**). Thereafter, each time-indexed IPS matrix was assigned to the most plausible centroid (i.e., state) based on its posterior probability of assignment (which is called hard assignment in the clustering literature).

Characterization of IPS Transient Behavior

Following hard/crisp clustering, every IPS matrix (at each time point) bears one state label $\in \{1:8\}$. The resultant sequence of states per subject (see **Figure 3A**, lower panel) was used to characterize the transient expression of phase synchrony modes over time. Specifically, we call the fraction of IPS patterns assigned to each state the *prevalence* of that state. Other properties of interest are the probability of *transition* between states, the probability of remaining in a particular state (called *dwell*), and the average period over which a state is held continuously (called *persistence*). These measures are summarized in **Table 2**. They were computed per subject and compared between (HC and SZ) groups using permutation-based two-sample t tests, followed by FDR correction for multiple comparisons.

We also investigated the relationship between the transition probability and the similarity of states. It has been suggested that transition probability between states is related to their connectomic similarity; i.e., the brain undergoes smooth connectivity changes over time (Vidaurre et al., 2017). Specifically, a significant positive correlation between the connectomic similarity of (correlation-based) functional states and the transition probability between them has been reported (Vidaurre et al., 2017). We intended to test the same hypothesis about phase synchrony modes and their transition profile. We tested, additionally, for any correlation between transition probabilities and state prevalence (and persistence) similarities.

[13]The Davies–Bouldin criterion is based on the ratio of within-cluster scatter to between-cluster distance. Hence, the optimal clustering solution would achieve the lowest value of this index.

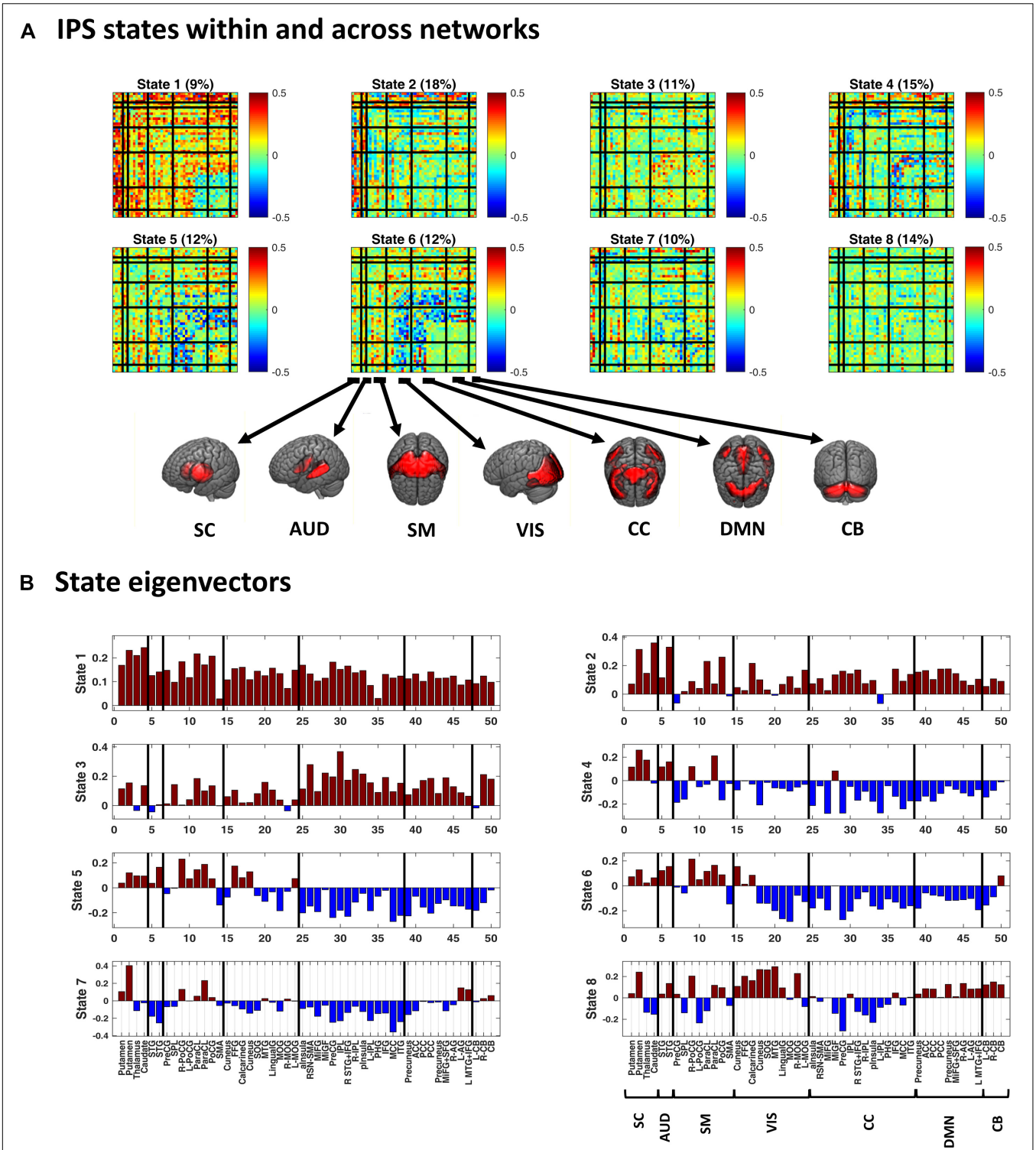


FIGURE 4 | Continued

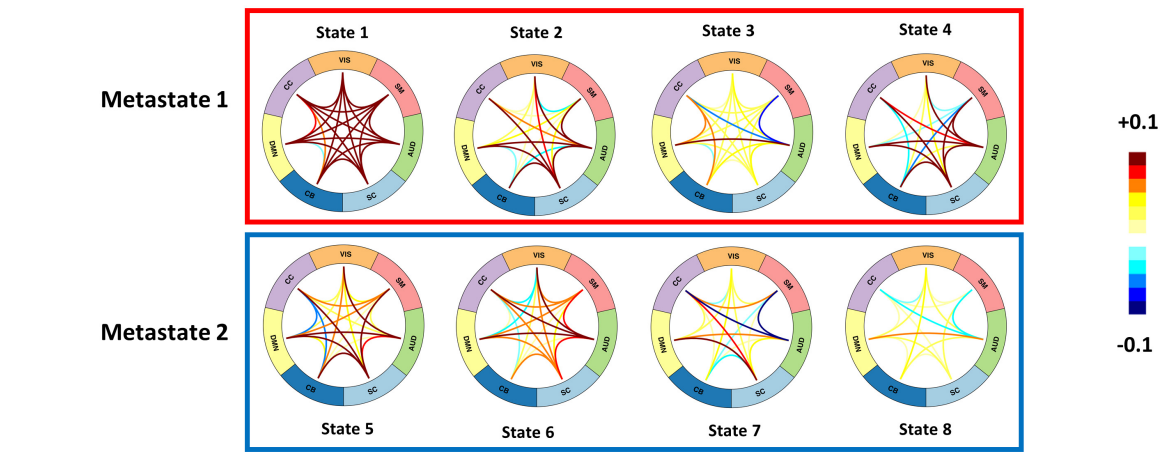


FIGURE 4 | Instantaneous phase synchrony (IPS) states, in three representations. **(A)** Eight cluster centroids, depicted in matrix format. Each centroid stands for an IPS state. The relative expression of each state (across time and subjects) is printed above it. Surface colors reflect IPS values (Eq. 3). Matrix representation offers a detailed account of the phase coupling structure of each state, within and across seven functional networks. **(B)** The first eigenvector of each state is presented as a bar plot. Positive and negative entries constitute two detached communities. **(C)** State connectograms are concise representations of average inter-network phase couplings, where the average is taken over sub-components of each network. For instance, the central role of the subcortical (SC) network in state 2 is readily recognized by noting the warm colored arcs extending from SC to other networks. SC, subcortical; AUD, auditory; SM, somatomotor; VIS, visual; CC, cognitive control; DMN, default mode network; CB, cerebellum. For the sub-network abbreviations, please refer to the caption of **Figure 2**.

TABLE 2 | Measures of instantaneous phase synchrony (IPS).

Measure	Description
Prevalence	Probability of occurrence of each state
Persistence	Mean lifetime of each state (in seconds)
Transition Probability	Probability of switching from one state to another
Dwell	Probability of remaining in a given state
Trajectory length	Total L1 distance between successive IPS patterns
Span	Maximum L1 distance between IPS patterns
Capacity	Average L1 distance between IPS patterns
Efficiency	Ratio of Capacity to Trajectory Length
Smoothness	Average L1 similarity of successive IPS patterns

The upper section contains measures based on cluster analysis. The lower section includes measures defined for the trajectory of IPS patterns. All indices are calculated per subject session.

The correlation values are reported alongside 95% confidence intervals (CIs), which were derived from 1000 bootstrap resamplings of the subjects.

Deep Temporal Organization of Phase Synchrony Modes

Recent DFC research has revealed a hierarchy in the temporal organization of connectivity states at rest (Vidaurre et al., 2017). In other words, the repertoire of connectivity modes comprises of groups of states (called *metastates*) such that within-metastate transitions are more probable than across. So, once the brain enters a metastate, it is more likely to circulate within that community of states for a while, before switching into another metastate. This speaks to a separation of temporal scales, i.e., a temporal hierarchy. That is to say, the dynamics of the metastates and states unfold on slower and faster timescales, respectively. So far, this deep temporal organization has been endorsed for correlation-based connectivity modes.

Similarly, to examine whether the temporal arrangement of phase synchrony states conforms to a hierarchy, we looked for communities within the transition matrix. By regarding the transition probabilities as adjacency values, the transition matrix can be treated as an adjacency matrix, with an associated graph. Due to the small number of states (i.e., 8), we looked into the coarsest division of the transition matrix (into two partitions). In spectral graph theory, bi-partitioning can be achieved using the Fiedler vector (which is the second smallest eigenvector) of the graph Laplacian (Fiedler, 1973), using the sign of the corresponding vector entries. In other words, positive and negative entries of the Fiedler vector determine how the graph nodes (here states) can be placed into two maximally disjoint partitions. This is an approximate solution to the minimum-cut partitioning problem on graphs (Riolo and Newman, 2014). The result was validated using hierarchical clustering. The presence of distinct well-connected communities in the transition matrix would demonstrate temporal order beyond sample to sample switching. If plausible metastates exist, the remaining task would be investigating their phase synchrony profiles and temporal characteristics – plus potential alterations in SZ.

Trajectory of Phase Coupling Evolution

Cluster analysis is a useful approach for capturing the overall composition and transience of phase couplings patterns. With the usual crisp assignment, all the IPS matrices that belong to a particular cluster collapse onto one dimension (which is the state

label), disregarding individual differences and proximity to the center or boundary of the cluster. Previous DFC work (Miller et al., 2016) has shown that this crude dimensionality reduction overlooks information in the *trajectory* of connectivity and could over/underestimate connectomic variations.

To circumvent this issue, Miller et al. (2016) have reparameterized time-dependent connectivity patterns by discretizing their probabilistic assignments to a few prototypical states (derived from data decomposition). In other words, they have projected time-dependent connectivity patterns on (five) representative connectivity modes and discretized the resultant coefficients; thus, summarizing each connectivity pattern with a five-element integer vector. This method reveals valuable trajectory information and is more indicative of time to time connectivity differences than hard (crisp) state assignment. However, this approach still depends on data decomposition, followed by projection and discretization, and the FC patterns are derived from signal correlations, over successive windows. Instead, we intended to follow the trajectory of IPS patterns, in a high-dimensional space and at every time point, without resorting to decomposition and discretization. Hence, we defined a number of intuitive measures to characterize the evolution of phase couplings. In the following, we describe these measures.

Due to the high-dimensional nature of the IPS matrices (with $50 \times 49/2 = 1225$ unique features), L1 norm[14] is a more effective dissimilarity measure than L2 (Euclidean) distance (Aggarwal et al., 2001). So, we computed *trajectory length* as the sum[15] of L1 distances between successive IPS matrices, per subject session; hence, trajectory length = $\sum_{i=1}^{N-1} L^m_{i(i+1)}$. Plus, the maximum L1 distance (over all IPS patterns in a session) was taken as the *span* of the trajectory: span = $\max_{i,j} L^m_{ij}$; $i, j \in \{1:N\}$. Although informative, this maximum value does not reflect dispersion of the realized IPS patterns over the hyperspace (see schematic trajectory in **Figure 1B**). Hence, by averaging over all pairwise L1 distances (per subject) we computed the *capacity* of the trajectory; i.e., capacity = $< L^m_{ij} >$. Furthermore, the trajectory *efficiency* was defined as the ratio of capacity to trajectory length, for each subject. In this sense, an efficient trajectory traverses a large portion of the state space (i.e., realizes diverse patterns) with relatively small steps; conversely, a very inefficient trajectory is a long one (with large misplanned steps) that eventually remains confined in a small hyperspace.

Another useful feature is the *smoothness* of the trajectory. Smoothness was defined as the average of L1 similarities between consecutive IPS patterns (per subject), where L1 similarity stands for the reciprocal of L1 distance; hence, smoothness = $\frac{1}{N-1} \sum_{i=1}^{N-1} \left(L^m_{i(i+1)} \right)^{-1}$. Using these measures, we looked for potential modulations of the phase coupling evolution in the patient group. For a summary of the measures defined in this section please refer to the lower section of **Table 2**.

Phase Synchrony Features as SZ Predictors

We used the phase synchrony measures, introduced in the previous sections, to predict the SZ label using a linear regression model. Age, gender, and MFD were considered as potential confounds. The model was set up as follows:

$$y_{\text{diagnosis}} = \beta_0 + \beta_{measure} X_{measure} + \beta_{age} X_{age} + \beta_{gender} X_{gender} + \beta_{\text{MFD}} X_{\text{MFD}} + \epsilon \quad (7)$$

where $y_{diagnosis}$ is a binary vector with SZ diagnosis coded as 1 and HC as 0. The linear relationship between each IPS measure ($X_{measure}$) and the diagnosis label ($y_{diagnosis}$) was inspected separately. Moreover, an F test was also conducted on each regression model to assess the significance of the linear relationship. p-values were FDR corrected and significance corresponds to $q < 0.05$. Hence, for each measure, we report the adjusted R^2, $\beta_{measure}$, and q_{FDR} values.

Validation Analyses

Subject-Specificity of Networks

To investigate individual differences in the estimated ICs and their correspondence to the aggregate networks, we followed the validation procedure in Lin et al. (2010). Namely, subject components were variance normalized and a voxel-wise one-sample t test was performed on each of the 50 components, across subjects. Each t-map was subsequently thresholded at an FDR corrected $q < 0.01$. The normalized spatial correlation of subject-specific (FDR-corrected) IC maps with the corresponding IC templates was computed. Average correlation for each subject, over all ICs, was in the (0.6–0.74) range. These values reflect high correspondence of individual networks to the templates [when compared with the (0.43–0.63) range in Lin et al. (2010)], as well as considerable amount of inter-subject variability. Notably, (normalized) spatial correlation to the template was always above 0.45 for any subject-specific IC, averaged around 0.69, and reached as high as 0.88. These results have been demonstrated in **Supplementary Figure S1**.

Genuineness of IPS Modes

To verify that the identified IPS patterns reflect genuine phase arrangements, the same cluster analysis was repeated on surrogate data. Relevant literature recommends phase shuffling (Theiler et al., 1992; Schreiber and Schmitz, 2000; Sun et al., 2012) and random circular shifting (Politis and Romano, 1992; Glerean et al., 2012) for generating surrogates. The former shuffles the phase spectra of the original time series while keeping the amplitude spectra unchanged. The latter applies a random circular shift to each of the original series, to disrupt the existing synchrony between them, while temporal dependencies within each series are preserved. We used both approaches and generated two sets of surrogates. If the clusters in the original dataset reflect meaningful and recurrent phase

[14] $L^m_{ij} = IPS^m(t_j) - IPS^m(t_i)_1$, where m is the subject index; $i, j \in \{1:N\}$ and N is the number of time points per subject session.

[15] In case of unequal session lengths (i.e., different number of time samples per subject), this measure can be normalized to the respective number of time points, to denote average step length.

synchrony modes, they should be distinct from patterns identified from surrogate data.

Motion Effect

To inspect whether head motion is related to the emergence of any state, we investigated the correlations between state occurrences and concomitant FDs. Moreover, we checked for systematic differences between FDs associated with different states. That is, we computed average FD for each state, per subject: $\overline{FD}^s(i)$, where s is the state label and i is the subject index. Then, we conducted a one-way repeated-measures ANOVA to test for difference in the means of these state-dependent average-FDs.

Drowsiness Effect

A prior DFC study (Allen et al., 2014) has reported one FC state associated with drowsiness or light sleep. This speculation was based on the increasing occurrence rate of that state over time. To inspect this possibility, we fitted a line to the occurrence rate (in five-sample windows) of each state over time, for each subject. The slopes of the fitted lines were examined for potential non-zero trend using one-sample t test. If a state is associated with drowsiness, its occurrence rate is likely to show a positive trend over the length of the scan.

Smoothing Kernel Effect

To inspect sensitivity of the findings to the width of the smoothing kernel (in pre-processing), we repeated our analysis for data smoothed using a Gaussian kernel with FWHM = 9 mm. Recent research suggests that kernels spanning 2–3 voxels are optimal for fMRI preprocessing prior to ICA analysis at the subject level (Chen and Calhoun, 2018). From this validation analysis, we found that our key results (i.e., the nature of the states and metastates and their temporal profiles, as well as the IPS trajectory features and their alterations in SZ) are robustly replicated even with this wider kernel. The results have been provided as **Supplementary Material**.

RESULTS

sFNC Indicates Disconnection in SZ

Figure 2A illustrates group-specific sFNC patterns, obtained for each subject and then averaged over each group. The average patterns reveal well-known modular organization within sensory systems and default mode components, as well as anticorrelation between them (Fox et al., 2005; Chang and Glover, 2010; Shirer et al., 2012). Group difference in sFNC (HC–SZ) is demonstrated in **Figure 2B** as $-10\log(q_{FDR}) \times sign(t - statistic)$. Significant differences have been marked in bold squares in the lower triangular part (corresponding to $q_{\mathrm{FDR}} < 0.05$). From 1225 unique sFNC entries, 18% are significantly different between the two groups (13% have higher mean value in the HC group and 5% are higher in SZ). Specifically, stronger correlation among sensory areas of normal subjects as well as pronounced subcortical-sensory anticorrelation (compared to SZ patients) is in accordance with previous findings (Damaraju et al., 2014). Moreover, a two-sample t test on the ensemble of the matrix entries showed that sFNC is globally stronger in normal subjects ($p < 1e{-}4$).

Recurrent IPS States Comprise Distinct and Diverse Patterns

The representative portion of the angular similarity matrix in **Figure 3A** indicates that phase coupling evolution is gradual (hence the heavy diagonal). Moreover, similar patterns emerge over non-adjacent epochs, creating the bright off-diagonal patches. The goal of cluster analysis on IPS patterns was to quantify this recurrence. Using spectral clustering, eight distinct IPS arrangements were identified, which recur over time and across subjects. By associating each IPS matrix with one state label, a sequence of states was formed (see **Figure 3A**, lower panel). The overall prevalence of each state has been imprinted on top of the corresponding matrix in **Figure 4A**.

Figure 4 shows the cluster centroids (i.e., the IPS states) in three formats: matrices, eigenvectors, and connectograms. Clearly, the states differ in their functional architectures. State 1 denotes high level of phase synchrony within and across most networks, while parts of DMN, CB, and CC are decoupled from each other. In state 2, SC and AUD show strong (anti)coupling with respect to the rest of the networks. State 3 reflects notable phase coupling within and across CC, DMN, and CB. In state 4, SC and AUD have strong synchrony with other components while SM, VIS, CC, DMN, and CB are mostly anticoupled or decoupled. State 5 shows an internally incoherent CC, which is also anticoupled to DMN and CB, while DMN is internally integrated. In state 6, (parts of) VIS and CC are anticoupled within and across each other and have mixed (coupling and anticoupling) relationship to DMN and CB. State 7 is noted for strong SC connections, disintegrated CC, and anticoupling of AUD to SM and CC. State 8 is a decoupled state on average, with faintly more synchronous VIS and slightly anticoupled CC. Note that the (unity) diagonal values have been removed from all matrices to improve image contrast.

Figures 4B,C provide alternative useful representations of the states. Bar plots in **Figure 4B** illustrate the first eigenvectors of the states. Hence, the complex coupling and anticoupling relations in each state make up the detached modes in the leading eigenvectors (Bonacich and Lloyd, 2004; Bonacich, 2007; Cabral et al., 2017). The connectograms in panel C show average across-network phase couplings, which serve as concise representations of large-scale IPS modes. Of note is the variety of the patterns identified from empirical data. We will see (in the Results section) that these functional architectures are distinct from accidental phase synchrony arrangements obtained from surrogate data.

Figure 4 also shows that, averaged over all subjects, empirical state 1 is the least common (with 11% prevalence) while state 2 is visited most often (18% prevalence). In the next section, we will inspect group-specific results and show that SZ state proportion is profoundly different from that of the HC.

IPS State Proportion Has Been Altered in SZ

By inspecting state prevalence values in both groups (**Figure 5A**), we note that patients express the first two states almost half as much as normal subjects (6% vs. 11% for state 1, and 12% vs. 22% for state 2). Instead, patients would rather spend more time in state 5 or 7 (15% vs. 10% for state 5, and 12% vs. 9% for state 7).

Importantly, the first two states are more globally connected than others (see **Figure 4B**, the leading eigenvectors), while state 5 is most notable for anticoupling within CC and negative coupling of CC to DMN and CB. State 7 also reflects de/anti-coupling among CC components. Corroborating these results, the persistence plot (in **Figure 5B**) shows that patients tend to remain in state 5 about 1.2 s longer than healthy subjects, on average; conversely, SZ patients cannot hold state 2 as long as the HC (persistence of state 2 = 7.5 s and 10 s, for SZ and HC, respectively). These differences were significant after FDR correction and show that the proportion and average duration of IPS states have been altered in SZ. Notably, this alteration favors less synchronous states that lack a cohesive cognitive control network.

In addition to these state occupation features, the switching pattern between states is another informative aspect of IPS transience, which we will examine in the next section.

IPS State Transitions Show Minimal Change in SZ

In **Figure 5C**, we have shown group-specific transition matrices and their difference (HC–SZ). Although permutation tests on the transition probabilities returned a few significant uncorrected p-values (for transitions from/into states 2 and 5), none of them survived FDR correction, except the probability of dwell in state 2 ($q_{\mathrm{FDR}} < 0.05$. This probability is on average 0.7 for the HC group and 0.6 for SZ patients. This result is in line with our previous finding that state 2 is almost twice as prevalent in healthy subjects (compared to the patients) and persists longer (**Figures 5A,B**).

Although the transition probabilities are not very informative for distinguishing patients from controls (in the present study), they still reveal important information about the spontaneous spatiotemporal reorganization of the brain, in general. We will demonstrate this in the following.

State Transition Probability Correlates With State Similarity

Previous research has suggested that the transition probability between connectivity modes correlates with the similarity of the corresponding connectivity patterns; that is, connectivity changes occur rather gradually over time (Vidaurre et al., 2017). However, the recurrent functional states in Vidaurre et al. (2017) reflect correlations. We tested the same hypothesis about fMRI instantaneous phase coupling variations. That is, we correlated the transition probabilities with the IPS state similarities. Notably, we assessed similarity of the states from both functional and temporal aspects.

To quantify functional similarity of the states, pairwise correlations were computed between the eight IPS states, over all subjects, resulting in an 8×8 symmetric state similarity matrix. Correspondence of this similarity matrix to the average transition matrix was quantified using correlation analysis. The same procedure was repeated using prevalence similarity, persistence similarity, and activity similarity of the states.

The results (in **Figure 5D**) show that, indeed, similar states are more likely to follow each other in time. Phase synchrony resemblance of the states significantly correlates with the transition probabilities [$r = 0.45$, p-value = 0.017, 95% CI = [0.31, 0.57]). The temporal similarity of the states, namely, their prevalence and persistence associations, are also linearly related to their transitional behavior. Specifically, for prevalence similarity vs. transition probability: $r = 0.56$, $p = 0.0018$, 95% CI = [0.47, 0.71]; and for persistence similarity against transition probability: $r = 0.50$, $p = 0.0065$, 95% CI = [0.45, 0.72]. However, activity pattern resemblance is not correlated with transition probability ($r = 0.073$, $p = 0.71$, 95% CI = [−0.25, 0.46]), corroborating the results in Vidaurre et al. (2017). In other words, it is the connectivity (be it amplitude dependencies or phase coupling) structure that evolves with more self-consistence over time – rather than the activity profile.

As noted, the transition matrix holds abundant information about the spatiotemporal organization of the brain at rest. In the next section, we will go through the results of our metastate analysis – again based on the transition pattern – which speaks to higher-order temporal structure in the phase coupling variations.

Metastates Induce Two Distinct Phase Synchrony Modes

To identify metastates (MS), we treated the average transition pattern as an adjacency matrix and bi-partitioned the associated graph using the Fiedler vector (as elaborated in the "Methods" section). **Figure 6A** (right panel) shows the Fiedler vector alongside the hierarchical clustering dendrogram. Both approaches show that states 1–4 and 5–8 constitute two plausible partitions, i.e., metastates. The blocks on the main diagonal of the transition matrix (**Figure 6A**, left panel) show that within-metastate transitions are more probable, than across. For the sake of presentation, each entry of this matrix denotes 0.6th quantile of the corresponding transition probabilities, across subjects.

The average MS profiles are illustrated in **Figure 6B**. MS1 is clearly more coherent, and both SC and AUD networks have central roles in its functional architecture. Conversely, MS2 has lower overall phase synchrony level (**Figure 6C**, bar plot) and shows a notable breakup and anticoupling among CC components. We tested for connectomic differences between these MSs using permutation-based paired t tests. The results in **Figure 6C** show that out of $(\frac{7\times6}{2} + 7 =)$ 28 unique connections within and across the main (seven) networks, 21 couplings are on average stronger in MS1, only 2 connection means are higher in MS2, while 5 connections are not significantly different (after FDR correction). The connectograms in **Figure 6C** show the mean values for the statistically different connections in the two MSs. Moreover, the bar plot (right panel) demonstrates that global phase coupling of MS1 is significantly higher than MS2 (p-value $<$ 1e−4). Having characterized the metastates, the outstanding task is to verify whether the temporal organization of these MSs has been altered in SZ.

The Less-Synchronous Metastate Dominates in SZ

We quantified the temporal profile of the metastates by computing their transition probabilities, as well as their

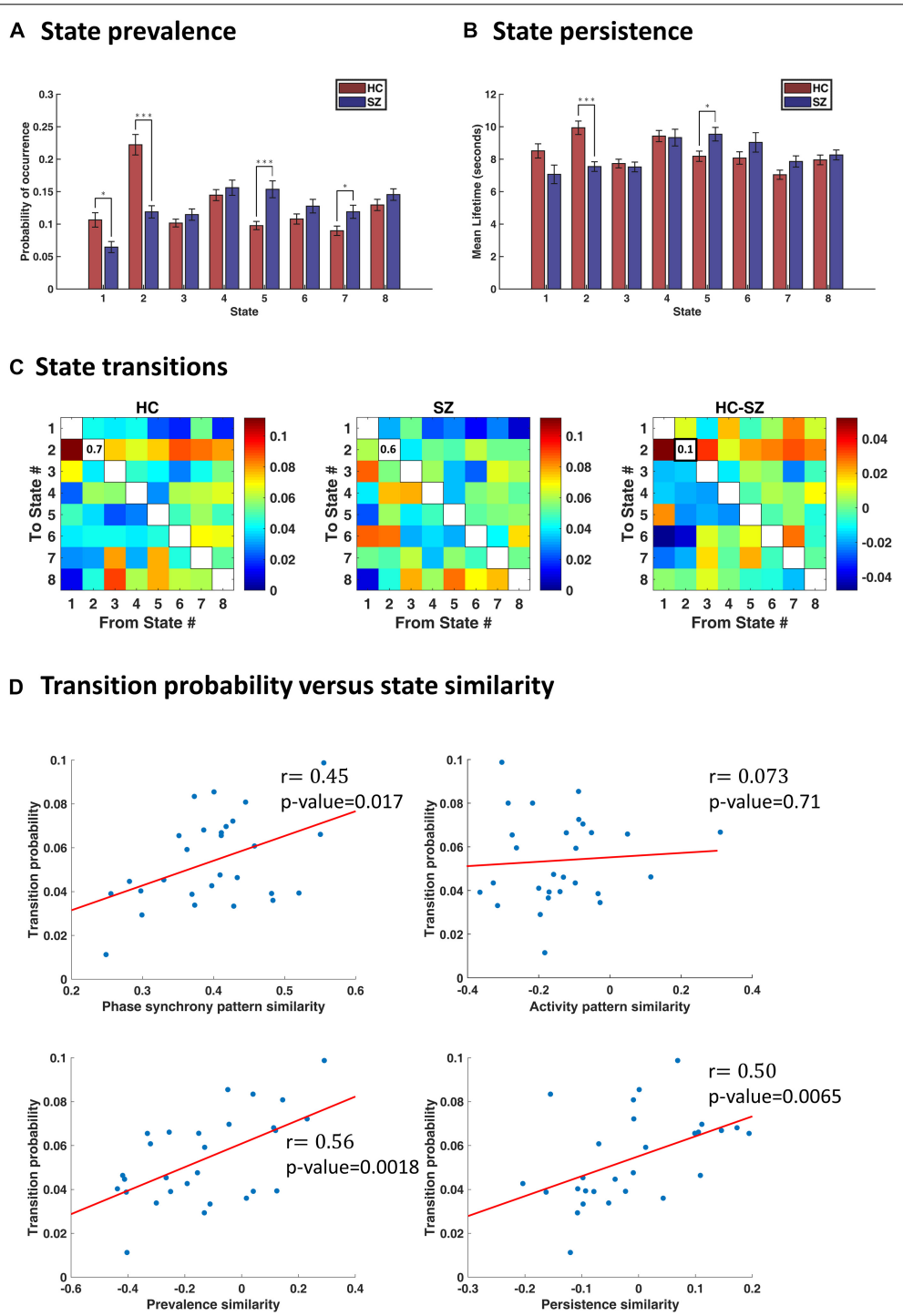


FIGURE 5 | Characterization of transience in IPS states. **(A)** State prevalence (i.e., probability of occurrence) for each state, in the healthy control (HC) and schizophrenic (SZ) group. Error bars denote standard errors of mean (SEM). Group differences were assessed using permutation-based *t* tests. States 1 and 2 are more frequently visited by the HC, whereas states 5 and 7 are more prevalent in the patient group. **(B)** Persistence (i.e., mean lifetime) of each state, per group. The HC remain longer (continuously) in state 2, whereas SZ subjects prefer to linger in state 5. **(C)** Group-average transition patterns and their difference (HC–SZ). Diagonal values have been removed to improve image contrast. Only one entry (i.e., the probability of dwell in state 2) is significantly different after FDR correction. **(D)** Relationship between transition probability and state similarity. State similarity is assessed from four aspects: IPS pattern similarity, activity map similarity, prevalence similarity, and persistence similarity. Activity map resemblance is uncorrelated with transition probability between states; however, similarity in the other three functional and temporal aspects of the states significantly correlates with the probability of switching between them. The correlation coefficient (*r*) is imprinted beside each plot, and red lines denote best fitted lines. In panels **(A,B)**, asterisks show statistically significant differences, after FDR correction (* indicates $p < 0.05$; *** indicates $p < 0.001$).

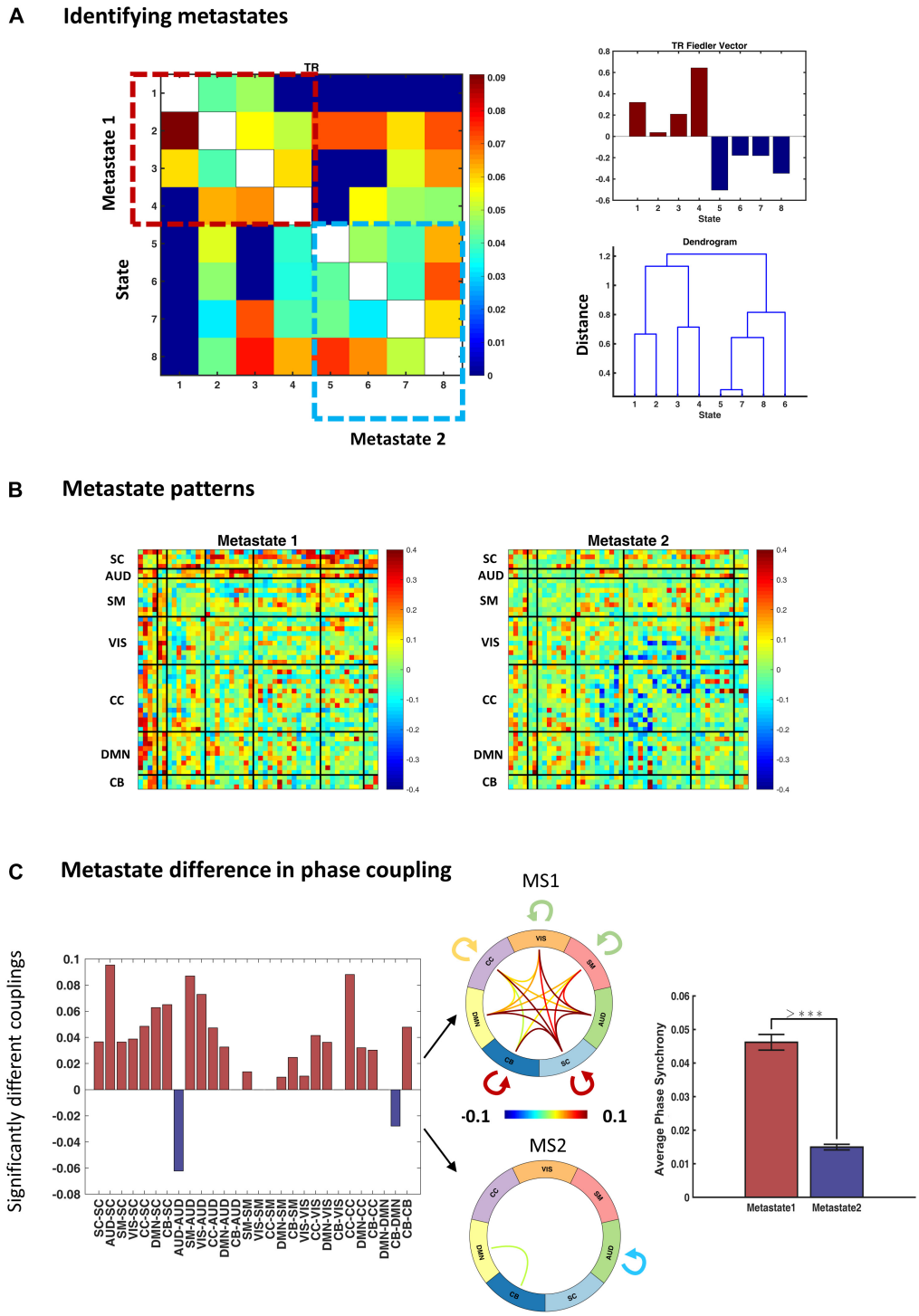


FIGURE 6 | Identifying and characterizing metastates (MS). **(A)** Metastates are groups of states with higher probability of within-group transitions, than across. States 1–4 and 5–8 show this property, as evidenced by diagonal blocks on the group transition matrix. Entries of the transition matrix (left panel) denote 0.6th quantiles, across subjects. The top right panel shows the Fiedler vector (i.e., the second smallest eigenvector of the graph Laplacian associated with the transition matrix) and the lower panel contains the dendrogram of hierarchical clustering (cophenetic correlation coefficient = 0.22). Both methods acknowledge the same metastates. **(B)** Average metastate patterns, across subjects. MS1 is clearly more coherent (than MS2) and MS2 is notable for anticoupling of the cognitive control network components. **(C)** Difference in the phase coupling pattern of the metastates. Left panel: connections that are on average stronger in MS1 (or MS2) were identified using permutation-based paired *t* tests. Positive (pink) bars show (average) phase couplings that are significantly higher in MS1 than MS2; negative (blue) bars denote the inverse. The two connectograms (middle panel) convey the same information graphically. The bar plot (on the right) shows that average phase coupling is significantly higher in MS1, than MS2 (permutation-based paired *t* test: $p < 1e{-}4$). Error bars stand for standard errors of mean, and $>^{***}$ indicates $p < 1e{-}4$.

prevalence and persistence values, in each group. **Figure 7A** shows that the probability of dwell in MS1 is higher in HC compared to SZ (0.85 versus 0.81, p-value < 1e−3). Conversely, the probability of dwell in MS2 is higher in SZ than HC (0.84 versus 0.80). Transition probabilities (from MS1 to MS2 and vice versa) are the complements of dwell values, so they are also significantly different between the two groups. Moreover, the dwell probabilities are significantly different *within* each group; that is, MS1 dwell is higher than MS2 dwell in HC subjects (p < 1e−3), and the opposite is true for SZ patients (p = 0.0162). All p-values were FDR corrected. These results become clearer as we inspect other temporal characteristics of the metastates.

Figures 7B,C show that MS1 is more prevalent and persists longer in the HC, compared to the SZ group (compare the pink bars); the inverse is true for MS2 (inspect the blue bars). All FDR-corrected p-values remained below 0.001 for these tests. But, more importantly, this plot is saying that the balance of MS expression for a typical SZ patient is the inverse of what is expected in a normal subject. That is, while MS1 is dominant in the HC (both in terms of prevalence and persistence), MS2 is the leading phase coupling mode in SZ. These within-group effects were investigated using paired (permutation-based) t tests[16]. Comparing MS ratio between the groups showed that both persistence ratio and prevalence ratio (computed as MS2/MS1) are significantly higher in SZ (p-values < 1e−4; Cohen's d = 0.86 and 0.85).

In short, although immediate IPS transition probabilities seem to be minimally changed in SZ (**Figure 5C**), deep temporal organization of the states has been gravely altered in the patient group. Notably, this altered balance is such that patients spend more time in a poorly connected MS (i.e., MS2), which promotes phase decoupling within and across most functional networks – and anticoupling within the cognitive control network. Overall, this metastate analysis suggests that large-scale functional disconnection in SZ could be mediated by distortions in the deep temporal structure of IPS connectivity modes, at the network level.

This concludes our cluster-based analysis. In the next section, we will see the results of our cluster-free approach to IPS assessment, using the trajectory of instantaneous phase couplings. We will inspect whether phase synchrony evolution has been modulated in SZ.

Phase Coupling Trajectory Is Less Efficient and Less Smooth in the Patient Group

In this section, we report the results of IPS trajectory characterization. We mentioned that collapsing IPS patterns onto one dimension (i.e., state labels) is a simplification that overlooks potentially useful information in the sample-to-sample connectomic changes (Miller et al., 2016). Hence, we used L1 distance to assess dissimilarity between time-indexed IPS matrices, and defined a number of measures to characterize the hyperspace and pathway traversed during each session. These measures are summarized in **Table 2**, and their empirical values are presented in **Table 3**.

The results show that the average *trajectory length* is higher in SZ (compared to HC), but the difference is weakly significant (p = 0.096). However, all the other indices have significantly higher means in the HC group. Namely, the *span*, *capacity*, *efficiency*, and *smoothness* of the trajectory are all higher in normal subjects (FDR corrected p-values = 0.023, 0.002, 0.032, and 0.024, respectively). These results speak to the modulation of ongoing phase coupling in SZ, at the network level. This modulation is such that consecutive patterns are more dissimilar in the patient brain (hence the reduced smoothness), but this jumpy trajectory fails to achieve the more diverse repertoire of IPS patterns that is realized by the typical healthy brain.

Figure 8 shows representative IPS trajectories from a normal subject (22-year-old male) and a SZ patient (33-year-old male), projected on the first two principal components of their IPS profiles. Hence, each circle (or cross) is a time-indexed IPS pattern presented on a low-dimensional manifold and the connecting lines show the progression of IPS from time to time. Even visual inspection confirms that the HC trajectory smoothly explores a larger space (left panel) as opposed to the patient trajectory (right panel) which seems to be confined in the top right corner of the space most of the time. The corresponding state sequences (lower panels) show that the normal subject spends longer periods in state 2 and 4 (which belong to MS1). Conversely, the patient sequence mostly avoids states 1 and 2, but visits states 5–8 more often (which are associated with MS2).

Equipped with IPS measures derived from our state, metastate, and trajectory analyses, we will now look into the explanatory power of these phase indices for predicting SZ in a regression model.

IPS Measures Predict SZ

A linear regression model was used to assess the predictive power of different IPS measures for SZ diagnosis, while treating age, gender, and MFD as confounds. The adjusted $^{R^2}$ values (i.e., explained variances) have been reported in **Figure 9**, alongside the regression coefficients and FDR corrected p-values (of the F tests).

Among the included states, state 2 features (namely, its prevalence, persistence, and dwell) are the better predictors (adjusted $R^2 = 0.10$). Notably, all the MS indices (i.e., MS prevalence, persistence, dwell, and MS2/MS1 ratio) have adjusted R^2 above 0.12, while MS2/MS1 persistence ratio alone can explain 19% of the variance in diagnosis label. As such, metastates seem to be distinctive features. In fact, previous correlation-based DFC research has shown that metastate profile is heritable, subject-specific, and related to behavioral traits (Vidaurre et al., 2017). Among the trajectory measures, capacity seems the most distinctive, with adjusted $R^2 = 0.10$. These results are particularly useful for selecting IPS features for classification studies.

IPS States Are Distinct From Random Phase Synchrony Patterns

Figure 10A shows IPS states derived from clustering surrogate data, alongside empirical states. Surrogate1 was generated by

[16]FDR-corrected p-values for test on prevalence: p < 1e−4 (HC), p = 0.0165 (SZ); test on persistence: p < 1e−4 (HC), p = 0.0135 (SZ).

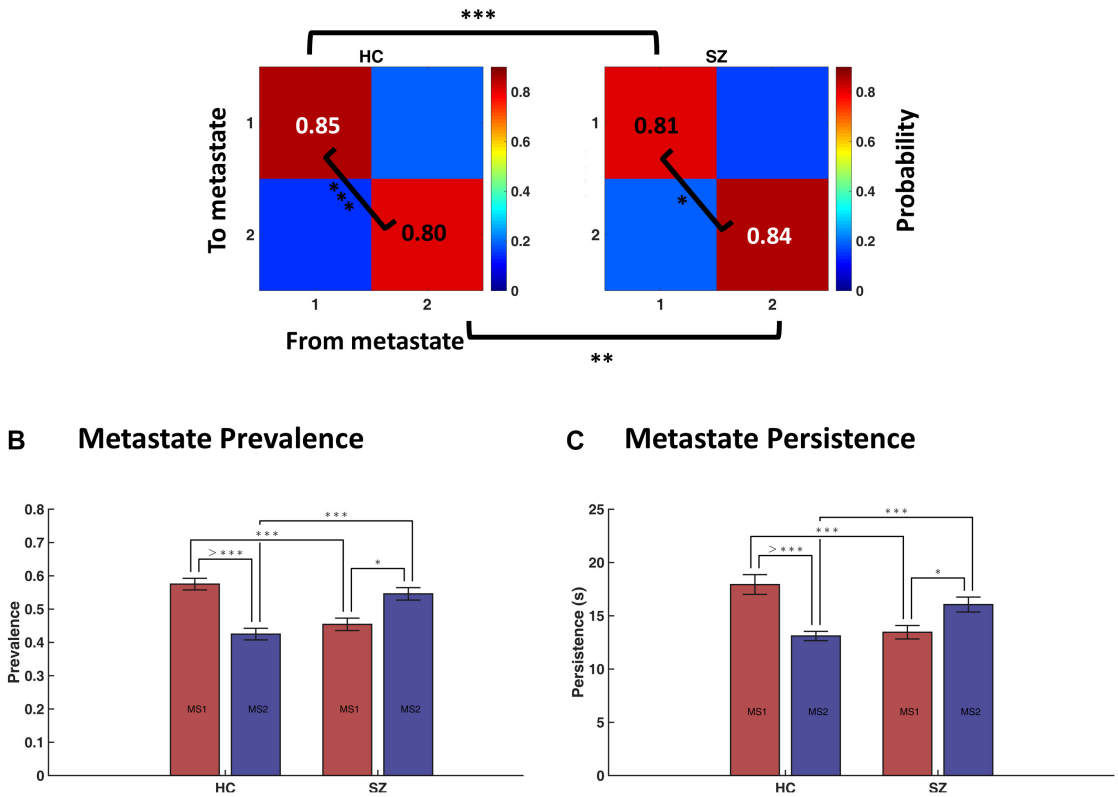


FIGURE 7 | Temporal characteristics of metastates (MS). **(A)** Average transition probability matrices of metastates, for the healthy control (HC, left panel) and schizophrenic (SZ, right panel) subjects. Dwell probabilities (i.e., the diagonal entries) were compared within and between groups. **(B)** Difference in the prevalence (i.e., probability of occurrence) of MS1 and MS2, within and across groups. **(C)** Difference in the persistence (i.e., uninterrupted occupancy) of metastates, within and across groups. The key finding here is that MS1 is dominant in the HC, whereas MS2 prevails in SZ. Statistical tests were permutation-based *t* tests (paired, when appropriate). Error bars denote standard errors of mean (SEM). Asterisks denote significant difference in mean, after FDR correction (* indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$; >*** indicates $p < 0.0001$).

TABLE 3 | Trajectory analysis results.

Measure	HC mean	SZ mean	*p*-value (uncorrected)	*p*-value (FDR corrected)	Effect size (Cohen's *d*)
Trajectory Length	3.93e+3	3.98e+3	0.096	0.096	−0.31
Span	63.25	61.94	0.0092	0.023	+0.50
Capacity	48.90	48.69	0.0004	0.002	+0.67
Efficiency	0.0125	0.0122	0.025	0.032	+0.36
Smoothness	3.35e−3	3.29e−3	0.0145	0.024	+0.46

The measures were defined in ***Table 2****.*

phase shuffling the original series and surrogate2 was produced by inducing random circular shifts in the original data. For surrogate data, the cluster validity index (i.e., Davies–Bouldin index, panel C) continues to decrease by increasing model complexity, as opposed to the empirical plot (in panel B) that endorses eight clusters. Subsequent inspection of surrogate centroids (assuming $k = 8$ clusters) revealed that the resultant states are indistinguishable and lack the strong (anti)coupling structure in the empirical arrangements. Furthermore, we depicted group-specific states to highlight the correspondence of IPS modes in the two groups, although SZ patterns (especially states 1 and 2) are somewhat fainter.

In short, we see that empirical IPS modes are well-defined, diverse, and distinct from random phase synchrony patterns. In other words, IPS states seem to reflect intrinsic order in the phase synchronization of functional networks, which further corroborates the idea that resting-state dynamics may be conceptualized as excursion though a bounded repertoire of metastable functional modes (Deco et al., 2011; Cabral et al., 2017).

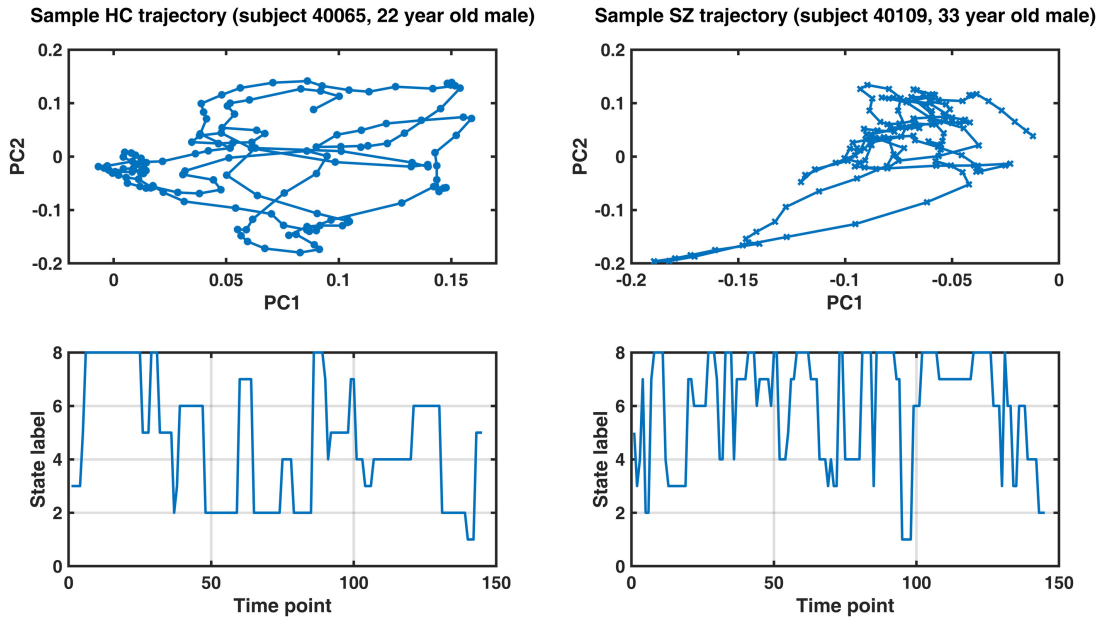


FIGURE 8 | Sample IPS trajectories from a healthy control subject (HC, left panel) and a schizophrenia patient (SZ, right panel). Top panels: IPS trajectories in low-dimensional manifolds (spanned by the first two principal components, PC). Notably, the HC subject covers a larger space and realizes a smoother trajectory (compared to the patient). Lower panels illustrate the corresponding state sequences. The healthy subject spends longer periods in states 2 and 4 (which belong to metastate 1), while the patient's sequence mostly avoids states 1 and 2, but visits states 5–8 more often (which are associated with metastate 2). Please refer to **Table 3** for group differences in IPS trajectory measures.

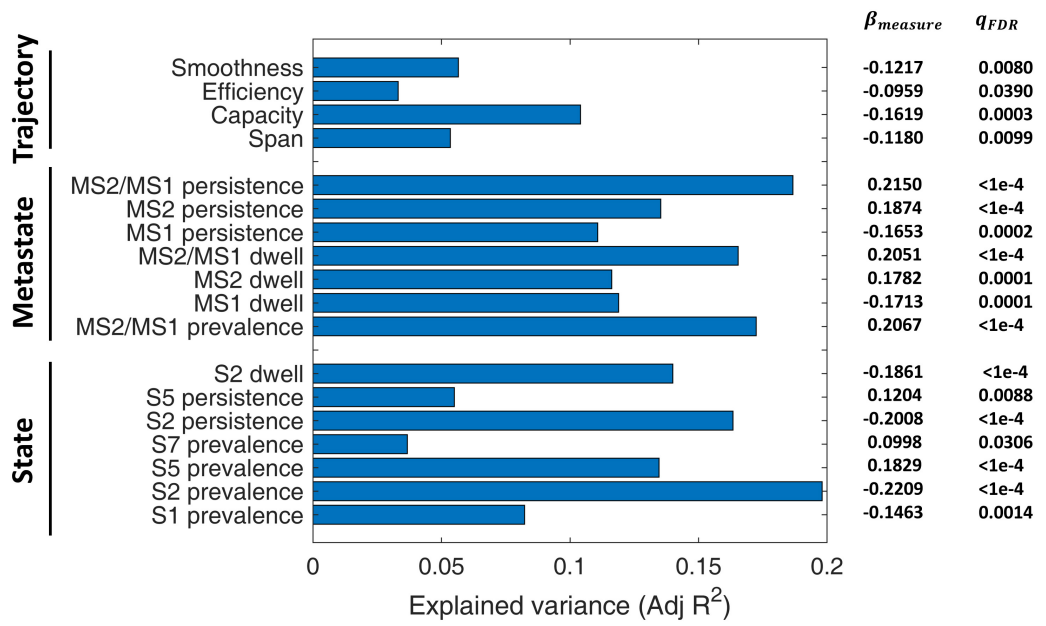


FIGURE 9 | Diagnostic value of IPS measures for schizophrenia (SZ). A linear regression model (Eq. 7) was set up for each measure, while treating age, gender, and mean framewise displacement (MFD) as confounds. Bars denote adjusted R^2 values (i.e., explained variances). Numbers on the right show regression coefficients ($\beta_{measure}$) and FDR-corrected p-values (q_{FDR}) of F tests conducted on regression models. The measures have been grouped into three categories: state-related measures, metastate-based measures, and trajectory measures. For definition of the measures, please refer to **Table 2**. S: State; MS: Metastate.

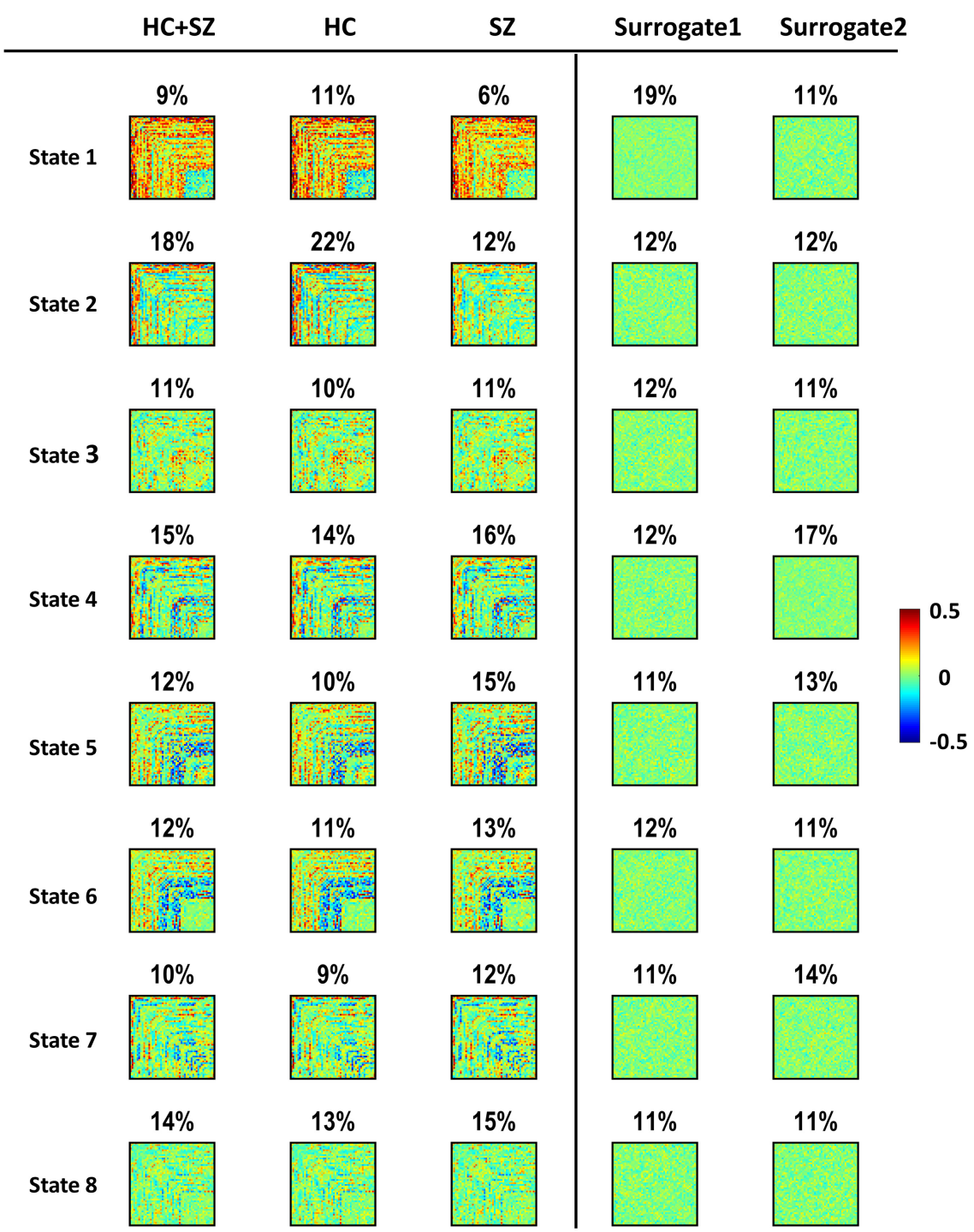


FIGURE 10 | Continued

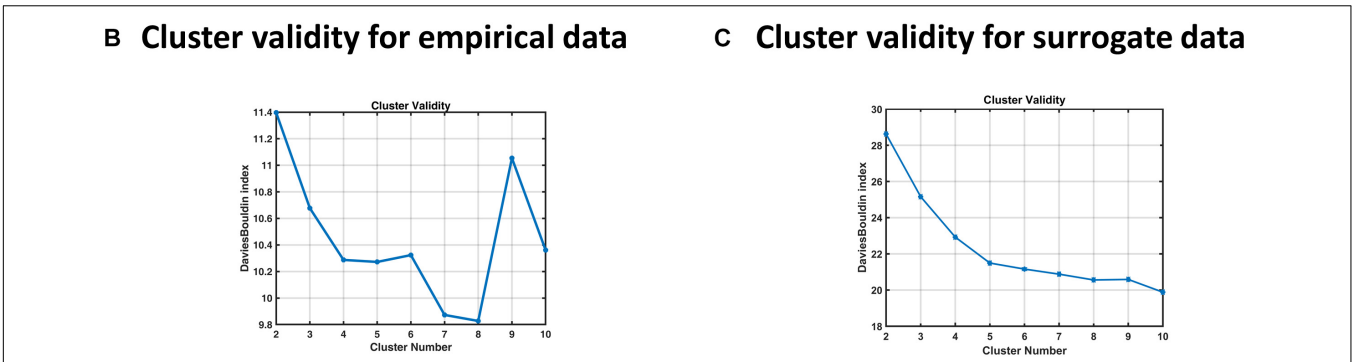


FIGURE 10 | Empirical versus surrogate IPS states. Panel **(A)** plots show that clustering IPS patterns from surrogate data do not result in the same well-defined and distinct states identified in empirical data. The first column (left) shows the empirical states (same as **Figure 4A**). The second and third columns reflect the high correspondence of IPS states in the two (HC and SZ) groups. Percentage values denote the prevalence of each IPS mode. Conversely, the patterns in the last two columns (on the far right) were derived from surrogate data. Surrogate1 was generated by phase shuffling the original series and surrogate2 was produced by inducing random circular shifts in the original time courses. Surrogate data were subjected to the same clustering procedure as empirical data. The cluster validity plot for surrogate data [in panel **(C)**] shows that the Davies–Bouldin validity index continues to decrease by increasing model complexity (i.e., the cluster number). This is in contrast to the trend observed in real data [panel **(B)** here, same as **Figure 3C**], which endorses an optimal model size of 8. The surrogate states depicted in panel **(A)** show the clustering solution for $k = 8$, for the sake of comparison with empirical modes. The resultant surrogate states are flat and lack the strong coupling and anticoupling structure in the original states. Healthy Control: HC; Schizophrenic: SZ.

Phase Synchrony States Are Not Driven by Head Motion

Correlation analysis showed that FD does not meaningfully covary with the occurrence of any state: r = 0.0323, 0.0198, 0.0023, −0.0286, 0.008, −0.0190, 0.0029, and −0.0141, for states 1–8, respectively. Moreover, the effect of state label on mean FD was insignificant [repeated-measures ANOVA: F (7,693) = 1.12, p = 0.35]. Hence, none of the states is associated with head movement, and the evidence does not support difference in the associated motion of the states.

To verify the potential relation of the states to wakefulness, we inspected state occurrence rates over time. It turned out that only the best fitted line to state 4 incidence rate has a significantly non-zero (positive) trend. The uncorrected p-values (of one-sample t tests on the line slopes, across subjects) were 0.66, 0.69, 0.13, 0.034, 0.91, 0.76, 0.12, and 0.35, for the eight states, respectively. The actual state occurrences have been depicted in **Figure 11**, for all subjects (in the upper surface plots). In each lower plot, state occurrence has been accumulated over subjects per time point; hence, the superimposed fitted lines show group trends. Specifically, the group trend for state 4 reflects 3.9% increase in the appearance of this state, per time sample. Notably, state 4 was not different in prevalence or persistence between HC and SZ subjects (see **Figures 5A,B**).

DISCUSSION

Today, resting-state DFC features are being rigorously examined as potential functional biomarkers for diagnosis and prognosis of brain disorders. The ease of task-free data acquisition (especially from patients) makes resting-state studies particularly appealing. Besides, recent evidence suggests that resting-state DFC can affect task performance and may explain inter-subject differences in perception, learning, and other cognitive abilities (Gonzalez-castillo and Bandettini, 2019). Among DFC measures, fMRI IPS has the advantage of evading the window problem (Hindriks et al., 2015; Leonardi and Ville, 2015; Zalesky and Breakspear, 2015) and its functional relevance has been revealed in health, disease, and pharmacological conditions (Glerean et al., 2012; Ponce-Alvarez et al., 2015; Cabral et al., 2017; Alderson et al., 2019; Lord et al., 2019; Zhang et al., 2019). In the present study, we extended IPS characterization to higher functional and temporal scales, and addressed an important analysis issue – that of finding natural order in this type of data. In the following, we will review the key contributions of this work and the significance of the findings.

We showed that the inherent spatiotemporal order in IPS data can be uncovered (using spectral clustering) without compromising the phase coupling details. Importantly, this approach improves interpretability of the identified states. For instance, state 5 (which is more prevalent in SZ patients) shows clear anticoupling between CC and CB networks (see **Figure 4**). In fact, fronto-cerebellar dissociation in SZ has been commonly reported and relates to negative and cognitive symptoms, as well as executive dysfunction in SZ (Ridler et al., 2006; Brady et al., 2019). Moreover, cluster analysis revealed a globally coherent state (i.e., state 1), which has been reported in prior IPS research as well. Specifically, Cabral et al. (2017) showed that this highly synchronous state is visited less often in older adults with poor cognitive scores. Similarly, Lord et al. (2019) showed that the occurrence of a very synchronous state decreases after psilocybin injection. We found that state 1 is significantly less frequent in the SZ group, alongside state 2. As for state 2, it reflects strong interactions of SC and AUD with other networks. Functional (and structural) alterations of the subcortical network in SZ and their relationship with social and cognitive performance have been extensively

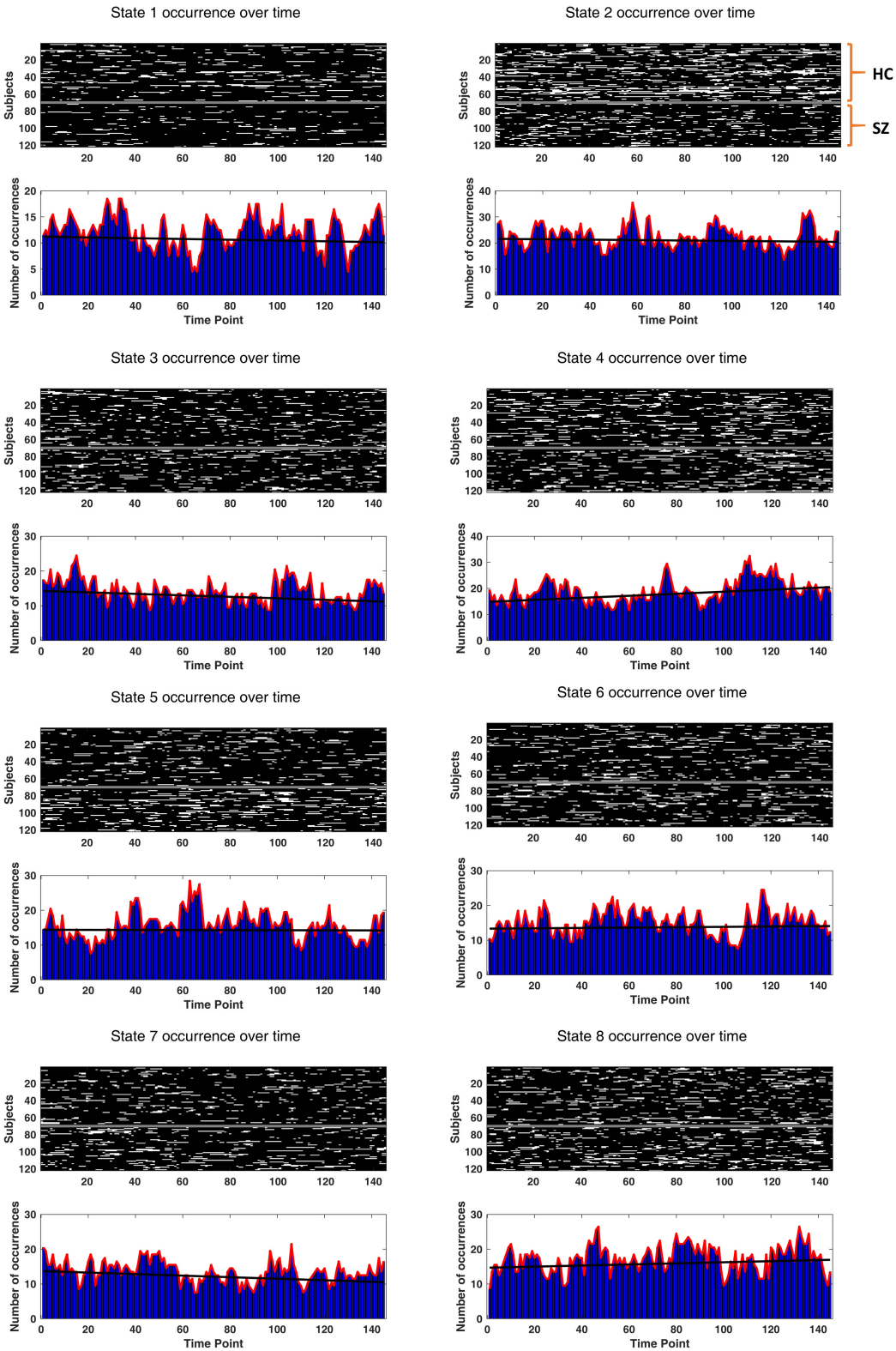


FIGURE 11 | Drowsiness effect. The occurrence of each state (over time) has been depicted for each subject. The two groups (HC and SZ) have been separated by a constant (white) line, in the surface plots. Other white patches indicate the emergence of that particular state over time. In the lower bar plots, state occurrences have been accumulated over subjects, for each time point. The black fitted lines show group trends. Group trend for state 4 reveals a 3.9% increase in the expression of this state, per time sample. Only state 4 has a significantly non-zero (positive) trend (p-value = 0.034).

studied in the literature (Koshiyama et al., 2018). It has even been suggested that reduced fronto-subcortical FC could be a functional biomarker of SZ (Vandevelde et al., 2017). As such, preserving the details of IPS patterns is quite informative and can reveal useful functional features (e.g., for diagnostic purposes).

Another important aspect of our research was focusing on IPS analysis at the network level (rather than regional), with a whole-brain scope, which speaks to higher-order functional integration in the brain. Notably, transient interaction of large-scale networks is a functional attribute that relates to personality traits (Vidaurre et al., 2017, 2019) and shows alterations in prior DFC accounts of SZ (Damaraju et al., 2014; Rashid et al., 2014; Salman et al., 2019). We used constrained ICA and verified (section "Subject-Specificity of Networks") that this approach respects subject specificity of the estimated networks and simultaneously maintains network correspondence at the group level. Moreover, the fine-grained functional parcellation (with 50 ICs) was useful for inspection of sub-network behavior. For instance, we notice that four states reflect anticoupling among CC sub-components (i.e., states 4–7). Remarkably, three of these states belong to metastate 2, which is significantly predominant in SZ (**Figure 7**). This phase coupling alteration in the CC network of the patients might be related to cognitive deficits in SZ and supports previous reports of dysfunctional CC connections (Lesh et al., 2010; Alústiza et al., 2017). Hence, including sub-components of the networks is valuable in that it provides more insight into the functional organization of the states.

A further product of our analysis was uncovering the relationship between transition probability and IPS state similarity. Specifically, we showed that similar states (in terms of functional pattern and temporal profile) are more likely to switch into each other (**Figure 5D**). This is the same phenomenon reported in Vidaurre et al. (2017) – about gradual connectivity progression at rest – despite the fact that our connectivity states reflect phase synchrony modes and their states reflect correlations. Hence, the tendency to explore local neighborhoods in the (functional and temporal) feature space of metastable connectivity structures seems to be a fundamental aspect of spatiotemporal reorganization of the brain, at rest[17].

Subsequently, we looked for deeper (slower) temporal order in the IPS transience. This analysis disclosed higher-level temporal organization of phase synchrony modes and the utility of metastate analysis. This hierarchical temporal arrangement seems to be another fundamental DFC trait that does not depend on the (correlation or phase based) nature of the connectivity states (Vidaurre et al., 2017). Moreover, metastate analysis turned out particularly useful in our clinical application. That is because patients and HCs do not differ much in their immediate transition profiles (**Figure 5C**); however, we noticed that MS balance has been gravely altered in SZ (**Figure 7**). In other words, it is the slower clock (which determines metastate occupation) that seems to be damaged in the SZ disorder. Importantly, metastate proportion turned out to be a strong predictor of SZ, in the regression analysis (**Figure 9**). Additionally, MS characterization showed that MS2 (which prevails in SZ) is significantly less coherent than MS1 (**Figures 6B,C**). Overall, we showed that SZ connectivity disorder is manifest on higher functional (i.e., network) and temporal (i.e., metastate) levels as well.

After state and metastate characterization, we focused on the trajectory of IPS evolution, free from the restrictions of cluster analysis. We defined novel measures (**Table 2**) to quantify different trajectory attributes. The results (in **Table 3**) showed that, despite the relatively large sample-to-sample jumps in SZ, the inefficient IPS trajectory of the patient brain precludes realization of a rich repertoire of FC patterns, compared to HC subjects. Hence, all four measures of span, capacity, efficiency, and smoothness were significantly lower in the patient group. Relevant research (Miller et al., 2016) has characterized the trajectory of (correlation-based) DFC patterns in SZ and showed that patient trajectories cover a smaller portion of the state space and realize fewer distinct patterns. These results seem also in line with the findings in Alderson et al. (2019), which indicate that cognitive performance is directly related to the *variety* of network configurations explored at rest. This makes more sense when we remember that cognitive deficits are among the core symptoms of SZ (Goldman-Rakic, 1994; Perlstein et al., 2001; Silver et al., 2003; Barch, 2005; Forbes et al., 2009).

Despite partial consistency of our findings with those in Miller et al. (2016) – about the reduced span of DFC space in SZ – IPS analysis shows that the connectivity trajectory length is actually longer in SZ, unlike the correlation-based result in Miller et al. (2016). A key difference is that we have followed the connectomic path in a higher dimensional space, and that IPS has inherently higher temporal resolution. This longer trajectory in SZ, alongside shorter span, results in the lower efficiency of FC evolution (**Table 3**), suggesting less structured navigation of the state space in SZ. Moreover, the recurrent inter-network IPS modes identified by our analysis do not resemble the complex-valued states (based on wavelet coherence) in Yaesoubi et al. (2015, 2017) or the correlation-based states in Allen et al. (2014), even though the same network parcellations were adopted in our studies[18]. This speaks to a well-defined identity for the IPS repertoire, beyond correlation-based[19] DFC states. This disparity is due to the different mathematical properties of IPS and sliding-window correlations (Pedersen et al., 2018).

Lastly, we investigated the diagnostic value of IPS measures, using regression analysis. The metastates turned out to be distinctive traits for SZ identification, together with the prevalence of state 2 and the capacity of the trajectory. There is hope that, developing neuroimaging-based biomarkers and (machine learning based) classifiers would facilitate more

[17]Note that if this gradual connectivity evolution were simply due to the continuous nature of fMRI signals or ICA time series, then the activity maps (associated with connectivity states) could have more readily reflected this gradual change – but they do not (**Figure 5D**, upper right panel). Moreover, the tendency of temporally similar connectivity states (in terms of prevalence and persistence profile) to follow each other in time cannot be explained by the continuity of fMRI/ICA time series (**Figure 5D**, lower panels).

[18]Once collapsed along the frequency dimension, Yeasoubi's FC states are almost identical to those in Allen et al. (2014), as depicted in **Figure 5** of Yaesoubi et al. (2015), and both are remarkably different from IPS states. This comparison is licensed because our adopted functional networks are identical to theirs.

[19]Coherence is also correlation in the frequency domain.

objective diagnosis of patient (sub)groups, to furnish more effective treatment selection and prognosis (Drysdale et al., 2017; Fernandes et al., 2017; Bzdok and Meyer-Lindenberg, 2018). Along this way, FC-based methods can provide useful (statistical) insight into the pathological alterations of brain connectivity, at the observation level; however, characterizing the underlying neuronal circuitry and revealing the mechanisms of functional integration would call for model-based (effective connectivity) approaches (Friston, 2011, 2016; Gilson et al., 2019), which are recently being integrated with machine learning techniques as well (Frässle et al., 2020).

To elucidate the long-term vision, we mention a couple of recent achievements. Lately, Brady et al. (2019) have used a data-driven approach to show that connectivity breakdown between the cerebellum (CB) and right dorsolateral prefrontal cortex (DLPFC) directly corresponds to the severity of negative symptoms in SZ, which are known to be resistant to medication. Notably, when these researchers applied transcranial magnetic stimulation (TMS) to the cerebellar midline of SZ patients and restored this specific CB-DLPFC functional connection, the negative symptoms were also relieved – reflecting a causal relationship that is useful for therapeutic purposes. In another prominent research, Deco et al. (2019) have proposed a generative whole-brain model that can predict (*in silico*) how direct electrical stimulation of different brain regions would change the *proportion* of IPS states in the brain (of the sort depicted in our prevalence plot, in **Figure 5A**). As a proof of concept, the authors showed that this method can be used to "awaken" the brain from deep sleep to wakefulness and vice versa, i.e., switching between two conditions that entail different state proportions. Accordingly, 1 day we may be able to modulate disease-specific circuitry and restore state (or metastate) balance in brain disorders – such as SZ – which might relieve the clinical symptoms and eventually improve the quality of life for these patients.

DATA AVAILABILITY STATEMENT

The data analyzed in this study was obtained from the COllaborative Informatics and Neuroimaging Suite Data Exchange tool (COINS; https://coins.trendscenter.org/). Data collection was performed at the Mind Research Network, and funded by a Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the NIH to Dr. Vince Calhoun.

AUTHOR CONTRIBUTIONS

TZ and G-AH-Z conceived the project. TZ analyzed the data and wrote the manuscript. G-AH-Z and FB discussed the results with TZ and reviewed and revised the manuscript. All authors approved this submission.

FUNDING

The Cognitive Sciences and Technologies Council of Iran supported the Ph.D. research of TZ.

ACKNOWLEDGMENTS

We would like to thank Dr. Vince Calhoun and Dr. Elena Allen for sharing group ICA templates (from Allen et al., 2014) online at http://trendscenter.org/data/. We are also grateful to Dr. Joana Cabral for inspiring discussions and comments at the inception of this project. We further appreciate the efforts and responsiveness of the editor while handling this manuscript. Finally, we are grateful to our two reviewers for their constructive comments and suggestions, which improved the presentation of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00214/full#supplementary-material

REFERENCES

Abou-Elseoud, A., Starck, T., Remes, J., Nikkinen, J., Tervonen, O., and Kiviniemi, V. (2010). The effect of model order selection in group PICA. *Hum. Brain Mapp.* 31, 1207–1216. doi: 10.1002/hbm.20929

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science*, Vol. 1973, eds J. Van den Bussche, and V. Vianu, (Berlin: Springer), 420–434. doi: 10.1007/3-540-44503-x_27

Alderson, T. H., Bokde, A. L. W., Kelso, J. A. S., Maguire, L., and Coyle, D. (2019). Metastable neural dynamics underlies cognitive performance across multiple behavioural paradigms. *bioRxiv* [Preprint]

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., and Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* 24, 663–676. doi: 10.1093/cercor/bhs352

Allen, E. A., Erhardt, E. B., Damaraju, E., Gruner, W., Segall, J. M., Silva, R. F., et al. (2011). A baseline for the multivariate comparison of resting-state networks. *Front. Syst. Neurosci.* 5:2. doi: 10.3389/fnsys.2011.00002

Alústiza, I., Radua, J., Pla, M., Martin, R., and Ortuño, F. (2017). Meta-analysis of functional magnetic resonance imaging studies of timing and cognitive control in schizophrenia and bipolar disorder : evidence of a primary time de fi cit. *Schizophr. Res.* 188, 21–32. doi: 10.1016/j.schres.2017.01.039

American Psychogeriatric Association, (2000). *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders, Text Revision*, Vol. 75 Washington, DC: American Psychogeriatric Association, 943. doi: 10.1002/jps.3080051129

Barch, D. M. (2005). The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.* 1, 321–353. doi: 10.1146/annurev.clinpsy.1.102803.143959

Beckmann, C., Mackay, C., Filippini, N., and Smith, S. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage* 47:S148. doi: 10.1016/s1053-8119(09)71511-3

Bedrosian, E. (1962). A product theorem for Hilbert transforms. *Proc. IEEE* 51, 868–869. doi: 10.1109/PROC.1986.13495

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal-Part1: Fundamentals. *Proc. IEEE* 80, 520–538. doi: 10.1109/5.135376

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Soc. Netw.* 29, 555–564. doi: 10.1016/j.socnet.2007.04.002

Bonacich, P., and Lloyd, P. (2004). Calculating status with negative relations. *Soc. Netw.* 26, 331–338. doi: 10.1016/j.socnet.2004.08.007

Brady, R. O., Gonsalvez, I., Lee, I., Öngür, D., Seidman, L. J., Schmahmann, J. D., et al. (2019). Cerebellar-prefrontal network connectivity and negative symptoms in schizophrenia. *Am. J. Psychiatry* 176, 512–520. doi: 10.1176/appi.ajp.2018.18040429

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 3, 223–230. doi: 10.1016/j.bpsc.2017.11.007

Cabral, J., Kringelbach, M. L., and Deco, G. (2014). Exploring the network dynamics underlying brain activity during rest. *Prog. Neurobiol.* 114, 102–131. doi: 10.1016/j.pneurobio.2013.12.005

Cabral, J., Vidaurre, D., Marques, P., Magalhães, R., Silva Moreira, P., Miguel Soares, J., et al. (2017). Cognitive performance in healthy older adults relates to spontaneous switching between states of functional connectivity during rest. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-05425-7

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* 14, 140–151. doi: 10.1002/hbm.1048

Calhoun, V. D., Miller, R., Pearlson, G., Adali, T., Adalı, T., and Adal, T. (2014). The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron* 84, 262–274. doi: 10.1016/j.neuron.2014.10.015

Çetin, M. S., Christensen, F., Abbott, C. C., Stephen, J. M., Mayer, A. R., Cañive, J. M., et al. (2014). Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia. *Neuroimage* 97, 117–126. doi: 10.1016/j.neuroimage.2014.04.009

Chang, C., and Glover, G. H. (2010). Time–frequency dynamics of resting-state brain connectivity measured with fMRI. *Neuroimage* 50, 81–98. doi: 10.1016/j.neuroimage.2009.12.011

Chen, Z., and Calhoun, V. (2018). Effect of spatial smoothing on task fMRI ICA and functional connectivity. *Front. Neurosci.* 12:15. doi: 10.3389/fnins.2018.00015

Chung, F. R. K., and Graham, F. C. (1997). *Spectral Graph Theory*. Providence, RI: American Mathematical Soc.

Collins, D. L., Zijdenbos, A. P., Kollokian, V., Sied, J. G., Kabani, N. J., Holmes, C. J., et al. (1998). Design and construction of a realistic digital brain phantom. *IEEE Trans. Med. Imaging* 17, 463–468. doi: 10.1109/42.712135

Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., et al. (2001). Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. *Am. J. Neuroradiol.* 22, 1326–1333.

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., et al. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage Clin.* 5, 298–308. doi: 10.1016/j.nicl.2014.07.003

Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 224–227. doi: 10.1109/TPAMI.1979.4766909

de Lacy, N., and Calhoun, V. D. (2019). Dynamic connectivity and the effects of maturation in youth with attention deficit hyperactivity disorder. *Netw. Neurosci.* 3, 195–216. doi: 10.1162/netn_a_00063

Deco, G., Cruzat, J., Cabral, J., Tagliazucchi, E., Laufs, H., Logothetis, N. K., et al. (2019). Awakening: predicting external stimulation to force transitions between different brain states. *Proc. Natl. Acad. Sci. U.S.A.* 116, 18088–18097. doi: 10.1073/pnas.1905534116

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2013). Resting brains never rest: computational insights into potential cognitive architectures. *Trends Neurosci.* 36, 268–274. doi: 10.1016/j.tins.2013.03.001

Deco, G., Kringelbach, M. L., Jirsa, V. K., and Ritter, P. (2017). The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core. *Sci. Rep.* 7, 1–14. doi: 10.1038/s41598-017-03073-5

Deshmukh, A. V., Shivhare, V., Gadre, V. M., Patkar, D. P., Shah, S., and Pungavkar, S. (2004). "A phase based method for investigating the functional connectivity in the fMRI data," in *Proceedings of the IEEE INDICON 2004 - 1st India Annual Conference*, Kharagpur, 272–277. doi: 10.1109/INDICO.2004.1497754

Díez-cirarda, M., Strafella, A. P., Kim, J., Peña, J., and Ojeda, N. (2018). Dynamic functional connectivity in Parkinson' s disease patients with mild cognitive impairment and normal cognition. *NeuroImage Clin.* 17, 847–855. doi: 10.1016/j.nicl.2017.12.013

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23, 28–38. doi: 10.1038/nm.4246

Du, Y., and Fan, Y. (2013). Group information guided ICA for fMRI data analysis. *Neuroimage* 69, 157–197. doi: 10.1016/j.neuroimage.2012.11.008

Du, Y., Fryer, S. L., Fu, Z., Lin, D., Sui, J., Chen, J., et al. (2018). Dynamic functional connectivity impairments in early schizophrenia and clinical high-risk for psychosis. *Neuroimage* 180(Pt B), 632–645. doi: 10.1016/j.neuroimage.2017.10.022

Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., et al. (2019). NeuroMark : an adaptive independent component analysis framework for estimating reproducible and comparable fMRI biomarkers among brain disorders. *medRxiv* [Preprint]. doi: 10.1101/19008631

Erhardt, E. B., Rachakonda, S., Bedrick, E. J., Allen, E. A., Adali, T., and Calhoun, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* 32, 2075–2095. doi: 10.1002/hbm.21170

Espinoza, F. A., Anderson, N. E., Vergara, V. M., Harenski, C. L., Decety, J., Rachakonda, S., et al. (2019). Resting-state fMRI dynamic functional network connectivity and associations with psychopathy traits. *NeuroImage Clin.* 24:101970. doi: 10.1016/j.nicl.2019.101970

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Med.* 15:80. doi: 10.1186/s12916-017-0849-x

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslov. Math. J.* 23, 298–305.

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders*. research version, patient edition (SCID-I/P). New York, NY: Biometrics Research, New York State Psychiatric Institute.

Forbes, N. F., Carrick, L. A., McIntosh, A. M., and Lawrie, S. M. (2009). Working memory in schizophrenia: a meta-analysis. *Psychol. Med.* 36, 889–905. doi: 10.1017/S0033291708004558

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *Neuroimage* 62, 2296–2314. doi: 10.1016/j.neuroimage.2011.12.090

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. Van, and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Frässle, S., Marquand, A. F., Schmaal, L., Dinga, R., Veltman, D. J., van der Wee, N. J. A., et al. (2020). Predicting individual clinical trajectories of depression with generative embedding. *NeuroImage Clin.* 102213. (in press) doi: 10.1016/j.nicl.2020.102213

Friston, K. J. (1998). The disconnection hypothesis. *Schizophr. Res.* 30, 115–125. doi: 10.1016/S0920-9964(97)00140-0

Friston, K. J. (2011). Functional and effective connectivity: a Review. *Brain Connect.* 1, 13–36. doi: 10.1089/brain.2011.0008

Friston, K. J. (2016). Computational nosology and precision psychiatry. *Comput. Psychiatry* 1, 2–23. doi: 10.7551/mitpress/9780262035422.003.0011

Fu, Z., Caprihan, A., Chen, J., Du, Y., Adair, J. C., Sui, J., et al. (2019). Altered static and dynamic functional network connectivity in Alzheimer's disease and subcortical ischemic vascular disease: shared and specific brain connectivity abnormalities. *Hum. Brain Mapp.* 40, 3203–3221. doi: 10.1002/hbm.24591

Gallier, J. (2016). Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *arXiv* [Preprint]; arXiv1601.04692

Ghosh, A., Rho, Y., McIntosh, A. R., Kotter, R., and Jirsa, V. K. (2008). Noise during rest enables the exploration of the brain's dynamic repertoire. *PLoS Comput. Biol.* 4:e1000196. doi: 10.1371/journal.pcbi.1000196

Gilbert, N., Bernier, R. A., Calhoun, V. D., Brenner, E., Grossner, E., Rajtmajer, S. M., et al. (2018). Diminished neural network dynamics after moderate and severe traumatic brain injury. *PLoS One* 13:e0197419. doi: 10.1371/journal.pone.0197419

Gilson, M., Zamora-López, G., Pallarés, V., Adhikari, M. H., Senden, M., Tauste Campo, A., et al. (2019). Model-based whole-brain effective connectivity to study distributed cognition in health and disease. *Netw. Neurosci.* 1–62. (in press) doi: 10.1162/netn_a_00117

Glerean, E., Salmi, J., Lahnakoski, J. M., Jääskeläinen, I. P., and Sams, M. (2012). Functional magnetic resonance imaging phase synchronization as a measure of dynamic functional connectivity. *Brain Connect.* 2, 91–101. doi: 10.1089/brain.2011.0068

Goldman-Rakic, P. S. (1994). Working memory dysfunction in schizophrenia. *J. Neuropsychiatry Clin. Neurosci.* 6, 348–357. doi: 10.1176/jnp.6.4.348

Gonzalez-castillo, J., and Bandettini, P. A. (2019). Task-based dynamic functional connectivity: recent findings and open questions. *Neuroimage* 180(Pt B), 526–533. doi: 10.1016/j.neuroimage.2017.08.006

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage* 22, 1214–1222. doi: 10.1016/j.neuroimage.2004.03.027

Hindriks, R., Adhikari, M. H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N. K., et al. (2015). Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage* 127, 242–256. doi: 10.1016/j.neuroimage.2015.11.055

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., et al. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* 80, 360–378. doi: 10.1016/j.neuroimage.2013.05.079

Keilholz, S. D. (2014). The neural basis of time-varying resting-state functional connectivity. *Brain Connect.* 4, 769–779. doi: 10.1089/brain.2014.0250

Keilholz, S. D., Magnuson, M. E., Pan, W. J., Willis, M., and Thompson, G. J. (2013). Dynamic properties of functional connectivity in the rodent. *Brain Connect.* 3, 31–40. doi: 10.1089/brain.2012.0115

Kitzbichler, M. G., Smith, M. L., Christensen, S. R., and Bullmore, E. (2009). Broadband criticality of human brain network synchronization. *PLoS Comput. Biol.* 5:e1000314. doi: 10.1371/journal.pcbi.1000314

Kiviniemi, V., Starck, T., Remes, J., Long, X., Nikkinen, J., Haapea, M., et al. (2009). Functional segmentation of the brain cortex using high model order group PICA. *Hum. Brain Mapp.* 30, 3865–3886. doi: 10.1002/hbm.20813

Koshiyama, D., Fukunaga, M., Okada, N., and Yamashita, F. (2018). Role of subcortical structures on cognitive and social function in schizophrenia. *Sci. Rep.* 8:1183. doi: 10.1038/s41598-017-18950-2

Lachaux, J. P., Rodriguez, E., Le Van Quyen, M., Lutz, A., Martinerie, J., and Varela, F. J. (2000). Studying single-trials of phase synchronous activity in the brain. *Int. J. Bifurcat. Chaos.* 10, 2429–2439. doi: 10.1142/S0218127400001560

Lachaux, J. P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* 8, 194–208. doi: 10.1002/(SICI)1097-019319998:4<194::AID-HBM4<3.0.CO;2-C

Laird, A. R., Rogers, B. P., Carew, J. D., Arfanakis, K., Moritz, C. H., and Meyerand, M. E. (2002). Characterizing instantaneous phase relationships in whole-brain fMRI activation data. *Hum. Brain Mapp.* 16, 71–80. doi: 10.1002/hbm.10027

Leonardi, N., and Ville, D. Van De (2015). On spurious and real fluctuations of dynamic functional connectivity during rest. *Neuroimage* 104, 430–436. doi: 10.1016/j.neuroimage.2014.09.007

Lesh, T. A., Niendam, T. A., Minzenberg, M. J., and Carter, C. S. (2010). Cognitive control deficits in schizophrenia : mechanisms and meaning. *Neuropsychopharmacology* 36, 316–338. doi: 10.1038/npp.2010.156

Liang, M., Zhou, Y., Jiang, T., Liu, Z., Tian, L., Liu, H., et al. (2006). Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport* 17, 209–213. doi: 10.1097/01.wnr.0000198434.06518.b8

Lin, Q., Liu, J., Zheng, Y., Liang, H., and Calhoun, V. D. (2010). Semiblind spatial ICA of fMRI using spatial constraints. *Hum Brain Mapp.* 31, 1076–1088. doi: 10.1002/hbm.20919

Lord, L., Expert, P., Atasoy, S., Roseman, L., and Rapuano, K. (2019). Dynamical exploration of the repertoire of brain networks at rest is modulated by psilocybin neuroimage dynamical exploration of the repertoire of brain networks at rest is modulated by psilocybin. *Neuroimage* 199, 127–142. doi: 10.1016/j.neuroimage.2019.05.060

Luxburg, U. Von (2007). A tutorial on spectrum clustering. *Stat. Comput.* 17, 1–32. doi: 10.1055/s-0034-1376191

Ma, S., Calhoun, V. D., Phlypo, R., and Adali, T. (2014). Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis. *Neuroimage* 90, 196–206. doi: 10.1016/j.neuroimage.2013.12.063

Mcclellan, J. H., Parks, T. W., and Rabiner, L. R. (1973). A computer program for designing optimum FIR linear phase digital filters. *IEEE Trans. Audio Electroacoust.* 21, 506–526. doi: 10.1109/TAU.1973.1162525

Miller, R. L., Yaesoubi, M., Turner, J. A., Mathalon, D., Preda, A., Pearlson, G., et al. (2016). Higher dimensional meta-state analysis reveals reduced resting fMRI connectivity dynamism in schizophrenia patients. *PLoS One* 11:e0149849. doi: 10.1371/journal.pone.0149849

Murphy, K., and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* 154, 169–173. doi: 10.1016/j.neuroimage.2016.11.052

Murray, C. J. L., Barber, R. M., Foreman, K. J., Ozgoren, A. A., Abd-Allah, F., Abera, S. F., et al. (2016). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990-2013: quantifying the epidemiological transition. *Lancet* 386, 2145–2191. doi: 10.1016/S0140-6736(15)61340-X

Nejad, A. B., Ebdrup, B. H., Glenthøj, B. Y., and Siebner, H. R. (2012). Brain connectivity studies in schizophrenia: unravelling the effects of antipsychotics. *Curr. Neuropharmacol.* 10, 219–230. doi: 10.2174/157015912803217305

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). "On spectral clustering: analysis and an algorithm," *NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 849–856.

Pedersen, M., Omidvarnia, A., Zalesky, A., and Jackson, G. D. (2018). On the relationship between instantaneous phase synchrony and correlation-based sliding windows for time-resolved fMRI connectivity analysis. *Neuroimage* 181, 85–94. doi: 10.1016/j.neuroimage.2018.06.020

Perlstein, W. M., Carter, C. S., Noll, D. C., and Cohen, J. D. (2001). Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia. *Am. J. Psychiatry*. 158, 1105–1113. doi: 10.1176/appi.ajp.158.7.1105

Politis, D. N., and Romano, J. P. (1992). *A Circular Block-Resampling Procedure for Stationary Data. ExploreLimits Bootstrap* 2635270. New York, NY: John Wiley & Sons.

Ponce-Alvarez, A., Deco, G., Hagmann, P., Romani, G. L., Mantini, D., and Corbetta, M. (2015). Resting-state temporal synchronization networks emerge from connectivity topology and heterogeneity. *PLoS Comput. Biol.* 11:e1004100. doi: 10.1371/journal.pcbi.1004100

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*. 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018

Preti, M. G., Bolton, T. A., and Van De Ville, D. (2017). The dynamic functional connectome: state-of-the-art and perspectives. *Neuroimage* 160, 41–54. doi: 10.1016/j.neuroimage.2016.12.061

Rabinovich, M. I., and Varona, P. (2011). Robust transient dynamics and brain functions. *Front. Comput. Neurosci.* 5:24. doi: 10.3389/fncom.2011.00024

Rashid, B., Blanken, L. M. E., Muetzel, R. L., Miller, R., Damaraju, E., Arbabshirani, M. R., et al. (2018). Connectivity dynamics in typical development and its relationship to autistic traits and autism spectrum disorder. *Hum. Brain Mapp.* 39, 3127–3142. doi: 10.1002/hbm.24064

Rashid, B., Damaraju, E., Pearlson, G. D., and Calhoun, V. D. (2014). Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects. *Front. Hum. Neurosci.* 8:897. doi: 10.3389/fnhum.2014.00897

Ridler, K., Veijola, J. M., Tanskanen, P., Miettunen, J., Chitnis, X., Suckling, J., et al. (2006). Fronto-cerebellar systems are associated with infant motor and adult executive functions in healthy adults but not in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15651–15656. doi: 10.1073/pnas.0602639103

Riolo, M. A., and Newman, M. E. J. (2014). First-principles multiway spectral partitioning of graphs. *J. Complex Netw.* 2, 121–140. doi: 10.1093/comnet/cnt021

Salman, M. S., Du, Y., Lin, D., Fu, Z., Fedorov, A., Damaraju, E., et al. (2019). Group ICA for identifying biomarkers in schizophrenia : 'Adaptive' networks via spatially constrained ICA show more sensitivity to group di ff erences than spatio-temporal regression. *NeuroImage Clin.* 22:101747. doi: 10.1016/j.nicl.2019.101747

Schreiber, T., and Schmitz, A. (2000). Surrogate time series. *Phys. D Nonlinear Phenom.* 142, 346–382. doi: 10.1016/S0167-2789(00)00043-9

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation part of the electrical and computer engineering commons recommended citation normalized cuts and image segmentation normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi: 10.1109/34.868688

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb. Cortex* 22, 158–165. doi: 10.1093/cercor/bhr099

Silver, H., Feldman, P., Bilker, W., and Gur, R. C. (2003). Working memory deficit as a core neuropsychological dysfunction in schizophrenia. *Am. J. Psychiatry* 160, 1809–1816. doi: 10.1176/appi.ajp.160.10.1809

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13040–13045. doi: 10.1073/pnas.0905267106

Sun, J., Hong, X., Tong, S., and Member, S. (2012). phase synchronization analysis of EEG signals : an evaluation based on surrogate tests. *IEEE Trans. Biomed. Eng.* 59, 2254–2263. doi: 10.1109/TBME.2012.2199490

Tagliazucchi, E., Laufs, H., and Garrett, D. D. (2015). Multimodal imaging of dynamic functional connectivity. *Front. Neurol.* 6:10. doi: 10.3389/fneur.2015.00010

Tagliazucchi, E., von Wegner, F., Morzelewski, A., Brodbeck, V., and Laufs, H. (2012). Dynamic BOLD functional connectivity in humans and its electrophysiological correlates. *Front. Hum. Neurosci.* 6:339. doi: 10.3389/fnhum.2012.00339

Tass, P., Rosenblum, M. G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., et al. (1998). Detection of n:m phase locking from noisy data: application to magnetoencephalography. *Phys. Rev. Lett.* 81, 3291–3294. doi: 10.1103/PhysRevLett.81.3291

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Doyne Farmer, J. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Phys. D Nonlinear Phenom.* 58, 77–94. doi: 10.1016/0167-2789(92)90102-S

Thompson, G. J., Pan, W. J., Magnuson, M. E., Jaeger, D., and Keilholz, S. D. (2014). Quasi-periodic patterns (QPP): large-scale dynamics in resting state fMRI that correlate with local infraslow electrical activity. *Neuroimage* 84, 1018–1031. doi: 10.1016/j.neuroimage.2013.09.029

Van Den Heuvel, M. P., and Fornito, A. (2014). Brain networks in schizophrenia. *Neuropsychol. Rev.* 24, 32–48. doi: 10.1007/s11065-014-9248-7

Vandevelde, A., Leroux, E., Delcroix, N., and Dollfus, S. (2017). Fronto-subcortical functional connectivity in patients with schizophrenia and bipolar disorder during a verbal fluency task. *World J. Biol. Psychiatry* 19(Suppl. 3), S124–S132. doi: 10.1080/15622975.2017.1349339

Vidaurre, A. D., Llera, A., Smith, S. M., and Woolrich, M. W. (2019). Behavioural relevance of spontaneous, transient brain network interactions in fMRI. *bioRxiv* [Preprint]

Vidaurre, D., Smith, S. M., and Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12827–12832. doi: 10.1073/pnas.1705120114

Yaesoubi, M., Allen, E. A., Miller, R. L., and Calhoun, V. D. (2015). Dynamic coherence analysis of resting fMRI data to jointly capture state-based phase, frequency, and time-domain information. *Neuroimage* 120, 133–142. doi: 10.1016/j.neuroimage.2015.07.002

Yaesoubi, M., Miller, R. L., Bustillo, J., Lim, K. O., Vaidya, J., and Calhoun, V. D. (2017). A joint time-frequency analysis of resting-state functional connectivity reveals novel patterns of connectivity shared between or unique to schizophrenia patients and healthy controls. *NeuroImage Clin.* 15, 761–768. doi: 10.1016/j.nicl.2017.06.023

Zalesky, A., and Breakspear, M. (2015). Towards a statistical test for functional connectivity dynamics. *Neuroimage* 114, 466–470. doi: 10.1016/j.neuroimage.2015.03.047

Zhang, R., Kranz, G. S., and Lee, T. M. C. (2019). Functional connectome from phase synchrony at resting state is a neural fingerprint. *Brain Connect.* 9, 519–528. doi: 10.1089/brain.2018.0657

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zarghami, Hossein-Zadeh and Bahrami. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 01 April 2020
doi: 10.3389/fnins.2020.00192

A Deep Learning Approach for Mild Depression Recognition Based on Functional Connectivity Using Electroencephalography

Xiaowei Li[1], Rong La[1], Ying Wang[1], Bin Hu[1,2,3] and Xuemin Zhang[4,5,6]*

[1] Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China, [2] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [3] Beijing Institute for Brain Disorders, Capital Medical University, Beijing, China, [4] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University, Beijing, China, [5] State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, [6] Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing, China

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Rui Li,
Institute of Psychology (CAS), China
Loris Nanni,
University of Padua, Italy

***Correspondence:**
Bin Hu
bh@lzu.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *07 June 2019*
Accepted: *24 February 2020*
Published: *01 April 2020*

Citation:
Li X, La R, Wang Y, Hu B and Zhang X (2020) A Deep Learning Approach for Mild Depression Recognition Based on Functional Connectivity Using Electroencephalography. Front. Neurosci. 14:192. doi: 10.3389/fnins.2020.00192

Early detection remains a significant challenge for the treatment of depression. In our work, we proposed a novel approach to mild depression recognition using electroencephalography (EEG). First, we explored abnormal organization in the functional connectivity network of mild depression using graph theory. Second, we proposed a novel classification model for recognizing mild depression. Considering the powerful ability of CNN to process two-dimensional data, we applied CNN separately to the two-dimensional data form of the functional connectivity matrices from five EEG bands (delta, theta, alpha, beta, and gamma). In addition, inspired by recent breakthroughs in the ability of deep recurrent CNNs to classify mental load, we merged the functional connectivity matrices from the three EEG bands that performed the best into a three-channel image to classify mild depression-related and normal EEG signals using the CNN. The results of the graph theory analysis showed that the brain functional network of the mild depression group had a larger characteristic path length and a lower clustering coefficient than the healthy control group, showing deviation from the small-world network. The proposed classification model obtained a classification accuracy of 80.74% for recognizing mild depression. The current study suggests that the combination of a CNN and functional connectivity matrix may provide a promising objective approach for diagnosing mild depression. Deep learning approaches such as this might have the potential to inform clinical practice and aid in research on psychiatric disorders.

Keywords: EEG, functional connectivity, convolutional neural network, mild depression, classification

INTRODUCTION

Depression is a global public health problem, which has a relatively high lifetime prevalence, ranging from 2 to 15%, and is associated with significant morbidity (Ustun and Chatterji, 2001; Üstün et al., 2004). According to the latest data from the World Health Organization (2017)[1], more than 300 million people are now living with depression[1]. Presently, the most widely used methods for depression diagnosis are based on Beck's Depression Inventory (BDI), the patient's self-report, the doctor's clinical experience, or some combination thereof. However, the accuracy of this diagnosis is often influenced by the doctor's proficiency and patient's cooperation, both of which are highly subjective. Critically, a subset of depression–mild depression–receives far less attention than does depression, despite being more common than depression and often increasing in severity over time (Volz and Laux, 2000). This lack of attention leads to missed early detection and treatment and increases the mortality risk and likelihood that mild depression will evolve into major depression (Fogel et al., 2006; Lyness et al., 2006). Additionally, mild depression is not only a mental illness but also often a social problem (Li et al., 2016b). Therefore, studies of methods that might improve the early detection and treatment of mild depression are both necessary and meaningful.

With the emergence of more and more studies of depression using functional brain imaging, it is becoming clear that this psychopathology might be relevant to the distributed properties of large-scale cortical systems across many functionally connected cortical regions (Spence, 1995; Dalgleish, 2004; Wang et al., 2012). Recent studies have provided further evidence that depressed individuals tend to have altered brain functional connectivity, such as significantly decreased functional connectivity between right posterior insula and a series of sensory cortices (Hu et al., 2019), decreased brain activity in the dorsolateral prefrontal cortex, superior temporal gyrus, posterior precuneus, and posterior cingulate (Zhong et al., 2016), and increased subgenual cingulate–thalamic connectivity (Ajilore et al., 2015). To examine the functional connectivity of the brain, various metrics have been used, including coherence, correlation, phase locking value, and phase lag index, among others.

Electroencephalography (EEG) coherence is an effective measure of functional cortical connectivity and is used to calculate linearly dependent interactions between the frequencies of EEG signals derived from two electrodes or brain regions (Andrew and Pfurtscheller, 1996; Pfurtscheller and Andrew, 1999). This measure results in a symmetrical, two-dimensional (2D) matrix. High coherence between two EEG signals reflects synchronized neuronal oscillations (suggesting functional integration between neural populations), while low coherence indicates independently active populations (suggesting functional segregation) (Murias et al., 2007). Using EEG coherence to study the brain activity of patients with depression (Li et al., 2017), as well as those with Alzheimer's disease (AD) (Wada et al., 1998) and Parkinson's disease (Teramoto et al., 2016), has been quite successful. For example, Murias et al. (2007) observed an elevated EEG theta coherence in the frontal and temporal regions of the left hemisphere in individuals with autism spectrum disorder. In a study that used EEG coherence to assess resting state functional connectivity, Teramoto et al. (2016) found a decreased resting state functional connectivity between the frontal and parietal cortices, especially in the left hemisphere, of patients with Parkinson's disease. Yingjie et al. (Li et al., 2015) further used EEG coherence to investigate differences in brain functional networks between patients with depression and healthy controls, while they were processing emotional stimuli. The authors found that global EEG coherence in the gamma band was significantly higher in patients with depression than it was in healthy controls.

Correlation is an alternative method of calculating functional connectivity matrices and is often used to estimate the level of linear dependence between two electrode channels. In a study by Zhang et al. (2018) that measured functional connectivity, the authors calculated the Pearson correlation coefficients of the power spectral densities for the delta, theta, alpha, beta, and gamma bands. Following this, they constructed a square correlation matrix for each participant and each frequency band. Finally, the functional brain networks of healthy controls and patients with major depressive disorder (MDD) were constructed using the correlation matrix. The results revealed that compared to healthy controls, the patients with MDD showed significant randomization of the global network metrics.

A statistical method known as "phase synchronization" can also be used to measure coordinated activation across different brain regions. In biological signals, such as EEG time series, synchronization is measured by calculating the phase locking value (PLV). Mormann et al. (2000) reported on the application of phase synchronization methods to biological time series data representative of brain electrical activity in patients with epilepsy. They observed characteristic spatial and temporal shifts in synchronization that appeared to be strongly related to pathological activity. In particular, the authors reported distinct differences in the degree of synchronization between recordings from seizure-free intervals and those from the intervals just before an impending seizure.

The phase lag index (PLI) is another measure of asymmetry in the distribution of phase differences between two signals. The PLI is an alternative measure of statistical interdependencies between time series that reflects the strength of their coupling by detecting consistent, non-zero phase lag between the two times series. Stam et al. (2007) found that PLI performed well for detecting relative increases in synchronization between the pre-seizure and seizure epochs. In addition, upon analyzing the EEG signals, the authors found that the average PLI in the beta band was significantly lower in patients with AD (15 subjects) than it was in healthy controls (15 subjects).

The current mainstream method for studying brain functional connectivity is to convert a functional connectivity matrix into a graph via graph theory analysis. After completing a characterization of the topological properties of the graph, the clustering coefficient and characteristic path length–two indices that characterize a graph and correspond to the two basic principles of brain functional organization, namely, functional

[1]https://www.who.int/mental_health/management/depression/en/

separation and integration (Friston, 2009), respectively–are used to distinguish between patients with neurological disorders and healthy controls. In addition, these two indices can comprehensively reflect the small-world characteristics of the network. Randomization of network topology and distribution of the small-world network architecture have been consistently shown in AD, schizophrenia, and depression. Stam et al. (2006) found that the characteristic path length in AD patients was significantly longer than that in healthy controls and demonstrated that AD is characterized by loss of small-world network characteristics. Rubinov et al. (2009) found a randomization of the small-world network structure in schizophrenia. One EEG study showed that the loss of small world characteristics in the sleep functional brain network in MDD indicates a disruption of topological organization caused by this disease (Zhang et al., 2011). In addition, Leistedt et al. (2009) reported that the healthy controls' neural network is closer to the ordered part of the rewiring scale, while the depressed patients' brain network during sleep is closer to the random part of the scale. Although changes in brain function connectivity in MDD have been known, the functional brain network structure of mild depression is unclear. Therefore, we analyzed the functional brain network of mild depression through graph theory.

Deep neural networks have recently achieved great success in the widespread application of large-scale image- (Hinton et al., 2012), video- (Karpathy et al., 2014), and text-based recognition tasks (Hermann et al., 2015; Zhang and Lecun, 2015). Convolutional neural networks (CNNs) lie at the core of the best current architecture processing methods for both image and video data, primarily due to the advantages of CNNs in processing 2D input data (Cecotti and Graser, 2011). CNN has good performance in the field of biological image classification, and the features learned from CNN are often better than handcrafted features (Bello-Cerezo et al., 2019). The studies of Nanni et al. (2019a; 2019b) show that ensemble system of handcrafted and learned features can boost the performance of CNN in bioimage classification. In neural signal classification, several studies have used different methods to convert EEG signals into image representations. One, in particular, used the short-term Fourier transformation method to convert EEG time series into 2D images and combined 1D CNNs and stacked autoencoders to classify EEG motor imagery signals (Tabar and Halici, 2016). This approach yielded a 9% improvement over the winning algorithm. A new representation of EEG signals was proposed by Bashivan et al. (2015) that preserves the structure of EEG data across space, time, and frequency bands. In this approach, 3D electrode locations were projected onto a 2D surface using azimuthal equidistant projection, and the spectral power within three prominent frequency bands was extracted for each location. This was then used to form topographical maps that were subsequently combined to form three-channel images. Finally, these three-channel images were input into a deep convolutional recurrent neural network to further classify the EEG signals. Inspired by these studies and the powerful ability of CNNs to process 2D data, we innovatively applied a CNN to the 2D data form of functional connectivity matrices and constructed a classification model for mild depression as a depression recognition method other than graph theory.

In addition, research on facial emotion processing plays a significant role in the study of emotion and cognition in patients with depression. For example, David et al. (Rubinow and Post, 1992) conducted an experiment where 17 patients with depression and 31 healthy controls were asked to recognize seven affective states within images of facial expressions. The results revealed that patients with depression were significantly impaired in their ability to recognize facial affect; specifically, they made significantly (or nearly significantly) fewer correct matches for sad, happy, and interested face items. Similarly, James et al. (Cavanagh and Geisler, 2006) examined mood-relevant emotion processing in individuals with depression using event-related potentials. Mixed-model analyses of variance revealed significantly reduced P3 amplitudes and P3 latencies for happy faces in participants with depression. The authors interpreted these findings as providing evidence for a diminished cognitive processing ability during emotion discrimination in individuals with depression. Collectively, these studies reveal differences in the brain activity of patients with depression and healthy individuals during common face-processing tasks. Therefore, in the current study, we used a facial expression paradigm to investigate differences in brain functional connectivity between individuals with mild depression and healthy controls.

In the present study, we first studied the differences in the brain functional network between the mildly depressed group and the normal control group using graph theory. We calculated the four functional connectivity matrices of coherence, correlation, PLV, and PLI and converted them into binary undirected graphs. We calculated the characteristic path length, clustering coefficient, and the small-world properties of the brain functional network of the two groups to investigate the differences in these indices between two groups. Then, we proposed a novel approach aimed at improving the ability to recognize individuals with mild depression. Inspired by the proven utility of CNNs in image processing, we used a CNN to process EEG signals. EEG time series signals were converted into images through calculating functional connectivity matrices, and then these images from the two groups were used to classify individuals with mild depression and healthy controls in a CNN.

MATERIALS AND METHODS

Network Analysis

Coherence Analysis

Coherence is defined as the spectral cross-correlation between two signals normalized by their power spectrum. Coherence is computed mathematically as:

$$Coh_{xy}(f) = \frac{\left|\frac{1}{n}\sum_{k=1}^{n} Ax(f,k)Ay(f,k)e^{i\left(\varphi_x(f,k)-\varphi_y(f,k)\right)}\right|}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n} A_x^2(f,k)\right)\left(\frac{1}{n}\sum_{k=1}^{n} A_y^2(f,k)\right)}} \quad (1)$$

where n is the number of data points in a trial. A and φ are the amplitude and phase of the signal, respectively. The

numerator term represents the cross-spectral density on a single trial between the signals x and y at frequency f. The denominator represents the square root of the product of the power estimates on a single trial of the signals x and y at frequency f.

The coherence can then be concisely defined as:

$$Coh_{xy}(f) = \frac{|Sxy(f)|}{\sqrt{Sxx(f)Syy(f)}} \quad (2)$$

where Sxy(f) is the cross-spectral density of the signal, and $Sxx(f)$ and $Syy(f)$ are the power spectral density of signals x and y, respectively. Values of Coh_{xy}will always satisfy $0 \leq Coh_{xy}(f) \leq 1$, where 0 represents no coupling, and 1 indicates maximum linear interdependence between two signals.

Correlation Analysis

Correlation (Pearson correlation coefficient) is used to estimate the level of linear dependence between two electrode channels in the time domain. The correlation is given by the following:

$$Corr_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} \quad (3)$$

where $Cov(x, y)$ is the covariance between electrodes x and$y$, $\sigma_x$and σ_y are the standard deviations of the electrodes x and y, respectively. $Corr_{xy}$ has a value between 1 and -1, where 1 is a total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation. The greater the absolute value of $Corr_{xy}$, the stronger the correlation.

Phase Locking Value Analysis

The PLV (Mormann et al., 2000) assumes that the signal amplitude and phase are statistically independent, and thus, only the phase synchronization is used to estimate a possible functional interaction between the EEG signals of two channels. When Ax(.,.) = Ay(.,.) = 1 in Eq. (1), the PLV is obtained (Lachaux et al., 1999), as follows:

$$PLV_{xy}(f) = \frac{\left|\frac{1}{n}\sum_{k=1}^{n} 1x(f,k)1y(f,k)e^{i\left(\varphi x(f,k)-\varphi y(f,k)\right)}\right|}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n} 1_x^2(f,k)\right)\left(\frac{1}{n}\sum_{k=1}^{n} 1_y^2(f,k)\right)}} = \left|\frac{1}{n}\sum_{k=1}^{n} e^{i\left(\varphi_x(f,k)-\varphi_y(f,k)\right)}\right| \quad (4)$$

The value of PLV is between 0 and 1, where 0 represents a lack of synchronization, and 1 represents perfect phase synchronization.

Phase Lag Index Analysis

The PLI (Stam et al., 2007) is a measure of the asymmetry of the distribution of phase differences between two EEG signals and is defined as follows:

$$PLI_{xy}(f) = |\frac{1}{n}\sum_{k=1}^{n} sign(\phi_x(f,k) - \phi_y(f,k))| \quad (5)$$

where n is the number of data points, and $\phi_x(f,k) - \phi_y(f,k)$ represents the phase synchronization between the signals in channels x and y at the frequency f. It is essential to know the instantaneous phase of the two signals involved, which can be achieved using the analytical signal based on the Hilbert transform (Bruns, 2004), to compute the phase synchronization. The $PLI_{xy}(f)$ ranges between 0 and 1, where 1 indicates perfect phase synchronization and 0 indicates either no coupling or coupling with a phase difference centered around 0 mod π. PLI, contrary to the other methods, basically disregards perfect coupling (zero phase) or phase opposition coupling, which cannot be distinguished from no coupling.

Four kinds of metrics, coherence, correlation, PLV, and PLI, were calculated using all 128 EEG channels through the Neurophysiological Biomarker Toolbox[2]. For each trial, four 128 × 128 matrices were obtained (128 was the number of EEG channels). For each kind of metric, we considered the following EEG bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–70 Hz).

Complex Network Analysis

Applying graph theoretical analysis to a functional connectivity matrix is to convert the matrix into a binary undirected graph. The functional connectivity matrix can be converted to a graph by considering a threshold T. If the functional connectivity metric between a pair of channels exceeds T, then there is an edge, otherwise there is no edge. The choice of T has an important impact on the constructed graph, i.e., a low T-value will result in a densely connected network, while a large T-value will result in a sparse network. Since there is no unique method to select the appropriate threshold, we studied the entire range of values of T ($0 < T < 1$, with increments of 0.025 for coherence, correlation, and PLV and 0.005 for PLI) and repeated the full analysis for each value of T.

Once we converted the functional connectivity matrix into a graph, the next step is to characterize the graph in terms of its characteristic path length L and its cluster coefficient C. These two indices correspond to the two basic principles of brain function organization, that is, functional integration and segregation (Friston, 2009). Functional integration reflects the brain's ability to organize and combine information from different brain regions, and could be well measured by the characteristic path length L, which is obtained by calculating the average shortest path length between all pairs of nodes. Functional segregation reflects the ability to process information in a specialized way. Clustering coefficient C is often used to measure functional segregation. "C quantifies the number of connections between the nearest neighbours of a node as proportion of the maximum possible number of connections" (Strogatz, 2001). After clustering coefficients of all nodes are calculated, the average is taken as the clustering coefficient of the network.

To establish the network topology characteristics, the ratios C/Cr and L/Lr are calculated as a function of the threshold T, where Cr and Lr represent the values of C and L for matched random networks. Random networks have the same nodes and connectivity as the original network, whereas their choice of connected nodes is completely random. Following the previous studies (Watts and Strogatz, 1998; Stam et al., 2006; Li et al., 2015) that generated 20 random networks, 20 matched random networks are generated here for each actual network of each

[2]http://www.nbtwiki.net/

subject, and their average Cr and Lr are calculated to compare with the actual network. When values of C/Cr are significantly greater than 1, while the values of L/Lr are close to the value of 1, the small-world network organization is evident. In addition, Humphries et al. (2005) defined a single scalar index S to quantify the "small-world-ness" of a network, $S = \frac{\gamma}{\lambda} = \frac{C/C_r}{L/L_r}$, and a value of S greater than 1 is often used as a simple indicator of small-world organization.

The software used in this study was the Brain Connectivity Toolbox (Rubinov and Sporns, 2010) for calculating two graph theoretical measurements.

Convolutional Neural Network

Input

We used only the upper triangular part of the 128 × 128 functional connectivity matrices due to the symmetry of matrices and computational efficiency of neural networks. This triangular part contains 8,256 elements, of which 128 are on the diagonal. As the elements on the diagonal are the coherence/correlation/PLV/PLI values between an electrode and itself, we excluded these 128 elements. Since the input to the CNN is generally square, we organized the remaining 8,128 elements into a square and made the size of the square as small as possible. In addition, considering that the functional connectivity metric located in the last part of the upper triangle is calculated based on the electrodes distributed on the face, they are more affected by myoelectricity and ocular electricity. We, thus, organized the 8,128 elements line by line into a 90 × 90 matrix and discarded the last 28 elements, which are calculated based on the eight electrodes E121, E122, E123, E124, E125, E126, E127, E128. The transformation process is depicted in **Figure 1**. A new 90 × 90 matrix was transformed into a single-channel image and served as the CNN's input. It is worth mentioning that the elements in the functional connectivity matrix are arranged according to the increasing electrode number, and the neighborhood relationship between the electrodes did not represent the neighborhood information of 128 electrodes on HydroCel Geodesic Sensor Net. In other words, the location information of the electrodes was not preserved in the functional connectivity matrix. In addition, studies such as those mentioned in this review (De Aguiar Neto and Rosa, 2019) involve little information on the location of the electrodes when performing depression recognition. Similarly, in our study, the key piece of information for depression recognition is the functional connectivity value between the electrode pairs rather than the location of these values. So the element rearrangement during this transformation process will not have any obvious influence on mild depression recognition.

In addition, inspired by the work of Bashivan et al. (2015), we merged the 90 × 90 single-channel functional connectivity matrices into an image with three channels. We used the three bands with the best classification performance for each functional connectivity matrix as the red, green, and blue channels to generate a three-channel image that was then fed into the CNN (**Figure 2**).

Architecture

It is known that CNNs are multi-layer neural networks with several convolution–pooling layer pairs and a fully connected output layer (Tabar and Halici, 2016). A standard CNN (LeCun et al., 1990) is designed to recognize the shape in the images. As shown in **Figure 3**, the CNN model used here mainly included the following types of layers: convolution, pooling, activation function, fully connected layer, and softmax layer. Below, we provide brief descriptions of each of these.

In simulating orientation-selective simple cells in the primary visual cortex of the brain (LeCun et al., 2010), convolution is considered to be an indispensable component of CNN frameworks. The input data x is a tensor with K channels. There are K′ filters of weights W generating K' output y, as shown in Eq. (6). In our network architecture, $K = 1$ or $K = 3$, $K' = 32$.

$$y_{i'j'k'} = \sum_{ijk} w_{ijkk'} x_{i+i',j+j',k} \tag{6}$$

In the present study, we adopted two stacked convolutional layers as the basic structure of the CNN. Each convolutional layer used 32 3 × 3 small filters with a convolution stride of one. Convolution layer inputs were padded with one pixel to preserve the spatial resolution after convolution.

Pooling was also an important operation in the CNN framework. Multiple pooling methods serve to simulate complex cells in the brain's visual cortex (LeCun et al., 2010). In practice, the pooling method known as max pooling has been shown to work better than does average pooling. Specifically, max pooling calculates the max response of each feature map in a patch of size $p \times p$ as follows:

$$y_{ijk} = \max\left\{y_{i'j'k} : i \le i' < i + p, j \le j' < j + p\right\} \tag{7}$$

Max pooling is performed over a 2 × 2-pixel window here with a stride of two.

Additionally, we selected a rectified linear unit as an activation function for our CNN model, since it has performed both accurately and quickly in other CNN models (Dahl et al., 2013; Donahue et al., 2014). Its output is given by the following formula:

$$y_{ijk} = \max(0, x_{ijk}) \tag{8}$$

The max pooling layer is followed by a fully connected layer with 512 hidden cells, the last layer of which is a two-way softmax layer. The softmax layer is usually used for classification in CNN, and several classifications use several ways of softmax. In our study, since our goal is to distinguish between depression and normal subjects, that is, two classifications, we used two-way softmax. Softmax changes the output of the original neural network into a probability output to represent the probability that a sample belongs to different categories. Assuming that the output of the original neural network is $y_1, y_2, \ldots y_n$, then the output after softmax processing is:

$$softmax(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{9}$$

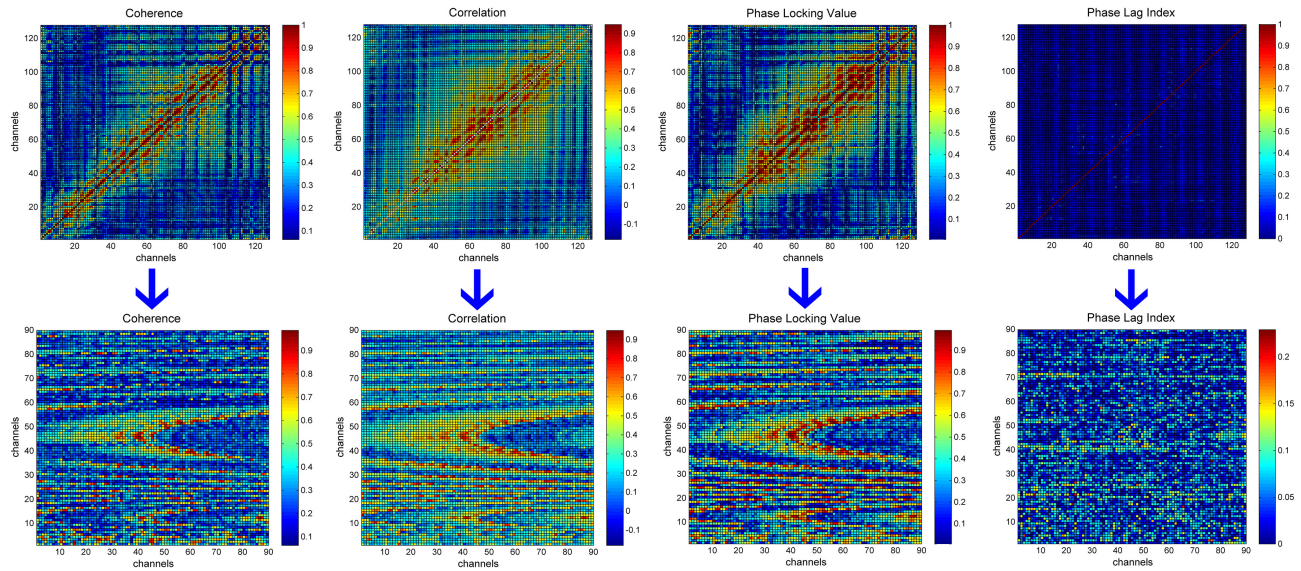


FIGURE 1 | The transformation process of functional connectivity matrices. The first row represents the original functional connectivity matrices corresponding to 128 electrodes. The second row represents the new 90 × 90 matrices obtained by rearranging the elements in the upper triangle of the original functional connectivity matrices. From left to right are, the functional connectivity matrices for coherence, correlation, phase locking value, and the phase lag index.

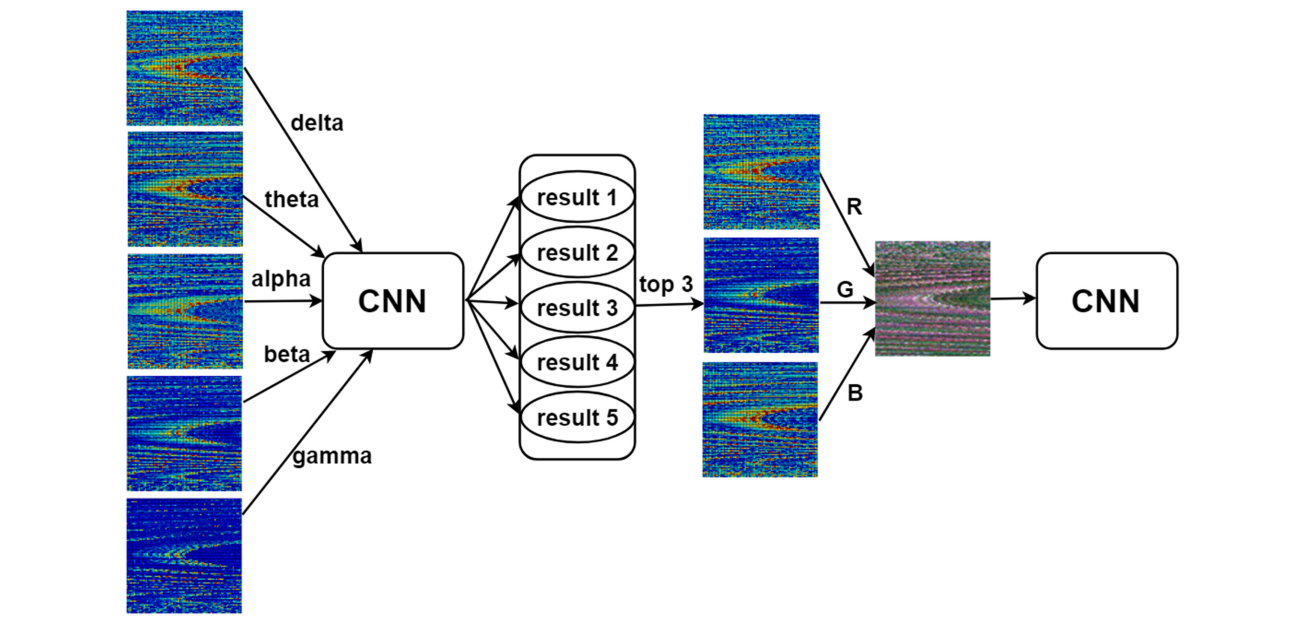


FIGURE 2 | Three-channel input of the convolutional neural network (CNN). The three bands with the best classification performance among the five bands (delta, theta, alpha, beta, and gamma) of each type of functional connectivity matrix were used as the red (R), green (G), and blue (B) channels of the image to generate a three-channel image that was then used as input for the CNN.

Training

Most of the code used in the present study was written in Python (version 3.5). The CNN model was trained and tested based on a lightweight library named Lasagne[3], which was used for building and training the neural networks in Theano (Al-Rfou et al., 2016).

The network was trained with the Adam algorithm (Kingma and Ba, 2014), which has demonstrated a competitively fast convergence rate in the training of neural network (Bashivan et al., 2015). The learning rate was 10^{-3}, and the decay rates of the first and second moments were 0.9 and 0.999, respectively.

[3]https://github.com/Lasagne/Lasagne

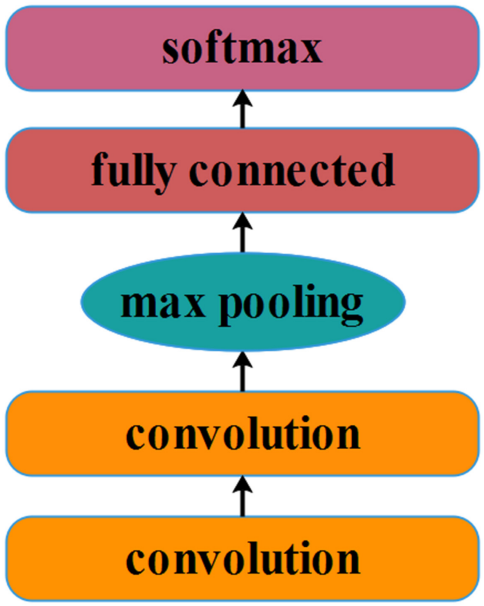


FIGURE 3 | Schematic structure of the convolutional neural network we constructed.

The weight initialization method we used was Xavier (Glorot and Bengio, 2010), which has been shown to be effective in network training (Jiao et al., 2018). We also used early stopping to monitor the performance of the model over the validation sets. Dropout (Hinton et al., 2012) with a probability of 0.5 was used across all fully connected layers. The effectiveness of this dropout to reduce overfitting in deep neural networks with millions of parameters has been shown previously (Krizhevsky et al., 2012) and in neuroimaging applications specifically (Plis et al., 2014). Because of implicit regularization imposed by smaller convolution filter sizes, the network requires fewer epochs to converge. Given this, our model was trained over 15 epochs with a batch size of 30.

EXPERIMENTS

Participants

Fifty-one students (36 males, 15 females) aged between 18 and 24 were recruited from Lanzhou University and participated in the study. All of them had no prior history of psychopathology and had normal or corrected-to-normal vision. All participants were interviewed by psychologists after completing an investigation in the psychological screening system of Lanzhou University. According to psychologists, 24 of them were considered depressed, and the remaining 27 were healthy. Also, participants were asked to finish the Beck Depression Inventory test-II (BDI-II) (Beck et al., 1996) before experiment. Analysis of BDI-II showed that the BDI-II scores in the depression group ranges from 14 to 28, corresponding to mild depression, whereas the BDI-II scores of the healthy group were all lower than 13. In order to ensure that the number of samples was balanced for the two groups, 24 healthy participants were selected to comprise our control group. **Table 1** shows the demographic characteristics and BDI-II scores of the two groups. This study was approved by the Ethics Committee of Lanzhou University Second Hospital (reference number: 2015A-037) and was conducted in full compliance with the ethical standards outlined in the Declaration of Helsinki. All participants signed an informed consent form before the experiment and received monetary compensation after study completion.

TABLE 1 | Characteristics of the individuals with mild depression and healthy controls.

	Mild depression (n = 24)	Healthy controls (n = 24)
Age (years)	20.96 ± 1.95	20 ± 2.02
BDI-II score	17.63 ± 3.41	4.63 ± 3.00
Sex		
Female	6	9
Male	18	15

Data are presented as the number or mean ± standard deviation. BDI-II, Beck Depression Inventory test-II.

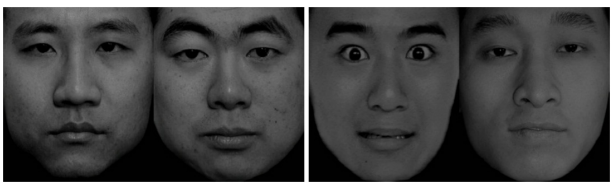


FIGURE 4 | Representative picture from the Neu_block **(Left)** and Emo_block **(Right)**.

Materials and Procedures

The stimuli used in the experiment were derived from the China Facial Affective Picture System (Gong et al., 2011), a subsystem in the standardized emotional stimuli picture system. We selected pictures of 45 neutral faces and 15 negative faces, including three each of angry, sad, disgusted, surprised, and fearful faces. The experiment consisted of two blocks, emotional block (Emo_block) and neutral block (Neu_block). Each block contained 15 trials. Each trial in the Emo_block contained an emotional expression and a neutral facial expression picture. Each trial in the Neu_block contained two pictures of neutral facial expressions. Each picture appeared randomly on either the right or left side of the screen. All faces were presented without hair, glasses, beards, or other facial accessories, and the two facial expressions used in each trial were combined into one image and presented on the screen. All 30 images were processed with the Adobe Photoshop CS6 software (Adobe Systems Incorporated, San Jose, CA, United States), and the size (1,280 × 738 pixels, 10.84 × 6.25 cm) and gradation were made uniform (**Figure 4**).

Instructions were displayed on the screen before the beginning of each block. Further, four practice trials identical to the real trials were included to ensure that all participants understood the experimental procedures. Each trial was presented for 6 s. A black background was also presented for 2 s between every two trials. The participants were comfortably seated 60 cm from a 17-inch

liquid crystal display monitor with a resolution of 1,024 × 768 and instructed to view each trial freely. A 2-min rest period followed the Neu_block, which was followed by the Emo_block. The whole protocol took approximately 7 min.

EEG Acquisition and Data Preprocessing

Electroencephalography signals were collected using a 128-channel HydroCel Geodesic Sensor Net (Electrical Geodesics, Inc.). The EEG electrodes were placed according to the HydroCel Geodesic Sensor Net128 Channel Map (Version 1.0) and referenced to Cz. The impedance of all electrodes was maintained below 60 kΩ (Ferree et al., 2001), and EEG signals were continuously recorded at a sampling frequency of 250 Hz. Because the signal gathered between two trials was not valid, each participant's continuous EEG signals were divided into 30 6-s segments according to marks in the time series. All EEG signals were high-pass filtered at a 0.5-Hz cutoff frequency and low-pass filtered at a 70-Hz cutoff frequency. Eye movement and muscle activity artifacts were discarded using Net Station Waveform Tools. Furthermore, because ocular artifacts are presented in the frequency band between 0 and 16 Hz, they overlap with the alpha rhythm frequency band (8–13 Hz). This study, therefore, used FastICA to eliminate ocular artifacts, as this approach has previously been shown to be effective in delineating between overlapping frequency bands (Hu et al., 2011). We used MATLAB R2010a (Mathworks, Natick, MA) to process all data.

RESULTS

Results of Network Analysis

In this section, our results show a significant difference in the small-world property between the two groups by showing the relationship between the small-world property and the thresholds of the two groups (Small-World Property section); statistical analysis of the mean clustering coefficient for each electrode was performed to determine the specific distribution of functional connectivity differences between the two groups in the brain (Statistical analysis of the mean clustering coefficient for each electrode section). In The Topological Structures of the Networks section, we visualized the functional connectivity differences between the two groups using coherence as an example.

Small-World Property

Figure 5 shows the mean small-world index S as a function of threshold T for the two groups. For coherence, a significant difference between the two groups was found in the delta band under Neu_block and in the theta band under Emo_block. The value of *S* in the mildly depressed group was significantly lower than that in the healthy control at most of the *T*-values points (indicated by asterisks). For the correlation, PLV, and PLI, we found a similar difference only in the delta band under both blocks.

In addition, from the separate analysis of γ and λ, we found that the γ of both groups was greater than 1, but the difference between the groups was not significant. The λ of both groups was less than 1 and close to 1, indicating that the brain networks of both groups showed small-world characteristics; however, the λ of the mild depression group was significantly higher than that of the healthy control group, showing that the functional brain network of the mild depression group deviated from the small-world network.

Statistical Analysis of the Mean Clustering Coefficient for Each Electrode

Table 2 shows the statistical analysis of the mean clustering coefficient based on independent samples *t*-tests for each electrode for coherence and PLV under Neu_block and Emo_block. For both metrics, we observed a similar pattern, that is, in the delta band, the clustering coefficient of the mild depression group was significantly higher than that of the healthy control group for the electrodes distributed in the temporal and frontal regions, but the results were exactly the opposite for the electrodes distributed in the central and parietal–occipital regions; in the beta band, the mild depression group had significantly lower clustering coefficient in the parietal–occipital region than the healthy controls under Neu_block and Emo_block, and the difference was mainly found in the right hemisphere. We observed a different pattern for correlation. Differences between the two groups were observed in all five bands, and we listed those common electrodes on which there are differences between the two groups in the five bands in **Table 3**. Similarly, the mild depression group had a significantly lower clustering coefficient than had healthy controls under Neu_block and Emo_block, and the differences were mainly found in the parietal–occipital region of the right hemisphere. The above statistical analysis was performed at a threshold T of 0.15. For the PLI, no differences between the two groups were found on all five bands.

The Topological Structures of the Networks

In order to visualize the distribution of coherence differences between the two groups, we plotted the topological structures of functional networks between the mild depression group and healthy controls at a threshold of 0.1 in the delta band under Neu_block. **Figure 6** was drawn based on the absolute value obtained by subtracting one coherence matrix from another one. When the absolute value was greater than the threshold of 0.1, there would be a connection in the figure. At a threshold of 0.1, the difference in coherence between the two groups in the delta band can be clearly seen. Most of the connections are distributed in the central, temporal regions and parietal–occipital region of the right hemisphere, and a few connections are distributed in the frontal region.

Classification Results

To evaluate the proposed CNN classification model, 24-fold cross-validation was adopted. In each fold, the functional connectivity matrices from one control and one participant with mild depression were used for testing, while matrices from remaining participants were utilized as training data and for validation. In each fold, the functional connectivity matrices of another mild depression and healthy control participants were used for validation. The functional connectivity matrices of the

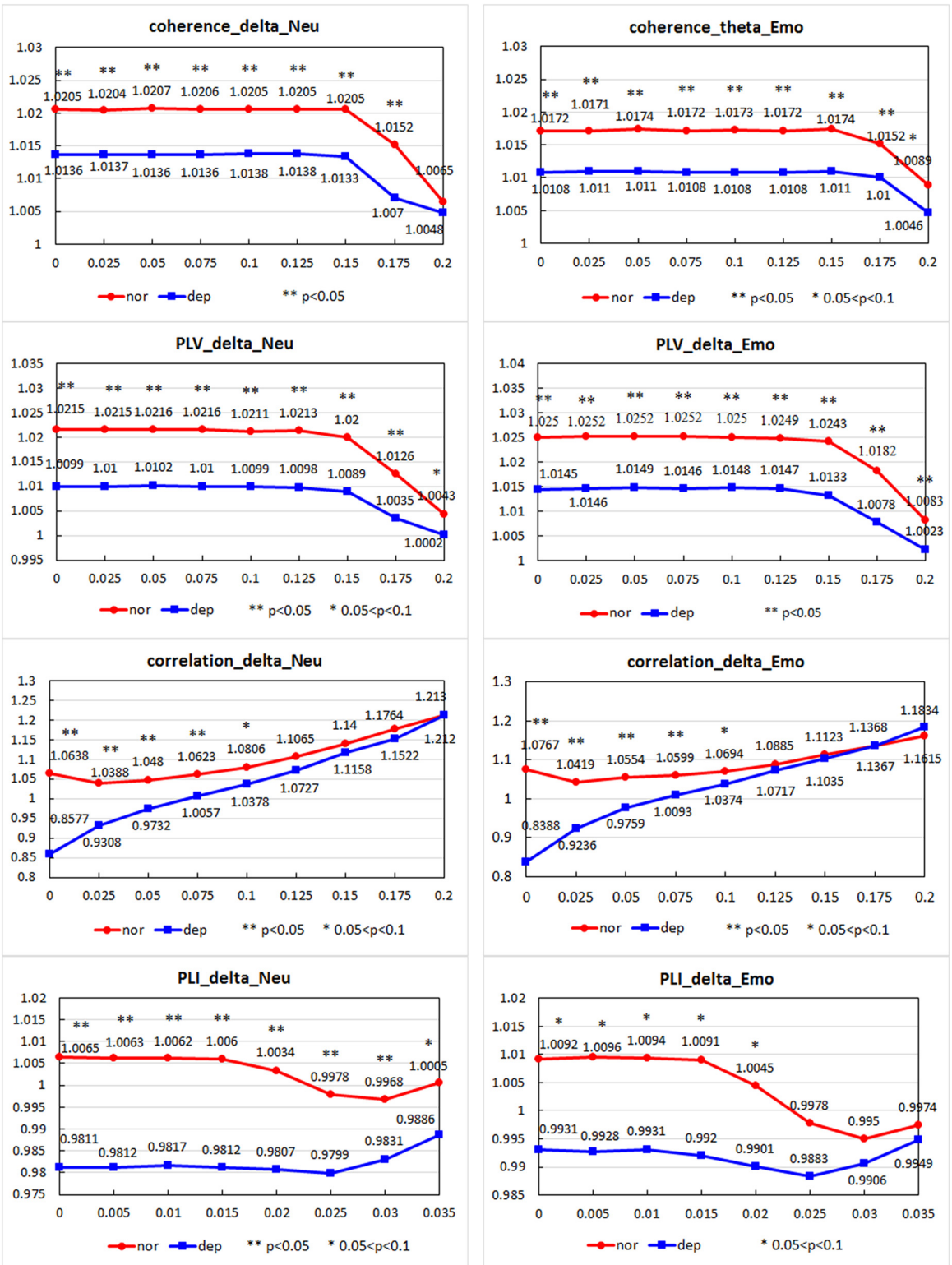


FIGURE 5 | The mean small-world index S as a function of the threshold T for the two groups under Neu_block and Emo_block. For the four metrics of coherence, correlation, phase locking value (PLV), and phase lag index (PLI), the values of S in the mildly depressed group were significantly smaller than those in healthy controls. The title of each subgraph is denoted as "metric_band_block." The horizontal axis presents different thresholds. The values under asterisks (*) denoted the values of *S* at different thresholds when the difference between two groups was significant at these thresholds. **The value of *p* smaller than 0.05; *the value of *p* between 0.05 and 0.1.

TABLE 2 | Statistical analysis of clustering coefficient using coherence and PLV (T = 0.15).

		Neu_block						Emo_block					
Metric	**Band**	**Electrode**	**N**	**D**	**F**	**P**	**region**	**Electrode**	**N**	**D**	**F**	**P**	**region**
Coherence	Delta	E31	0.326	0.3026	0.007	0.039	LC	*E2*	*0.2719*	*0.2851*	*2.699*	*0.022*	*RF*
		E60	0.3855	0.3646	0.016	0.047	LPO	E7	0.3045	0.2837	0.315	0.02	LC
		E78	0.3777	0.3543	0.014	0.039	RPO	E31	0.3287	0.3026	0.032	0.018	LC
		E79	0.3658	0.3394	0.138	0.033	RC	E60	0.38	0.3585	0.311	0.039	LPO
		E89	0.3859	0.3436	19.833	0.017	RPO	E78	0.3729	0.3503	0.101	0.045	RPO
		E111	0.3392	0.3183	0.525	0.043	RC	E89	0.38	0.341	9.471	0.022	RPO
		E119	*0.3155*	*0.3341*	*0.25*	*0.043*	*RT*	*E119*	*0.3096*	*0.3308*	*0.07*	*0.017*	*RT*
		E126	*0.2842*	*0.2995*	*1.457*	*0.04*	*RF*	*E126*	*0.279*	*0.2984*	*0.045*	*0.013*	*RF*
								E127	*0.2751*	*0.2943*	*1.959*	*0.012*	*LF*
	Beta	E63	0.3572	0.3238	2.605	0.023	LT	E56	0.3515	0.3177	14.499	0.034	LT
		E67	0.3541	0.3313	1.274	0.03	LPO	E60	0.3569	0.3293	7.942	0.033	LPO
		E68	0.3526	0.3303	1.337	0.032	LPO	E63	0.3588	0.319	10.884	0.017	LT
		E69	0.3575	0.329	0.527	0.035	LPO	E67	0.3577	0.3298	5.523	0.019	LPO
		E71	0.3616	0.3381	2.142	0.046	LPO	E68	0.326	0.3291	5.474	0.023	LPO
		E73	0.3492	0.3185	3.887	0.031	LPO	E69	0.3587	0.3254	4.372	0.028	LPO
		E74	0.3593	0.3319	1.998	0.042	LPO	E71	0.3651	0.336	6.753	0.03	LPO
		E77	0.3543	0.3301	2.533	0.047	RPO	E73	0.3512	0.3148	10.285	0.026	LPO
		E78	0.443	0.3195	3.511	0.038	RPO	E74	0.3619	0.3277	5.502	0.025	LPO
		E84	0.3624	0.3353	1.701	0.044	RPO	E77	0.3575	0.3293	4.993	0.038	RPO
		E85	0.3536	0.3255	1.354	0.027	RPO	E78	0.3471	0.3205	6.769	0.049	RPO
		E89	0.3612	0.3148	0.402	0.006	RPO	E84	0.3635	0.3335	4.796	0.043	RPO
		E91	0.3587	0.3293	1.92	0.036	RPO	E85	0.3552	0.3254	3.586	0.032	RPO
		E98	0.3495	0.3223	7.493	0.044	RC	E86	0.3461	0.3155	6.666	0.033	RC
								E87	0.3275	0.3016	7.723	0.043	RC
								E89	0.3622	0.3093	0.288	0.004	RPO
								E90	0.3622	0.3304	4.187	0.036	RPO
								E91	0.3598	0.3281	3.06	0.036	RPO
								E97	0.3606	0.3266	8.773	0.034	RT
								E98	0.3519	0.3187	10.346	0.027	RC
								E110	0.3393	0.3068	12.954	0.043	RC
PLV	Delta	E1	*0.302*	*0.3188*	*0.235*	*0.034*	*RF*	E7	0.3408	0.3149	0.277	0.044	LC
		E2	*0.3069*	*0.3248*	*0.117*	*0.025*	*RF*	E60	0.4267	0.4026	0.17	0.037	LPO
		E7	0.3437	0.3193	0.007	0.046	LC	E78	0.4173	0.3903	0.099	0.034	RPO
		E26	*0.3106*	*0.3283*	*2.938*	*0.028*	*LF*	E79	0.4054	0.3772	0.08	0.037	RC
		E31	0.3734	0.3447	0.464	0.018	LC	E80	0.3682	0.3414	0.781	0.048	RC
		E60	0.4225	0.4007	2.591	0.047	LPO	E85	0.4253	0.4017	0.04	0.038	RPO
		E78	0.416	0.3908	0.546	0.042	RPO	E86	0.4168	0.3908	0.55	0.026	RC
		E89	0.4224	0.3796	9.518	0.022	RPO	E87	0.395	0.3682	0	0.037	RC
		E111	0.3775	0.3555	1.77	0.043	RC	E89	0.427	0.3807	14.249	0.016	RPO
		E119	*0.352*	*0.3779*	0.001	0.017	*RT*						
		E126	*0.3178*	*0.3389*	0.326	0.029	*RF*						
		E127	*0.3124*	*0.333*	0.013	0.195	*LF*						
	Beta	E7	0.2751	0.234	2.879	0.036	LC	E78	0.3583	0.3124	0.582	0.017	RPO
		E23	0.2908	0.2573	0.951	0.043	LF	E79	0.3336	0.2936	1.278	0.031	RC
		E46	0.3411	0.2941	3.461	0.041	LT	E82	0.3742	0.3331	0.008	0.042	RPO
		E60	0.3642	0.3238	2.789	0.045	LPO	E85	0.3675	0.329	0.146	0.046	RPO
		E66	0.3705	0.3344	1.202	0.048	LPO	E89	0.3686	0.3038	1.561	0.012	RPO
		E67	0.3717	0.3335	2.427	0.044	LPO						
		E68	0.3713	0.3323	1.961	0.038	LPO						
		E70	0.3745	0.3379	3.355	0.049	LPO						

(Continued)

TABLE 2 | Continued

		Neu_block						Emo_block					
Metric	**Band**	**Electrode**	**N**	**D**	**F**	**P**	**region**	**Electrode**	**N**	**D**	**F**	**P**	**region**
		E71	0.381	0.337	1.719	0.019	LPO						
		E74	0.3693	0.3232	1.418	0.023	LPO						
		E76	0.3765	0.3367	1.187	0.031	RPO						
		E77	0.3706	0.3262	2.14	0.022	RPO						
		E78	0.3567	0.3066	3.725	0.016	RPO						
		E79	0.3321	0.2893	3.849	0.037	RC						
		E82	0.3733	0.3189	2.234	0.009	RPO						
		E83	0.3784	0.334	1.224	0.018	RPO						
		E84	0.376	0.3317	1.153	0.03	RPO						
		E85	0.365	0.3128	1.226	0.013	RPO						
		E86	0.3526	0.3048	1.947	0.024	RC						
		E87	0.3243	0.2815	5.833	0.024	RC						
		E88	0.3529	0.303	3.609	0.026	RPO						
		E89	0.371	0.2894	0.101	0.001	RPO						
		E90	0.3708	0.3206	0.907	0.012	RPO						
		E91	0.365	0.321	1.094	0.027	RPO						
		E95	0.3565	0.3096	2.015	0.04	RPO						
		E96	0.3577	0.3104	3.751	0.032	RT						
		E97	0.3551	0.3087	4.479	0.034	RT						
		E98	0.346	0.304	6.075	0.042	RC						
		E105	0.3103	0.2674	7.592	0.045	RC						
		E111	0.316	0.2876	10.572	0.031	RC						

N, mean clustering coefficient of normal control group; D, mean clustering coefficient of mild depression group; F and p, mean F- and P-value of independent samples t-tests. The electrode distribution of the 128-channel Geodesic Sensor Net is in the Section ***Appendix****. The electrodes are divided into eight regions: the left frontal (LF), right frontal (RF), left temporal (LT), right temporal (RT), left central (LC), right central (RC), left parietal–occipital (LPO), and right parietal–occipital (RPO) regions. The italics indicate that the mean clustering coefficient of mild depression group at this electrode is lower than that of healthy control group.*

remaining 44 participants were used for training. This strategy allowed us to avoid records from one participant being divided into both training and test sets and, thus, yielding a falsely high classification accuracy. The classification performance of the validation data was used for selecting the hyperparameters (such as learning rate, learning rate decay, regularization coefficient, number of iterations, weight initialization, etc.) and as a stopping criterion in training to avoid overfitting of the training data.

Classification Performance of the Four Functional Connectivity Matrices Using the CNN

We assessed the classification accuracy of the functional connectivity matrices using our CNN method. Since the initial weights of CNN were randomly initialized and different initial weights will train a slightly different model, we repeated the training and testing procedures nine times and calculated the mean and standard deviation of nine test accuracy for further analysis. Coherence performed best under the Neu_block and Emo_block (77.78% for the two blocks). The second highest classification accuracy was achieved with the PLV (74% in the gamma band for the Neu_block, 73.33% in the delta band for the Emo_block). Correlation yielded a 71.46% accuracy for the Neu_block and a 65.83% accuracy for the Emo_block. The PLI achieved the lowest accuracy for both the Neu_block (63.41%) and the Emo_block (55.60%). The classification accuracy of each functional connectivity matrix is shown in **Table 4A**.

Receiver operating characteristic (ROC) curves are commonly used to present the results of binary decision problems in machine learning (Davis and Goadrich, 2006). AUC is the area under an ROC curve and has a value between 0 and 1, with a greater AUC value indicating a better classification ability of the model. In order to compare the classification performance of the four input forms more intuitively, ROC curves for four functional connectivity matrices were obtained in the delta, theta, alpha, beta, and gamma bands. These are shown in **Figures 7A–E**. From these ROC curves, we were able to obtain the same results as those displayed in **Table 4A**.

Performance Comparison Between Our Method and the Self-Established Baseline

In order to further demonstrate the efficiency of our model, we input the four functional connectivity metrics into the four classic classifiers that are widely used in computer aided detection systems for depression (Hosseinifard et al., 2013; Li et al., 2016a; Cai et al., 2018), namely, BayesNet (BN), logistic regression (LR), k-nearest-neighbor (kNN), and random forest (RF), to classify participants into one of two classes, mild depression or healthy. The accuracy obtained from these four classifiers served as the baseline. These classifiers all used the default parameter values

TABLE 3 | Statistical analysis of clustering coefficient using correlation (T = 0.15).

Neu_block				Emo_block			
Electrode	***F***	***P***	**Region**	**Electrode**	***F***	***P***	**Region**
E30	2.118	0.042	LC	E60	0.74	0.036	LPO
E36	0.068	0.039	LC	E61	0.843	0.04	LPO
E41	0.187	0.043	LC	E65	1.056	0.034	LPO
E42	0.421	0.041	LC	E66	1.638	0.02	LPO
E46	0.152	0.02	LT	E67	0.879	0.013	LPO
E47	0.024	0.04	LC	E68	0.694	0.012	LPO
E53	0.049	0.029	LC	E69	0.159	0.03	LPO
E60	0.005	0.038	LPO	E70	0.928	0.009	LPO
E61	0.066	0.034	LPO	E71	0.506	0.026	LPO
E64	0.134	0.047	LPO	E73	0.009	0.047	LPO
E65	0.01	0.033	LPO	E74	1.822	0.012	LPO
E66	0.313	0.026	LPO	E76	1.521	0.01	RPO
E67	0.06	0.023	LPO	E77	0.87	0.006	RPO
E68	0.012	0.026	LPO	E78	0.02	0.006	RPO
E69	0.088	0.018	LPO	E79	0.102	0.01	RC
E70	0.044	0.006	LPO	E82	1.055	0.019	RPO
E71	0.463	0.016	LPO	E83	3.964	0.029	RPO
E73	0.64	0.022	LPO	E84	1.443	0.01	RPO
E74	0.012	0.007	LPO	E85	0.561	0.008	RPO
E76	0.212	0.008	RPO	E86	1.404	0.009	RC
E77	0.112	0.005	RPO	E87	0.156	0.046	RC
E78	0	0.007	RPO	E88	1.072	0.043	RPO
E79	0.002	0.016	RC	E89	7.387	0.005	RPO
E82	0.097	0.009	RPO	E90	2.974	0.023	RPO
E83	0.839	0.005	RPO	E91	1.522	0.015	RPO
E84	0.005	0.015	RPO	E92	1.002	0.016	RC
E85	0.135	0.014	RPO	E93	0.404	0.045	RC
E86	0.03	0.023	RC	E94	0.374	0.047	RPO
E87	0.268	0.047	RC	E97	1.489	0.036	RT
E88	0.068	0.028	RPO	E98	0.631	0.042	RC
E89	3.717	0.002	RPO	E104	0.133	0.045	RC
E90	0.048	0.024	RPO	E107	1.632	0.04	RT
E91	0	0.019	RPO				
E92	0.241	0.038	RC				
E93	0.021	0.031	RC				
E94	0.011	0.03	RPO				
E95	0.512	0.028	RPO				
E97	0.027	0.042	RT				
E98	0.022	0.031	RC				
E104	0.115	0.033	RC				
E105	1.097	0.031	RC				
E107	0.019	0.034	RT				
E110	0.694	0.023	RC				

F and p, mean F- and P-value of independent samples t-tests. The electrodes are divided into eight regions: the left frontal (LF), right frontal (RF), left temporal (LT), right temporal (RT), left central (LC), right central (RC), left parietal–occipital (LPO), and right parietal–occipital (RPO) regions.

implemented in Weka (Witten et al., 1999). The values of k used in kNN were 1, 5, and 10. The evaluation method we used herein for classification accuracy also underwent 24-fold cross-validation. Since the inputs for the above four classifiers were feature vectors, we calculated the mean value of all elements in the functional connectivity matrix for each trial, and the mean value for all trials constituted an N × 1 (N represents the number of trials) vector that was then input into the above four classifiers for classification. We only show the highest classification accuracies among the four classifiers in **Table 4B**.

Table 4B revealed that when functional connectivity metrics were input to the classic classifiers in the form of a feature

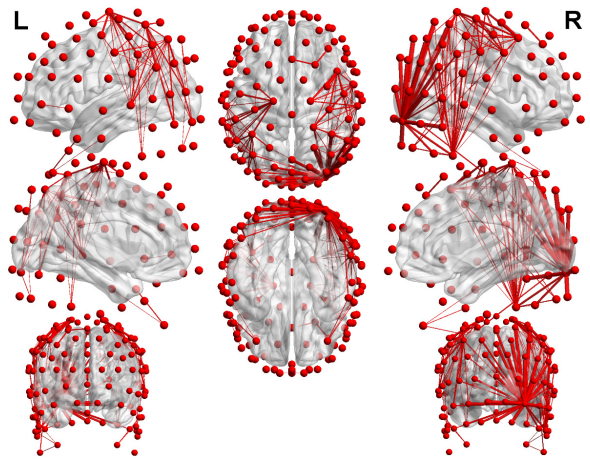


FIGURE 6 | The results of the distribution which was the difference of the coherence in the delta band. Red nodes represent 128 electrodes, and the red lines between nodes show the difference in coherence at the threshold of 0.1. The thicker the red line is, the more the threshold is exceeded.

vector, the recognition rates obtained were obviously lower than the accuracy of the coherence, correlation, and PLV obtained using the CNN. This fully demonstrated the improvements in classification accuracy that our method provided compared to the self-established baseline when the functional connectivity metrics were input into the CNN in the form of 2D data for classification. However, this improvement did not appear in the PLI of the Emo_block.

Classification Performance of the Three-Channel Functional Connectivity Matrix

As described above, for each functional connectivity matrix, we used the three bands of the top three classification performance as the R, G, and B channels to generate a three-channel image. We then input this three-channel image into the CNN for it to learn and classify. The bands used for each functional connectivity matrix are shown in **Table 5**.

The classification accuracy and ROC curves for the three-channel functional connectivity matrix obtained using the CNN are shown in **Table 6A** and **Figure 7F**. Compared with the results of the single channel, the three-channel functional connectivity matrix, which integrated information from three frequency bands, slightly increased the classification accuracy of coherence and correlation in the Neu_block and Emo_block. However, there was no increase in the PLV or PLI accuracy, which performed similarly to a single-channel construct. **Figure 8** displayed the comparison of the results of the delta, theta, alpha, beta, and gamma bands, as well as the three-channel coherence that performed the best among the four functional connectivity matrices. The obvious accuracy improvement of coherence through integrating three EEG frequency bands can be observed in **Figure 8**.

TABLE 4 | The classification accuracy of each band.

(A) The mean accuracy and standard deviation in each functional connectivity matrix using the CNN for the Neu_block and Emo_block.

		Accuracy (acc) % and standard deviation (std) %									
		Neu_block					**Emo_block**				
		Delta	**Theta**	**Alpha**	**Beta**	**Gamma**	**Delta**	**Theta**	**Alpha**	**Beta**	**Gamma**
Coherence	acc	77.78	75.88	68.64	75.91	75.77	74.48	72.10	72.67	75.57	77.78
	std	1.87	2.17	3.65	1.76	1.86	1.35	4.69	1.86	2.18	2.22
Correlation	acc	70.69	71.33	71.17	70.93	71.46	65.83	60.59	60.94	63.60	63.60
	std	4.30	3.92	3.68	4.19	3.38	3.92	5.51	5.26	5.18	4.94
PLV	acc	71.23	70.93	71.87	66.82	74.00	73.33	71.71	62.64	56.00	64.75
	std	2.12	2.83	3.37	4.50	5.05	1.94	2.88	4.08	3.18	7.11
PLI	acc	61.79	62.27	63.41	61.88	62.16	51.61	50.68	50.73	52.85	55.60
	std	3.89	3.55	3.02	3.33	4.21	2.37	1.10	0.66	1.61	2.82

(B) The highest classification accuracy obtained by inputting the feature vector corresponding to each functional connectivity metric into the four classic classifiers (BN, LR, kNN, and RF) in the Neu_block and Emo_block.

	Accuracy %									
	Neu_block					**Emo_block**				
	Delta	**Theta**	**Alpha**	**Beta**	**Gamma**	**Delta**	**Theta**	**Alpha**	**Beta**	**Gamma**
Coherence	57.22	50	51.67	56.67	61.53	53.47	54.86	51.94	57.78	61.94
Correlation	47.08	49.03	46.81	51.39	52.64	48.19	47.78	47.5	51.39	47.78
PLV	53.19	50	49.03	58.61	60.14	54.31	53.75	49.17	53.06	50.97
PLI	54.72	52.08	52.50	54.03	53.75	56.39	51.67	51.81	53.61	51.25

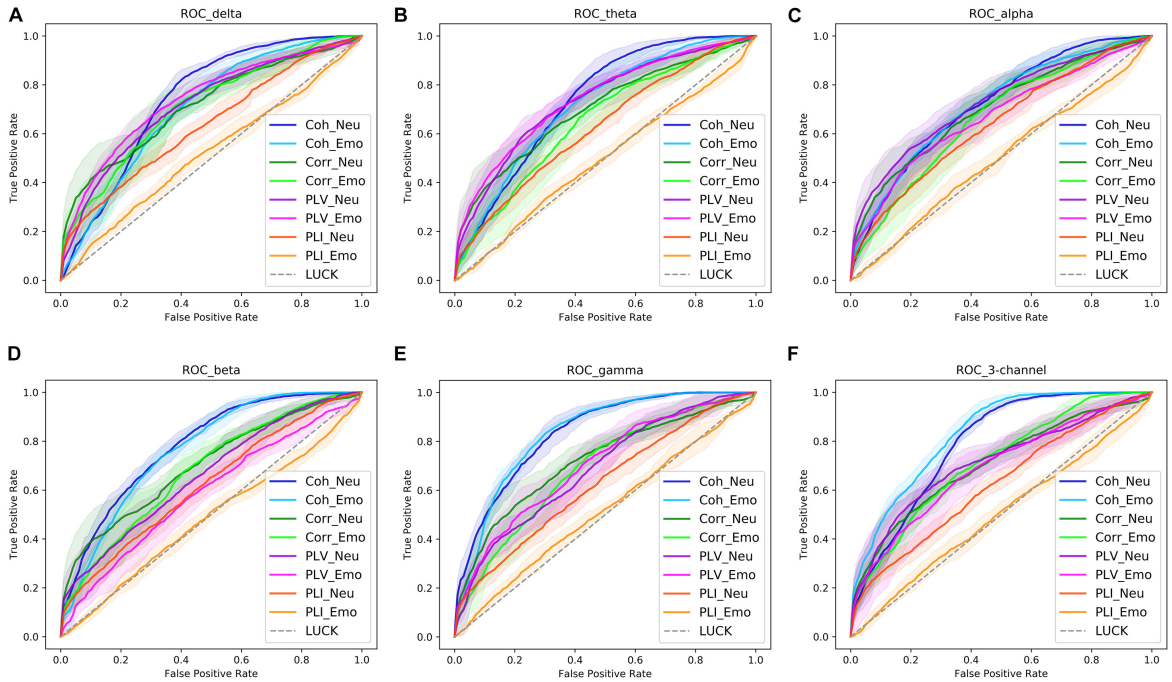


FIGURE 7 | The receiver operating characteristic (ROC) curves of four functional connectivity matrices [coherence (Coh), correlation (Corr), phase locking value (PLV), and phase lag index (PLI)] obtained in the delta **(A)**, theta **(B)**, alpha **(C)**, beta **(D)**, and gamma bands **(E)**, as well as the three-channel input **(F)**. The shaded area in the graph is the result of each measure, and the solid line in the shadow is the average result of nine measures. Neu, neutral picture condition; Emo, emotional picture condition.

TABLE 5 | The bands used for each functional connectivity matrix in the Neu_block and Emo_block.

	Coherence	Correlation	PLV	PLI
Neu_block	Delta, beta, theta	Gamma, theta, alpha	Gamma, alpha, delta	Alpha, theta, gamma
Emo_block	Gamma, beta, delta	Delta, beta, gamma	Delta, theta, gamma	Gamma, beta, delta

We also applied the three-channel strategy to the functional connectivity metrics in the form of feature vectors. The functional connectivity metrics on the three bands shown in **Table 5** formed an N × 3 feature vector (N represents the number of trials, and three represents the three bands) that was input into the four classic classifiers for classification. The classification accuracies (the highest classification accuracy among the four classifiers) are shown in **Table 6B**. Compared to the accuracy of the single band that was obtained using the classic classifiers shown in **Table 4B**, no improvement was provided. This demonstrated that the feature vector form of the functional connectivity metrics did not exhibit the advantages that integrating the three channels in the classic classifiers did and further confirmed the benefits of combining the CNN with functional connectivity matrices.

DISCUSSION

In the present study, we first illustrated the abnormal organization of functional connectivity network in mild depression by graph theory. Second, we proposed a novel approach of using a CNN to process functional connectivity matrix data in order to identify individuals with mild depression.

The Differences in the Small-World Network Between the Mildly Depressed Group and the Normal Control Group

Through graph theory, we found that the small-world index of the mild depression group was significantly lower than that of the healthy control group. Small-world networks are characterized by a high clustering coefficient and a short path length (Watts and Strogatz, 1998), while the mild depression group has a lower clustering coefficient and a larger characteristic path length than the healthy control group, indicating that the functional brain networks of the mild depression group deviate from small-world networks. These findings were consistent with previous neuroimaging studies using graph analysis to study depression (Leistedt et al., 2009). The lower clustering coefficient of the mild depression group implies that the local connectedness of networks in mild depression is relatively spared. In addition, short path lengths ensure effective interregional integrity or prompt information transmission in brain networks, which

TABLE 6 | The classification accuracy of the three-channel input form.

(A) The mean classification accuracy and standard deviation of the three-channel functional connectivity matrix obtained using the CNN.

	Accuracy % ± standard deviation %			
	Coherence	**Correlation**	**PLV**	**PLI**
Neu_block	**78.39** ± 2.62	**71.68** ± 2.51	71.54 ± 4.79	63.12 ± 3.10
Emo_block	**80.74** ± 1.48	**66.93** ± 3.06	65.80 ± 4.57	53.63 ± 2.48

(B) The highest classification accuracy obtained by inputting the three-band feature vector corresponding to each functional connectivity metric into the four classic classifiers (BN, LR, kNN, and RF) in the Neu_block and Emo_block.

	Accuracy %			
	Coherence	**Correlation**	**PLV**	**PLI**
Neu_block	57.92	51.81	49.72	54.03
Emo_block	54.72	46.81	53.47	51.39

*Bold indicates that the accuracy obtained based on the three-channel functional connectivity matrix is higher than that of single channel in **Table 4 (A)**.*

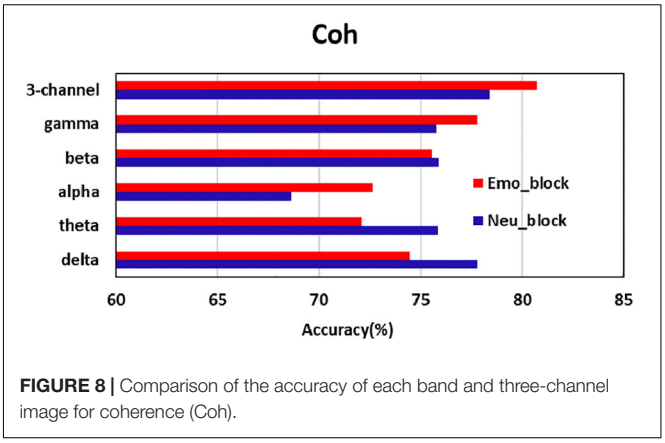


FIGURE 8 | Comparison of the accuracy of each band and three-channel image for coherence (Coh).

constitutes the basis of cognitive processes. Thus, the increase in the path length associated with the disease may be attributed to the degeneration of fiber bundles for information transmission (Bai et al., 2012).

In addition, the statistical analysis results of the clustering coefficients for each electrode showed that the difference between the two groups was mainly found in the parietal–occipital region of the right hemisphere. The study has shown that the right hemisphere is hyperactive in depression (Hecht, 2010). The parietal region and the right occipitotemporal cortex are the locations of the amygdala and hippocampus. Positron emission tomography studies have shown that the resting blood flow of the amygdala in patients with major depression increased by about 6% (Drevets et al., 1992). Jin et al. (2011) observed an abnormal hyperactive amygdala in depressed adolescents compared to healthy controls. In addition, research reported that the hippocampal volume of depressed patients was significantly reduced compared with healthy controls and the hippocampal neurons of these patients atrophy (Lee et al., 2002). Our findings are consistent with these previous results.

However, it is worth noting that, in the delta band, the clustering coefficient of the mild depression group was higher than that of the healthy control group in the frontal area. Some researchers have suggested that one function of prefrontal cortex is to modulate or inhibit amygdala activity (Davidson, 2004). Ochsner et al. (2002) reported a strong inverse relationship between activation of the prefrontal cortex and the amygdala when subjects were requested to voluntarily downregulate their negative affect. Our results and previous findings suggest abnormal frontal activity in mild depression.

Moreover, through the analysis of small-world properties and clustering coefficients, it is found that the difference between the two groups was mainly in the delta band. Knyazev (2012) reported that delta oscillations are prominent only in early human development stages and during slow-wave sleep. In waking adults, delta oscillations are overshadowed by more advanced processes associated with higher-frequency oscillations. However, delta oscillations are more pronounced in pathological states caused by detrimental environmental factors, developmental pathology, or damage to brain tissue, such as depression.

Coherence Recognizes Mild Depression Best

In the present study, we used different methods to construct functional connectivity matrices and compared the performance of four of these to identify cases of mild depression. Our results demonstrated that coherence was most effective in identifying mild depression using a CNN.

By using Eqs. (1) and (3), we determined that the PLV was the amplitude-normalized coherence. There are two perspectives on coherence and PLV. Researchers who support the use of the PLV often claim that the phase synchronization reflected by the PLV is more stringent than that for coherence because the latter confuses the consistency of the phase difference with amplitude correlation. This may be true from a mathematical point of view, but one might argue that when there are no amplitude correlations, it would be more "difficult" to get a meaningful non-zero coherence value in the absence of consistent phase differences. For example, when none of the separately observed cross-spectral density estimates has phase synchrony, even if there is perfect amplitude correlation, the expected value of the vector averages will be relatively small. On the other hand, if cross-spectral densities of all individuals are estimated to be strongly phase-synchronized, the expected value of their vector averages is still perceptible even in the absence of amplitude correlations. In addition, those in support of using the coherence also believe that in the case of coherence, the observations with large-amplitude products are given stronger weight, meaning they are favoring those observations that have a higher-quality phase difference estimate. This actually assumes that a higher amplitude reflects a higher signal-to-noise ratio of the source of interest, and thus a better-quality phase estimate (Bastos and Schoffelen, 2016).

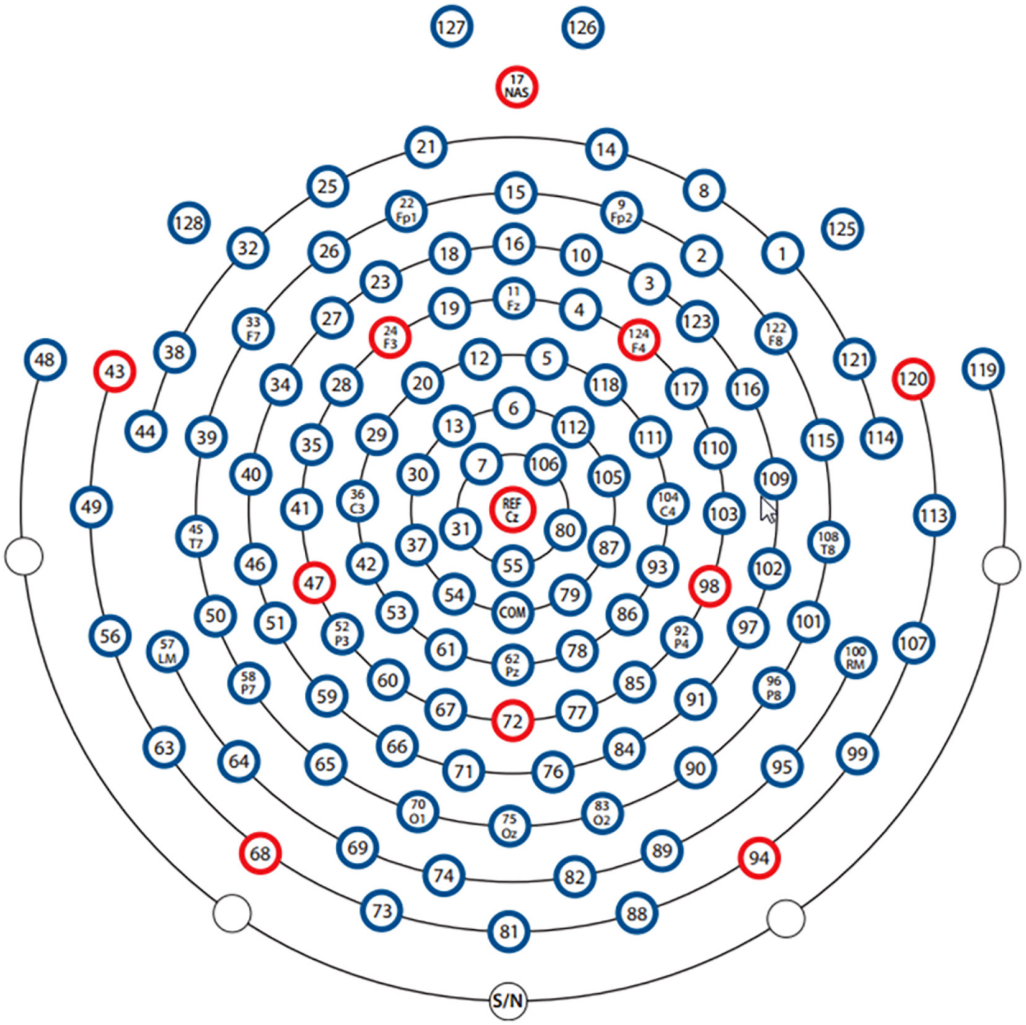


FIGURE 9 | The electrode distribution of 128-channel Geodesic Sensor Net.

A better-quality phase estimate may result in coherence achieving better classification performance than the PLV when using a CNN. However, the classification performance of the PLI in our CNN was relatively poor. This may be because the PLI is sensitive to the chosen length of the selected epochs (Fraschini et al., 2016), which was 6 s in the present study. This may have been too short to allow the PLI to stabilize.

The Three-Channel Functional Connectivity Matrix Improved Recognition Performance

Based on the data shown in **Tables 4A**, **6A**, it is clear that the three-channel functional connectivity matrix used herein improved the accuracy of coherence and correlation slightly, achieving an accuracy similar to that of single channels for the PLV and PLI. We suspect that human cognitive processes involve different EEG rhythms, making it reasonable that neural networks learn the information contained across several integrated EEG spectrums. However, this conjecture requires further, future verification using additional data sets.

CONCLUSION AND FUTURE WORK

In summary, the present study first illustrated that some abnormal organizations in the functional connectivity network of patients with depression also appeared in individuals with mild depression. Specifically, compared with healthy controls, the mild depression group has a larger characteristic path length and a lower clustering coefficient, indicating that the brain functional network of mild depression deviated from the small-world network. Second, we proposed a computer-aided method by which a CNN was used to learn information relevant to the functional connectivity matrices evident in individuals with mild depression such that they could be readily identified. This

is an innovative approach other than the existing graph theory for the use of functional connectivity matrices for depression recognition. We primarily considered functional connectivity matrices that reflect altered brain functional connectivity in patients with mental illnesses using a 2D data structure given the advantages of CNNs in processing 2D datasets. The classification results of our method showed that coherence, correlation, and the PLV can effectively recognize mild depression using a CNN and that the recognition performance of coherence was superior to the other functional connectivity metrics, obtaining a classification accuracy of 80.74%. The proposed method can provide an auxiliary diagnosis of mild depression and offers great promise. In the future, we are committed to implement this method as an online depression detection system. Once an individual's EEG signal is collected, it is used to determine whether the individual has mild depression, a disease that is not easily detectable and diagnosable. This approach may be used to improve and hasten the detection of individuals with mild depression, ultimately permitting quicker treatment.

While the present study offers significant benefits, it has some limitations that warrant discussion. First, in addition to the functional connectivity matrices used in the present study, there are other various connectivity metrics available, such as the imaginary part of coherency (Nolte et al., 2004), partial directed coherence (Baccalá and Sameshima, 2001), directed transfer function (Kamiński et al., 2001), phase slope index (Nolte et al., 2008), and Geweke's extension of Granger causality to the frequency domain (Brovelli et al., 2004). Further investigation of these metrics and their ability to detect the signatures of mental illnesses such as depression must continue in the future. Second, we used the functional connectivity metrics generated by all 128 pairs of electrodes available to us here. It is necessary to further explore which electrodes most obviously dictate functional connectivity matrix differences between healthy control individuals and those with mild depression. This would allow greater reductions to the number of electrodes used for classification and pave the way for real-time, online depression detection.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Lanzhou University Second Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and RL conceived the project idea. BH supervised the project. XZ provided critical suggestions for the statistical analysis. RL and YW performed the experiments and surveyed the literature, to check whether our findings were consistent with previous publications, and wrote the manuscript.

FUNDING

This work was supported by the National Basic Research Program of China (973 Program) (No. 2014CB744600), National Natural Science Foundation of China (Grant Nos. 61632014, 61210010, 61402211, 661300231, and 61627808), Fundamental Research Funds for the Central Universities (Grant Nos. lzujbky-2017-it74 and lzujbky-2017-it75), and the Program of Beijing Municipal Science and Technology Commission (Grant No. Z171100000117005).

REFERENCES

Ajilore, O., Vizueta, N., Walshaw, P., Zhan, L., Leow, A., and Altshuler, L. L. (2015). Connectome signatures of neurocognitive abnormalities in euthymic bipolar I disorder. *J. Psychiatr. Res.* 68, 37–44. doi: 10.1016/j.jpsychires.2015.05.017

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., et al. (2016). Theano: a python framework for fast computation of mathematical expressions. *arXiv* [Preprint]. arXiv:1605.02688.

Andrew, C., and Pfurtscheller, G. (1996). Event-related coherence as a tool for studying dynamic interaction of brain regions. *Electroencephalogr. Clin. Neurophysiol.* 98, 144–148. doi: 10.1016/0013-4694(95)00228-6

Baccalá, L. A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biol. Cybernet.* 84, 463–474. doi: 10.1007/pl00007990

Bai, F., Shu, N., Yuan, Y., Shi, Y., Yu, H., Wu, D., et al. (2012). Topologically convergent and divergent structural connectivity patterns between patients with remitted geriatric depression and amnestic mild cognitive impairment. *J. Neurosci.* 32, 4307–4318. doi: 10.1523/JNEUROSCI.5061-11.2012

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv* [Preprint]. arXiv: 1511.06448.

Bastos, A. M., and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. syst. Neurosci.* 9:175. doi: 10.3389/fnsys.2015.00175

Beck, A. T., Steer, R. A., Ball, R., and Ranieri, W. F. (1996). Comparison of beck depression inventories-IA and-II in psychiatric outpatients. *J. Personal. Assess.* 67, 588–597. doi: 10.1207/s15327752jpa6703_13

Bello-Cerezo, R., Bianconi, F., Di Maria, F., Napoletano, P., and Smeraldi, F. (2019). Comparative evaluation of hand-crafted image descriptors vs. off-the-shelf cnn-based features for colour texture classification under ideal and realistic conditions. *Appl. Sci.* 9:738. doi: 10.3390/app9040738

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9849–9854. doi: 10.1073/pnas.0308538101

Bruns, A. (2004). Fourier-, Hilbert-and wavelet-based signal analysis: are they really different approaches? *J. Neurosci. Methods* 137, 321–332. doi: 10.1016/j.jneumeth.2004.03.002

Cai, H., Han, J., Chen, Y., Sha, X., Wang, Z., Hu, B., et al. (2018). A pervasive approach to EEG-based depression detection. *Complexity* 2018, 1–3.

Cavanagh, J., and Geisler, M. W. (2006). Mood effects on the ERP processing of emotional intensity in faces: a P3 investigation with depressed students. *Int. J. Psychophysiol.* 60, 27–33. doi: 10.1016/j.ijpsycho.2005.04.005

Cecotti, H., and Graser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transact. Pattern Anal. Mach. Intell.* 33, 433–445. doi: 10.1109/TPAMI.2010.125

Dahl, G. E., Sainath, T. N., and Hinton, G. E. (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, BC: IEEE), 8609–8613.

Dalgleish, T. (2004). The emotional brain. *Nat. Reviews . Neurosci.* 5, 583–589.

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. *Biol. Psychol.* 67, 219–234. doi: 10.1016/j.biopsycho.2004.03.008

Davis, J., and Goadrich, M. (2006). "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning* (New York, NY: ACM), 233–240.

De Aguiar Neto, F. S., and Rosa, J. L. G. (2019). Depression biomarkers using non-invasive EEG: a review. *Neurosci. Biobehav. Rev.* 105, 83–93. doi: 10.1016/j.neubiorev.2019.07.021

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). "Decaf: a deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on Machine Learning* (New York, NY: ACM), 647–655.

Drevets, W. C., Videen, T. O., Price, J. L., Preskorn, S. H., Carmichael, S. T., and Raichle, M. E. (1992). A functional anatomical study of unipolar depression. *J. Neurosci.* 12, 3628–3641. doi: 10.1523/jneurosci.12-09-03628.1992

Ferree, T. C., Luu, P., Russell, G. S., and Tucker, D. M. (2001). Scalp electrode impedance, infection risk, and EEG data quality. *Clin. Neurophysiol.* 112, 536–544. doi: 10.1016/s1388-2457(00)00533-2

Fogel, J., Eaton, W., and Ford, D. (2006). Minor depression as a predictor of the first onset of major depressive disorder over a 15−year follow−up. *Acta Psychiatr. Scand.* 113, 36–43. doi: 10.1111/j.1600-0447.2005.00654.x

Fraschini, M., Demuru, M., Crobe, A., Marrosu, F., Stam, C. J., and Hillebrand, A. (2016). The effect of epoch length on estimated EEG functional connectivity and brain network organisation. *J. Neural Eng.* 13:036015. doi: 10.1088/1741-2560/13/3/036015

Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science* 326, 399–403. doi: 10.1126/science.1174521

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th international conference on artificial intelligence and statistics* (Sardinia: Chia Laguna Resort), 249–256.

Gong, X., Huang, Y.-X., Wang, Y., and Luo, Y.-J. (2011). Revision of the Chinese facial affective picture system. *Chinese Ment. Health J.* 25, 40–46.

Hecht, D. (2010). Depression and the hyperactive right-hemisphere. *Neurosci. Res.* 68, 77–87. doi: 10.1016/j.neures.2010.06.013

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., et al. (2015). "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1*, (Cambridge, MA: MIT Press), 1693–1701.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* [Preprint]. arXiv: 1207.0580.

Hosseinifard, B., Moradi, M. H., and Rostami, R. (2013). Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Comput. Methods Prog. Biomed.* 109, 339–345. doi: 10.1016/j.cmpb.2012.10.008

Hu, B., Majoe, D., Ratcliffe, M., Qi, Y., Zhao, Q., Peng, H., et al. (2011). EEG-based cognitive interfaces for ubiquitous applications: developments and challenges. *IEEE Intell. Syst.* 26, 46–53. doi: 10.1109/mis.2011.58

Hu, L., Xiao, M., Ai, M., Wang, W., Chen, J., Tan, Z., et al. (2019). Disruption of resting-state functional connectivity of right posterior insula in adolescents and young adults with major depressive disorder. *J. Affect. Disord.* 257, 23–30. doi: 10.1016/j.jad.2019.06.057

Humphries, M. D., Gurney, K., and Prescott, T. J. (2005). The brainstem reticular formation is a small-world, not scale-free, network. *Proc. R. Soc. B Biol. Sci.* 273, 503–511. doi: 10.1098/rspb.2005.3354

Jiao, Z., Gao, X., Wang, Y., Li, J., and Xu, H. (2018). Deep convolutional neural networks for mental load classification based on EEG data. *Pattern Recogn.* 76, 582–595. doi: 10.1007/s11517-019-01959-2

Jin, C., Gao, C., Chen, C., Ma, S., Netra, R., Wang, Y., et al. (2011). A preliminary study of the dysregulation of the resting networks in first-episode medication-naive adolescent depression. *Neurosci. Lett.* 503, 105–109. doi: 10.1016/j.neulet.2011.08.017

Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. *Biol. Cybernet.* 85, 145–157. doi: 10.1007/s004220000235

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, (Columbus, OH: IEEE), 1725–1732.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv* [Preprint]. arXiv: 1412.6980.

Knyazev, G. G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neurosci. Biobehav. Rev.* 36, 677–695. doi: 10.1016/j.neubiorev.2011.10.002

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 60, 1097–1105.

Lachaux, J. P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* 8, 194–208. doi: 10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., et al. (1990). "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, Vol. 2, ed. D. S. Touretzky (San Francisco, CA: Morgan Kaufmann), 396–404.

LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (Paris: *IEEE*), 253–256.

Lee, A. L., Ogle, W. O., and Sapolsky, R. M. (2002). Stress and depression: possible links to neuron death in the hippocampus. *Bip. Disord.* 4, 117–128. doi: 10.1034/j.1399-5618.2002.01144.x

Leistedt, S. J., Coumans, N., Dumont, M., Lanquart, J. P., Stam, C. J., and Linkowski, P. (2009). Altered sleep brain functional connectivity in acutely depressed patients. *Hum. Brain Mapp.* 30, 2207–2219. doi: 10.1002/hbm.20662

Li, X., Hu, B., Sun, S., and Cai, H. (2016a). EEG-based mild depressive detection using feature selection methods and classifiers. *Comput. Methods Prog. Biomed.* 136, 151–161. doi: 10.1016/j.cmpb.2016.08.010

Li, X., Jing, Z., Hu, B., and Sun, S. (2016b). "An EEG-based study on coherence and brain networks in mild depression cognitive process," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Shenzhen: IEEE), 1275–1282.

Li, X., Jing, Z., Hu, B., Zhu, J., Zhong, N., Li, M., et al. (2017). A resting-state brain functional network study in MDD based on minimum spanning tree analysis and the hierarchical clustering. *Complexity* 2017:9514369.

Li, Y., Cao, D., Wei, L., Tang, Y., and Wang, J. (2015). Abnormal functional connectivity of EEG gamma band in patients with depression during emotional face processing. *Clin. Neurophysiol.* 126, 2078–2089. doi: 10.1016/j.clinph.2014.12.026

Lyness, J. M., Heo, M., Datto, C. J., Ten Have, T. R., Katz, I. R., Drayer, R., et al. (2006). Outcomes of minor and subsyndromal depression among elderly patients in primary care settings. *Ann. Int. Med.* 144, 496–504.

Mormann, F., Lehnertz, K., David, P., and Elger, C. E. (2000). Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Phys. D Nonlin. Phenom.* 144, 358–369. doi: 10.1016/j.clinph.2013.09.047

Murias, M., Webb, S. J., Greenson, J., and Dawson, G. (2007). Resting state cortical connectivity reflected in EEG coherence in individuals with autism. *Biol. psychiatry* 62, 270–273. doi: 10.1016/j.biopsych.2006.11.012

Nanni, L., Brahnam, S., Salvatore, C., Castiglioni, I., and Initiative, A. S. D. N. (2019a). Texture descriptors and voxels for the early diagnosis of Alzheimer's disease. *Artif. Intell. Med.* 97, 19–26. doi: 10.1016/j.artmed.2019.05.003

Nanni, L., Brahnam, S., Ghidoni, S., and Maguolo, G. (2019b). General Purpose (GenP) bioimage ensemble of handcrafted and learned features with data augmentation. *arXiv* [Preprint]. arXiv:1904.08084.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin. Neurophysiol.* 115, 2292–2307. doi: 10.1016/j.clinph.2004.04.029

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., et al. (2008). Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* 100:234101.

Ochsner, K. N., Bunge, S. A., Gross, J. J., and Gabrieli, J. D. (2002). Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *J. Cogn. Neurosci.* 14, 1215–1229. doi: 10.1162/089892902760807212

Pfurtscheller, G., and Andrew, C. (1999). Event-related changes of band power and coherence: methodology and interpretation. *J. Clin. Neurophysiol.* 16, 512–519.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Rubinov, M., Knock, S. A., Stam, C. J., Micheloyannis, S., Harris, A. W., Williams, L. M., et al. (2009). Small−world properties of nonlinear brain activity in schizophrenia. *Hum. Brain Mapp.* 30, 403–416. doi: 10.1002/hbm.20517

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Rubinow, D. R., and Post, R. M. (1992). Impaired recognition of affect in facial expression in depressed patients. *Biol. Psychiatry* 31, 947–953. doi: 10.1016/0006-3223(92)90120-o

Spence, S. (1995). Descartes' error: emotion, reason and the human brain. *BMJ Br. Med. J.* 310:1213. doi: 10.1136/bmj.310.6988.1213

Stam, C. J., Jones, B., Nolte, G., Breakspear, M., and Scheltens, P. (2006). Small-world networks and functional connectivity in Alzheimer's disease. *Cereb. Cortex* 17, 92–99. doi: 10.1093/cercor/bhj127

Stam, C. J., Nolte, G., and Daffertshofer, A. (2007). Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. *Hum. Brain Mapp.* 28, 1178–1193. doi: 10.1002/hbm.20346

Strogatz, S. H. (2001). Exploring complex networks. *Nature* 410, 268–276. doi: 10.1038/35065725

Tabar, Y. R., and Halici, U. (2016). A novel deep learning approach for classification of EEG motor imagery signals. *J. Neural Eng.* 14:016003. doi: 10.1088/1741-2560/14/1/016003

Teramoto, H., Morita, A., Ninomiya, S., Akimoto, T., Shiota, H., and Kamei, S. (2016). Relation between resting state front-parietal EEG coherence and executive function in parkinson's disease. *Biomed Res. Int.* 2016: 2845754.

Ustun, T., and Chatterji, S. (2001). *Global Burden of Depressive Disorders and Future Projections. Depression: Social and Economic Timebomb*. London: BMJ Books.

Üstün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., and Murray, C. J. (2004). Global burden of depressive disorders in the year 2000. *Br. J. Psychiatry* 184, 386–392. doi: 10.1192/bjp.184.5.386

Volz, H.-P., and Laux, P. (2000). Potential treatment for subthreshold and mild depression: a comparison of St. John's wort extracts and fluoxetine. *Compr. Psychiatry* 41, 133–137. doi: 10.1016/s0010-440x(00)80019-x

Wada, Y., Nanbu, Y., Kikuchi, M., Koshino, Y., Hashimoto, T., and Yamaguchi, N. (1998). Abnormal functional connectivity in Alzheimer's disease: intrahemispheric EEG coherence during rest and photic stimulation. *Eur. Arch. Psychiatry Clin. Neurosci.* 248, 203–208. doi: 10.1007/s004060050038

Wang, L., Hermens, D., Hickie, I., and Lagopoulos, J. (2012). A systematic review of resting-state functional-MRI studies in major depression. *J. Affect. Disord.* 142, 6–12. doi: 10.1016/j.jad.2012.04.013

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature* 393, 440–442. doi: 10.1038/30918

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., and Cunningham, S. J. (1999). *Weka: Practical Machine Learning Tools and Techniques With Java Implementations*. Hamilton: University of Waikato, Department of Computer Science.

Zhang, J., Wang, J., Wu, Q., Kuang, W., Huang, X., He, Y., et al. (2011). Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder. *Biol. Psychiatry* 70, 334–342. doi: 10.1016/j.biopsych.2011.05.018

Zhang, M., Zhou, H., Liu, L., Feng, L., Yang, J., Wang, G., et al. (2018). Randomized EEG functional brain networks in major depressive disorders with greater resilience and lower rich-club coefficient. *Clin. Neurophysiol.* 129, 743–758. doi: 10.1016/j.clinph.2018.01.017

Zhang, X., and Lecun, Y. (2015). Text understanding from scratch. *arXiv* [preprint]. arXiv:1502.01710.

Zhong, X., Pu, W., and Yao, S. (2016). Functional alterations of fronto-limbic circuit and default mode network systems in first-episode, drug-naive patients with major depressive disorder: a meta-analysis of resting-state fMRI data. *J. Affect. Disord.* 206, 280–286. doi: 10.1016/j.jad.2016.09.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Li, La, Wang, Hu and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

The electrode distribution of 128-channel Geodesic Sensor Net (**Figure 9**).

Left frontal region (LF, number = 13): E32, E26, E23, E19, E12, E25, E22, E18, E21, E27, E24, E127, E128;

Right frontal region (RF, number = 13): E1, E2, E3, E4, E5, E8, E9, E10, E14, E123, E124, E126, E125;

Left temporal region (LT, number = 17): E33, E34, E38, E39, E40, E43, E44, E45, E46, E48, E49, E50, E51, E56, E57, E58, E63;

Right temporal region (RT, number = 17): E122, E116, E121, E115, E109, E120, E114, E108, E102, E119, E113, E101, E97, E107, E00, E96, E99;

Left central region (LC, number = 16): E7, E13, E20, E28, E29, E30, E31, E35, E36, E37, E41, E42, E47, E52, E53, E54;

Right central region (RC, number = 16): E106, E112, E118, E117, E111, E105, E80, E110, E104, E87, E103, E93, E98, E92, E86, E79;

Left parietal–occipital region (LPO, number = 13): E59, E60, E61, E64, E65, E66, E67, E68, E69, E70, E71, E73, E74;

Right parietal–occipital region (RPO, number = 13): E91, E85, E78, E95, E90, E84, E77, E94, E89, E83, E76, E88, E82.

frontiers in Neuroscience

METHODS
published: 15 April 2020
doi: 10.3389/fnins.2020.00221

A Constrained ICA-EMD Model for Group Level fMRI Analysis

Simon Wein[1,2]*, *Ana M. Tomé*[3], *Markus Goldhacker*[1,2], *Mark W. Greenlee*[2] and *Elmar W. Lang*[1]

[1] CIML, Biophysics, University of Regensburg, Regensburg, Germany, [2] Experimental Psychology, University of Regensburg, Regensburg, Germany, [3] IEETA/DETI, Universidade de Aveiro, Aveiro, Portugal

Independent component analysis (ICA), being a data-driven method, has been shown to be a powerful tool for functional magnetic resonance imaging (fMRI) data analysis. One drawback of this multivariate approach is that it is not, in general, compatible with the analysis of group data. Various techniques have been proposed to overcome this limitation of ICA. In this paper, a novel ICA-based workflow for extracting resting-state networks from fMRI group studies is proposed. An empirical mode decomposition (EMD) is used, in a data-driven manner, to generate reference signals that can be incorporated into a constrained version of ICA (cICA), thereby eliminating the inherent ambiguities of ICA. The results of the proposed workflow are then compared to those obtained by a widely used group ICA approach for fMRI analysis. In this study, we demonstrate that intrinsic modes, extracted by EMD, are suitable to serve as references for cICA. This approach yields typical resting-state patterns that are consistent over subjects. By introducing these reference signals into the ICA, our processing pipeline yields comparable activity patterns across subjects in a mathematically transparent manner. Our approach provides a user-friendly tool to adjust the trade-off between a high similarity across subjects and preserving individual subject features of the independent components.

Keywords: independent component analysis, ICA, empirical mode decomposition, EMD, Green's-function - based EMD, fMRI

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Shijie Zhao,
Northwestern Polytechnical University, China
Vesa J. Kiviniemi,
University of Oulu, Finland

***Correspondence:**
Simon Wein
simon.wein@psychologie.uni-regensburg.de

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *22 August 2019*
Accepted: *28 February 2020*
Published: *15 April 2020*

Citation:
Wein S, Tomé AM, Goldhacker M, Greenlee MW and Lang EW (2020) A Constrained ICA-EMD Model for Group Level fMRI Analysis. Front. Neurosci. 14:221. doi: 10.3389/fnins.2020.00221

1. INTRODUCTION

Independent component analysis (ICA) is a data-driven tool that is widely employed for functional magnetic resonance imaging (fMRI) data analysis. Based on a linear mixture model, either spatially (McKeown et al., 1998) or temporally (Biswal and Ulmer, 1999) independent components (ICs) can be obtained with ICA, without the requirement of prior information about anatomical regions of interest or temporal activation profiles. One problem of ICA is that, because of inherent indeterminacies, in general, it is not suitable for group studies. Different subjects have different time courses and spatial maps, and the extracted components will be sorted differently. This can make it difficult to find comparable activation patterns between subjects and draw inferences from subject groups. So far, various approaches have been proposed to overcome these shortcomings of ICA (Calhoun et al., 2009). Combining components obtained by single-subject ICA based on spatial correlation or clustering was proposed by Calhoun et al. (2001a) and Esposito et al. (2005). Another possibility is the spatial or temporal concatenation of the individual datasets to obtain components in a single ICA step from a group dataset, as well as the employment of back-reconstruction

approaches to obtain subject-specific components (Calhoun et al., 2001b; Svensén et al., 2002). These concatenation-based approaches were compared with a simple across-subject averaging by Schmithorst and Holland (2004). In a sophisticated approach Beckmann and Smith (2005) proposed a tensorial extension of ICA. The authors extended the probabilistic ICA (PICA) model by adding a third dimension representing subject-related dependencies in addition to the spatio-temporal dimensions. The model represents a three-way factor analysis similar to the well-known PARAFAC model (Harshman and Lundy, 1994).

A popular paradigm used to acquire data for the above-mentioned exploratory matrix factorization techniques is the so-called resting-state. Under this paradigm, subjects either rest with eyes open fixating a fixation cross or with eyes closed. Usually, subjects are instructed not to fall asleep and to let their mind wander. Contrary to the simplicity of the paradigm, the generated database has a complex spatial structure and temporal dynamics, which arise from low-frequency fluctuations in the BOLD signal (Biswal et al., 1995; Fox and Raichle, 2007). Furthermore, the data are characterized by large number of spatial dimensions and they lack the temporal structure usually found for task-based fMRI investigations. Because of these two aspects, exploratory matrix factorization techniques are appropriate to analyze the large amount of data and to explore the complex spatial and temporal structure in the data, as shown by Kiviniemi et al. (2003). In this context, ICA-based pipelines have emerged as a state-of-the-art approach to investigate rs-fMRI data (Allen et al., 2011; Remes et al., 2011). This decomposition of resting-state fMRI data results in a so-called parcellation of the cortex into brain networks composed of functionally connected brain areas. In the literature, common brain networks are default-mode, cognitive control, visual, somatomotor, sub-cortical, auditory, or cerebellar, depending on the function of the brain areas included in each network (Allen et al., 2014), which can be successfully extracted from the data with ICA (Beckmann et al., 2005).

In this paper, a hybrid method is proposed for extracting resting-state networks (RSNs) from fMRI data based on constrained ICA (cICA) and empirical mode decomposition (EMD). This constrained extension of ICA optimizes the statistical independence and additionally the similarity to a given reference signal. In the framework of an augmented Lagrangian approach, the incorporation of a reference into ICA helps more robust ICs to be obtained while eliminating the ambiguities of the ICA approach (Lu and Rajapakse, 2006; Lin et al., 2009; Rodriguez et al., 2014). In this paper, like (Lin et al., 2009), spatial reference maps were employed to extract resting-state networks from fMRI data. Besides analyzing temporal time series (Huang et al., 1998), the EMD framework can be extended for the analysis of two-dimensional spatial maps (Nunes et al., 2003; Al-Baddai et al., 2016a,b), and in this study, we focus on the latter variant. Based on a preliminary study Wein et al. (2019), it is shown that an EMD-based image decomposition technique, denoted as Green's function in tension based bi-dimensional ensemble EMD (GiT-BEEMD) (Al-Baddai et al., 2016b), produces suitable references for cICA. This two-dimensional variant of EMD allows us to decompose images into so-called bi-dimensional intrinsic mode functions (BIMFs) and can also be used to slice-wise decompose volumetric fMRI images. Because of its inherent natural ordering of the extracted intrinsic modes according to their spatial frequencies, EMD can easily generate prototypical spatial maps. Similar spatial maps obtained with the EMD for each subject can be identified and averaged across subjects. In the next step, these prototypical spatial maps can serve as reference signals for a constrained ICA applied in parallel to the entire group of subjects. In this workflow, the references are obtained from the same dataset as used for the analysis, so no prior information is required. We extend previous constrained ICA methods (Lu and Rajapakse, 2006; Lin et al., 2009; Rodriguez et al., 2014) by showing that intrinsic mode functions generated by EMD are suitable references which help to extract resting-state networks in a purely data-driven fashion. The proposed workflow intrinsically adapts to the statistics of the given data, thereby avoiding any bias toward external references. We compare our new method to another data-driven ICA approach, an established group ICA method, based on temporal concatenation (Calhoun et al., 2001b), using a resting-state fMRI dataset from the *Human Connectome Project* (Essen et al., 2012). The potential benefits of this hybrid cICA-EMD method are emphasized by showing that this approach allows the user to actively shape the extracted resting-state networks. The trade-off between enforcing a certain similarity across subjects and preserving individual subject features can be determined and can be adapted to optimally fulfill the requirements of different studies.

2. MATERIALS AND METHODS

The following subsections introduce the dataset employed and describe the data analysis techniques, which combine cICA and GiT-BEEMD, as well as the processing steps of the gICA approach used for comparison. Also, a flowchart of the proposed signal processing chain is provided.

2.1. Dataset

This study employed a data set from the *Human Connectome Project* (Essen et al., 2012). The S1200 release includes data from subjects who participated in four resting-state fMRI sessions, lasting 14.4 min each and resulting in 1200 volumes per session. Customized Siemens *Connectome Skyra* magnetic resonance imaging (MRI) scanners at Washington University with a field strength of $B_0 = 3$ Tesla were employed for data acquisition, using a multi-band (factor 8) technique (Feinberg et al., 2010; Moeller et al., 2010; Setsompop et al., 2012; Xu et al., 2012). The data were collected by gradient-echo echo-planar imaging (EPI) sequences with a repetition time $TR = 720ms$ and an echo time $TE = 31.1\ ms$, using a flip angle of $\theta = 52^{\circ}$. The field of view was $FOV = 208\ mm \times 180\ mm$, and $N_s = 72$ slices with a thickness of $d_s = 2\ mm$ were obtained, containing voxels with a size of $2\ mm \times 2\ mm \times 2\ mm$. The preprocessed version, which had been subjected to motion-correction, structural preprocessing, and ICA-FIX denoising, was chosen (Jenkinson et al., 2002, 2012; Fischl, 2012; Glasser et al., 2013; Smith et al., 2013; Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). In the S1200 release, the FIX-classifier was trained on a labeled subset of the provided

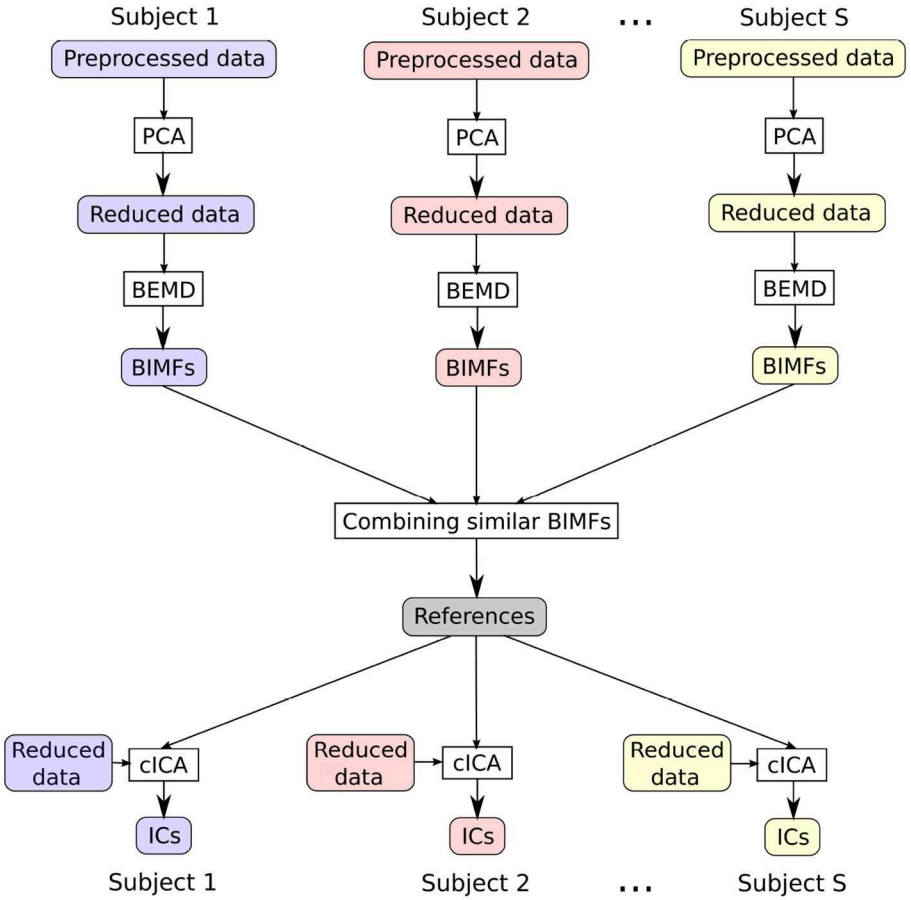


FIGURE 1 | The flowchart sketches the main steps of the presented approach: first, reducing the data with PCA, then extracting BIMFs with spatial BEMD from the reduced data, and then combining the BIMFs of each subject in order to get shared references for cICA, which finally help to obtain comparable ICs across subjects.

data (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). For comparison of the two approaches, 10 sessions from 10 different subjects were selected from the database. Gaussian smoothing with a half-width $FWHM = 5\ mm$ was then applied by using the *SPM12* software package[1], and the first five images were discarded to account for magnetic saturation effects.

2.2. A Hybrid cICA-EMD Approach

In this section a new approach to deal with an ICA analysis across a group of subjects will be described. The flowchart in **Figure 1** presents an overview of the various steps of the data analysis. All processing steps were performed in *MATLAB 9.3 Release 2017b*.

2.2.1. Preprocessing

The data as obtained from the data repository will be further pre-processed, as explained in the following.

- In the first step, the voxel time series are linearly detrended and the voxel values are transformed to have zero mean and unit variance.
- Next a mask common to all subjects is used to exclude voxels that are located outside of the brain. The mask is created by employing the *GIFT* toolbox[2], using the "Generate mask" option.
- The data of subject s is stored in a $K \times L$-dimensional matrix $\mathbf{X}^{(s)}$ containing in its columns $\mathbf{x}_l(k)$ the temporal evolution of L brain voxels at K time points.
- Principal component analysis (PCA) related dimension reduction is then performed based on the singular value decomposition (SVD) of $\mathbf{X}^{(s)} = \mathbf{U}^{(s)}\mathbf{\Sigma}^{(s)}\mathbf{V}^{(s)T}$. If the row mean of $\mathbf{X}^{(s)}$ is removed, then $\mathbf{U}^{(s)}$ contains the eigenvectors of the covariance matrix $\mathbf{C}^{(s)} \propto \mathbf{X}^{(s)}\mathbf{X}^{(s)T}$ in its columns. The eigenvectors with the largest eigenvalues indicate the directions of greatest variance, which are denoted as principal components (PCs) (Jolliffe, 2014).
- Next, the fMRI images of each subject are projected onto the first $M = 20$ PCs $\mathbf{X}_M^{(s)} = (\mathbf{U}_M^{(s)})^T\mathbf{X}^{(s)}$. This reduces the number of images per session to $M = 20 < K$.
- Finally, from the reduced data sets, the image slices are reconstructed to enter into the GiT-BEEMD analysis, while the

[1]https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[2]http://mialab.mrn.org/software/gift/

reduced data $\mathbf{X}_M^{(s)}$ enters directly into the cICA processing path. The number of selected principal components M determines the number of sources, estimated in the cICA step. A relatively low order of $M = 20$ was chosen, to obtain robustly observed, large-scale resting-state networks (van den Heuvel and Pol, 2010) and to make it easy to identify extracted networks, which are suitable for comparison of results in section 3.

2.2.2. Green's Function-Based Ensemble Empirical Mode Decomposition

For the next step, the extraction of suitable reference signals for cICA, the GiT-BEEMD technique was employed (Al-Baddai et al., 2016b) to extract intrinsic activity patterns from spatial volumetric fMRI images. The idea of this technique is to decompose a two-dimensional brain slice $I(\mathbf{r}) = I(x, y)$ into bi-dimensional intrinsic mode functions (BIMFs):

$$I(x, y) = \sum_{j=1}^{J} b_j(x, y) \tag{1}$$

Here $b_j(x, y)$ denotes the j-th BIMF, which is estimated iteratively as described in **Appendix 1**, and, in our notation, we include the residuum $r(x, y)$ as intrinsic mode $b_J(x, y)$. The first extracted BIMF contains the highest spatial frequency, which will decrease in every additionally extracted BIMF (Al-Baddai et al., 2016b).

Each brain slice was decomposed into five intrinsic modes and one residuum by repeating the sifting step five times. The ensemble step was repeated only twice, whereby noise was either added or subtracted from the data once at each step. The assisting noise was generated with a noise amplitude of $a_\eta = 0.2$. The tension parameter was initialized to $T_1 = 0.9$ and reduced after the extraction of the j-th BIMF b_j to $T_{j+1} = T_j - \frac{1}{J}$. This avoids blob-like artifacts in low frequency modes if the tension parameter is set too high (Al-Baddai et al., 2016b). An example of a decomposition is provided in **Figure 2**. These intrinsic modes represent characteristic spatial textures of the activity distribution in the brain. BIMFs with high spatial frequencies (see BIMF 1 and BIMF 2, for example) show highly localized spatial activation patterns that, however, are spread all over the brain slice, while BIMFs with lower spatial frequencies (see BIMF 4 and BIMF 5, for example) concentrate the activity in few highly localized areas in the brain. In this example, the residuum reveals focused activity in the temporal brain network. Note that this reflects the high activity in this area seen in the original activity distribution of the chosen brain slice. A combination of lowest spatial frequency intrinsic mode plus the residuum, i. e., $b_5(x, y) + b_6(x, y) \equiv b_{56}(x, y)$, was used as reference for cICA in the further evaluations. The analysis in section 3 reveals that low-frequency modes are especially suitable to serve as references in order to obtain consistent resting-state networks across subjects. By decomposing the activity patterns and pointing toward specific brain areas, these intrinsic modes can help the ICA to converge to these specific areas for all subjects. **Figure 4** shows that, for the most demanding similarity constraint, the most consistent results were achieved by combining BIMF 5 and the residuum to reference networks. This combination is also depicted in **Figure 2**. From all the decomposed two-dimensional slices, the corresponding modes $b_{56}(x, y)$ were organized into a three-dimensional data array, which then was concatenated into a 3D volume intrinsic mode function (VIMF). Decomposing the $M = 20$ brain volumes per subject in PC subspace results in M VIMFs per subject. For the next processing steps, the voxels inside of the brain are sorted into an $M \times L$ matrix again, denoted as $\mathbf{V}_M^{(s)}$.

In order to extract from each subject RSNs that are consistent across the proband cohort, corresponding intrinsic modes need to be identified. As part of the minimal preprocessing pipeline

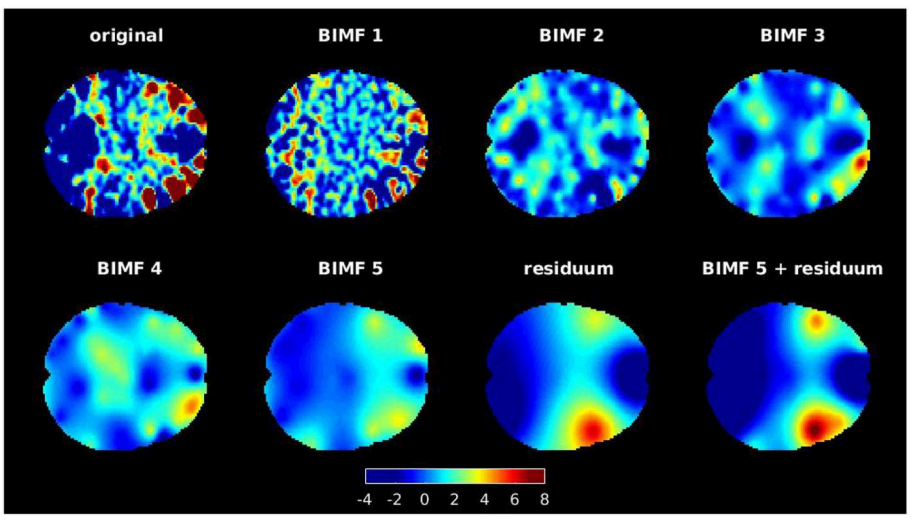


FIGURE 2 | An example of decomposing a brain slice with GiT-BEEMD in the transverse anatomical plane. The slice was decomposed into five intrinsic modes and one residuum. Note that BIMF 1 contains the highest spatial frequencies, which gradually decrease for the other extracted BIMFs. Also, the combination of the fifth BIMF and residuum is illustrated, as this combination is used for further evaluations. A mask was applied after the decomposition to set all intensity values that were located outside of the scanned brain to zero.

of the *Human Connectome Project* (Glasser et al., 2013), the fully preprocessed fMRI data were transformed to the MNI standard space for every subject, and therefore a good correspondence of activity patterns between the healthy subjects was given. Note that if two different subjects cohorts were to be analyzed, the reference networks should be formed individually for each group, because in patients with neurological diseases, resting-state activity could fundamentally differ from that of healthy controls. In our work, studying only healthy probands, averaging most similar modes between subjects yields proper common reference signals for all subjects that have been employed in a cICA of fMRI datasets. A visual inspection of the VIMFs of any two randomly chosen subjects showed that corresponding spatial patterns might occur in different rows of $\mathbf{V}_M^{(s)}$. Hence, the extracted VIMFs first need to be ordered according to their similarity between subjects. An efficient way to assign similar VIMFs between subjects is offered by the assignment algorithm proposed by Munkres (1957) based on the Hungarian method developed by Kuhn (1955). After computing the proper correspondences between the VIMFs of all subjects, the references $\mathbf{R}$ to be used in the cICA algorithm are then obtained by averaging the corresponding VIMFs across all subjects. To summarize, the reference signals are computed as follows:

1. Initialize the reference as $\mathbf{R} = \mathbf{V}_M^{(1)}$
2. For $s = 2, \ldots, S$, do:
   a. Apply the Hungarian algorithm and re-order the rows of $\mathbf{V}_M^{(s)} \rightarrow \tilde{\mathbf{V}}_M^{(s)}$
   b. Update the reference $\mathbf{R} \leftarrow \frac{s-1}{s}\mathbf{R} + \frac{1}{s}\tilde{\mathbf{V}}_M^{(s)}$

To apply the Hungarian algorithm, a cost function $1 - \rho(\mathbf{R}, \mathbf{V}_M^{(s)})$ is defined that, if optimized, results in an ordering of the rows in $\mathbf{V}_M^{(s)}$ such that the sum of the correlation coefficients between pairs of rows in the two matrices is maximized. Note that the algorithm achieves the re-ordering without calculating all M possible assignments. Finally, each row of $\mathbf{R}$ is normalized, having entries with zero mean and unit variance, and $M = 20$ references for the cICA algorithm are obtained.

2.2.3. Constrained ICA

Sources $\mathbf{Y}_M = [\mathbf{y}_1, \ldots, \mathbf{y}_L], \mathbf{y} \in \mathbb{R}^M$ can be blindly estimated from the mixtures $\mathbf{X}_M = [\mathbf{x}_1, \ldots, \mathbf{x}_L], \mathbf{x} \in \mathbb{R}^M$ according to:

$$\mathbf{Y}_M = \mathbf{W}\mathbf{X}_M \tag{2}$$

with the demixing matrix defined as $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]^T$, where $\mathbf{w}_m^T$ are the rows of the demixing matrix and $\mathbf{X}_M$ collects all L samples of the projected data after being spatially transformed to zero mean and unit variance.

Finding a demixing matrix is solved by designing an optimization problem where inequality and equality constraints are integrated in an augmented Lagrangian formulation. The inequality constraint terms in the Lagrange function are re-written as equality constraints with the help of a slack variable (Lu and Rajapakse, 2006). After finding the optimal value of these slack variables, the modified version of the augmented Lagrangian function is written as

$$\mathcal{L}(\mathbf{W}, \boldsymbol{\mu}) = J(\mathbf{W}) + \sum_{m=1}^{M} \frac{1}{2\gamma_m} \left[(\max\{0, \gamma_m h_m(\mathbf{w}_m^T) + \mu_m\})^2 - \mu_m^2 \right] \tag{3}$$

where μ_m are the Lagrangian parameters, while γ_m represents a user-defined penalty. The first term $J(\mathbf{W})$ reflects the cost function of ICA, and the second term in the Lagrangian is related with the inequality constraint, which compares the m-th extracted component with the corresponding reference signal:

$$h(\mathbf{w}_m^T) = \varsigma_m - \epsilon(\mathbf{w}_m^T\mathbf{x}, r_m) \leq 0 \tag{4}$$

where $\epsilon(\cdot)$ is a similarity measure and ς_m is a threshold parameter. Similarity is conventionally expressed either through a correlation measure $\mathbb{E}\{y_m r_m\}$ or the mean squared error $\mathbb{E}\{(y_m - r_m)^2\}$, with $y_m = \mathbf{w}_m^T\mathbf{x}$. The expected value is approximated by an average over the available data.

Estimating the demixing matrix $\mathbf{W}$, given the constraint introduced above, can be achieved in different ways, based either on negentropy-like cost functions (points 1–3) or on a maximum likelihood estimate (point 4):

1. Simply one IC, most similar to the given reference signal, can be extracted. This approach can easily be extended to a multi-reference cICA. However, this additionally requires a decorrelation of the weights during each iteration to prevent different weights from converging to identical estimations (Lu and Rajapakse, 2006).
2. Lu and Rajapakse (2006) introduced an objective function for cICA that contained an additional equality constraint to bound the weights. Later, a simplification was introduced by Lin et al. (2007) where equality constraints were omitted; rather, the weight vectors were normalized at each iteration instead.
3. Also, cICA based on fixpoint learning (Lin et al., 2009) was proposed, which should overcome the limitations of the second-order Newton-like learning used in the cICA algorithm of Lu and Rajapakse (2006).
4. Finally, yet another version, using a cost function $J(\mathbf{W})$ based on a maximum likelihood estimate has been proposed (Rodriguez et al., 2014) according to

$$J(\mathbf{W}) \approx \mathbb{E}\left\{ \sum_{m=1}^{M} \log(p(\mathbf{w}_m^T\mathbf{x})) \right\} + \log|\det(\mathbf{W})| \tag{5}$$

An iterative procedure is then derived to update the parameters $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_M]^T$ and the de-mixing matrix $\mathbf{W}$. Thereby, a decoupling scheme based on a Gram-Schmidt orthogonalization is proposed, finally yielding the following objective function:

$$J(\mathbf{w}_m) \propto \mathbb{E}\left\{ \log(p(\mathbf{w}_m^T\mathbf{x})) \right\} + \log|\mathbf{d}_m^T\mathbf{w}_m| \tag{6}$$

where the decoupling vector $\mathbf{d}_m \in \mathbb{R}^{M\times 1}$ is defined through $\tilde{\mathbf{W}}_m\mathbf{d}_m = 0$ and where $\tilde{\mathbf{W}}_m \in \mathbb{R}^{(M-1)\times M}$ denotes the de-mixing matrix without entries to the m-th row.

It has been shown by Cardoso (1997) that a maximum likelihood approach to ICA is equivalent to the seminal Infomax approach put forward by Bell and Sejnowski (1995). Thus, in analogy to the maximum likelihood approach, a constrained and decoupled version of the extended Infomax algorithm can be obtained (Rodriguez et al., 2014). The extended Infomax algorithm is often used in an analysis of fMRI data (Correa et al., 2007) and was also used in this study as the basis of the cICA. A more detailed description of this algorithm and a proper metacode are given in **Appendix 2** for the convenience of the reader.

The data, projected onto the first $M = 20$ PCs, and the M references, transformed to zero mean and unit variance, together enter the cICA algorithm to finally extract $M = 20$ ICs. The weights are initialized with small random values, and the learning rate for the weights is set to $\eta = 0.5$. The scalar penalty can be set to 3 (Rodriguez et al., 2014). The influence of the references can be well determined by adjusting the threshold parameter. Therefore different settings have been studied using the correlations $\mathbb{E}\{y_m r_m\}$ as distance measures, and the results are presented in section 3.

2.3. Group ICA

Generally, fMRI data are compared across a group of subjects by employing the gICA algorithm put forward by Calhoun and his group (Calhoun et al., 2001a). This gICA is made available in the *GIFT* toolbox[3] and was incorporated in this study for comparison. Voxel time series were preprocessed by variance normalization through linearly detrending and transforming the data to zero mean and unit variance. The single-subject data matrices $\mathbf{X}^{(s)}_{K\times L}$ enter the first PCA step, with the temporal evolution of L brain voxels at K time points in columns. The subject datasets were then projected onto the first $M' = 1.5 \cdot M = 30$ PCs in this step by applying an SVD to the data matrix. This follows the recommendation of the GIFT toolbox, projecting the data onto 1.5 times the components used in the group reduction step. Note that in the latter step, the number of projections was chosen to be $M = 20$. The S reduced $M' \times L$ matrices $\mathbf{X}^{(s)}_{M'}$ on subject level were concatenated to an $(S \cdot M') \times L$ group matrix $\mathbf{X}^{(g)}_{S\cdot M'}$ entering the second PCA step. The group matrix is projected onto $M = 20$ PCs, resulting in a reduced $M \times L$ matrix $\mathbf{X}^{(g)}_{M}$. The 20 group spatial maps $\mathbf{S}_M$ are extracted from $\mathbf{X}^{(g)}_{M} = \mathbf{A}\mathbf{S}_M$ by the extended Infomax algorithm (Lee et al., 1999) and by additionally employing the ICASSO option (Himberg et al., 2004), running the ICA algorithm ten times with different initializations to assure greater stability. Finally $M = 20$ subject-specific spatial $\mathbf{S}^{(s)}_M$ maps were obtained by the GICA3 (Erhardt et al., 2011) back-reconstruction approach. We compared these 20 brain networks, with the 20 networks we obtained from our approach. For visualization purposes, M mean networks $\langle \mathbf{S}_{m*} \rangle$, $m = 1, \ldots, M$ were obtained by averaging $\mathbf{S}^{(s)}_M$ over the subjects, i. e., $\langle \mathbf{S}_M \rangle = \frac{1}{S}\sum_{s=1}^{S} \mathbf{S}^{(s)}_M$.

[3]http://mialab.mrn.org/software/gift/

3. RESULTS

The goal of the study was to compare RSNs obtained with the newly proposed cICA-EMD approach as opposed to RSNs resulting from the conventional gICA approach. RSNs denote functionally connected brain areas that, however, are anatomically separated but maintain a high level of activity in a resting state of the proband. They are represented in this study by the ICs extracted with the discussed techniques. In this study, $M = 20$ ICs were extracted with either method. Comparable RSNs obtained by the different approaches were identified by visual inspection and are depicted in **Figure 3**. There, references used for cICA are shown in the first row, while in the second row, the ICs obtained therewith are presented, computed as mean ICs over subjects. In the third row of **Figure 3**, the mean ICs obtained by gICA are exhibited. The significance of the resulting ICs was tested with a one-sample student's T-test by employing the *SPM12* software package[4]. The resulting spatial maps of t-values are depicted in the fourth and fifth row in **Figure 3**. Spatial maps were thresholded at a significance level of $p < 0.001$ ($t = 4.30$, $df = 9$).

Most prominent brain areas are in IC 10/6 (obtained by cICA-EMD/gICA), the left inferior parietal lobule, in IC 17/8, the right angular/supramarginal gyrus, in IC 4/11, the superior occipital gyrus, in IC 9/13, the right inferior frontal gyrus, in IC 20/18, the anterior cingulate cortex, in IC 6/1, the precentral gyrus, in IC 2/9, the paracentral lobule, in IC 8/2, the middle occipital gyrus, and in IC 3/7, the middle temporal gyrus. The independent networks obtained represent well-observed RSNs (van den Heuvel and Pol, 2010) and can be further grouped based on their functions. The corresponding attentional and default mode networks are depicted in the first five columns, while the extracted auditory and sensorimotor networks are shown in the sixth and seventh column. Next, in the eighth and ninth column, visual networks are represented, and, in the last column, the cerebellum is shown. The similarity threshold for cICA was set to $\varsigma = 0.5$ in this example. All resting-state networks obtained with both approaches, including the employed references, are provided in the **Supplementary Material**.

The motivation of different group ICA approaches is to make this explorative analysis technique suitable for studies where it is necessary to compare extracted networks between different subjects. This means that issues with permutation indeterminacy and reproducibility of ICA have to be overcome to obtain well-comparable networks. Therefore, a measure of interest for the evaluation of the two different approaches could be the consistency of activation patterns across subjects. This consistency was quantified by measuring how much a resting-state network $(\mathbf{y}^{(s)})^T_m \in \mathbb{R}^L$ from one subject s differs on average from the mean network $\langle \mathbf{y} \rangle^T_m = \frac{1}{S}\sum_{s=1}^{S}(\mathbf{y}^{(s)})^T_m$ across subjects. Pearson's correlation $\rho\left((\mathbf{y}^{(s)})^T_m, \langle \mathbf{y} \rangle_m\right)$ was used to measure the correlation between standardized subject networks $\mathbf{y}^{(s)}_m$ and

[4]https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

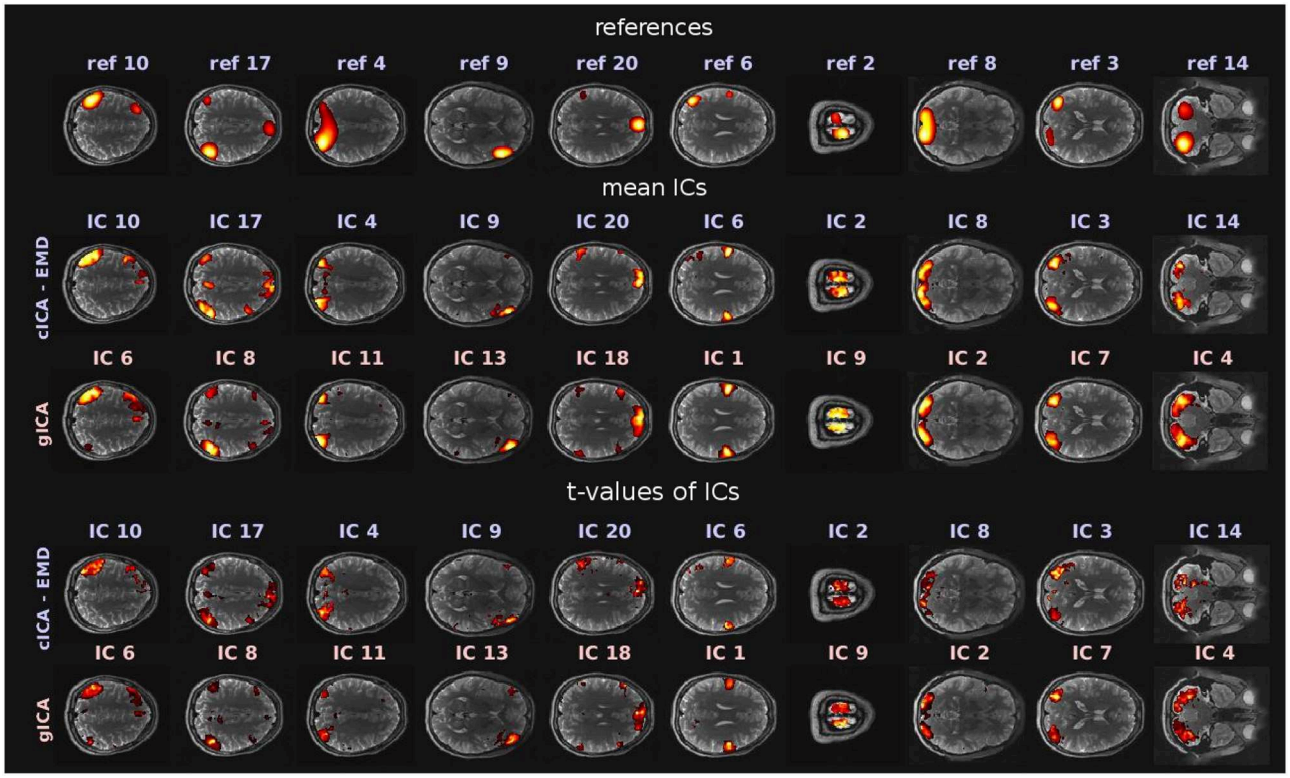


FIGURE 3 | RSNs obtained by the two different approaches. The first row shows the references used for cICA, while the second row exhibits mean ICs, averaged over the subject cohort and computed with the newly proposed cICA-EMD method. These ICs are contrasted in the third row with mean ICs obtained by gICA. In the fourth and fifth row of this figure, *t*-values of the RSNs can be found. All depicted slices were chosen such that they intersect the peak activation voxel of the corresponding ICs obtained by gICA. For visualization purposes, the activations of the networks shown in the first three rows are normalized to zero mean and unit variance. Furthermore, in accordance with common usage, the voxel intensities $\hat{I}(\mathbf{r})$ were thresholded by $\hat{I}(\mathbf{r}) > 2$, and in the sixth and ninth column, the threshold was adjusted to $\hat{I}(\mathbf{r}) > 1.5$ for better recognizability of the networks. The color range of the heatmap was adjusted to the largest intensity value in every pictured slice.

the related mean networks $\langle \mathbf{y} \rangle_m^T$. The following consistency measure is used:

$$K_{per}(\mathbf{y}_m) = \frac{1}{S} \sum_{s=1}^{S} \rho\left((\mathbf{y}^{(s)})_m^T, \langle \mathbf{y} \rangle_m \right) \tag{7}$$

In order to find the most suitable references, the consistency measure was evaluated for different BIMFs or combinations of them. Following the process described in section 2.2.2, the GiT-BEEMD algorithm was used to decompose the activity patterns of the fMRI data into BIMFs, reflecting intrinsic patterns on different frequency scales.

Figure 4 depicts the consistency measure, as defined in Equation 7, with BIMF 1 to 6 used as references and its dependence on similarity threshold ζ. Extracted BIMFs were also gradually summed up again, and for the most demanding threshold, $\zeta = 0.7$, a combination of BIMF 5 + BIMF 6 (the residuum) yielded the best results. This combination was used for further evaluations of our approach, and for the comparison with gICA (Calhoun et al., 2001b). For the comparison, the consistency measure was computed for all ICs obtained with the cICA-EMD approach and again for different settings of the similarity threshold ς. An equivalent procedure was followed using the results from applying the gICA algorithm. The results are listed in **Table 1**.

By adjusting the threshold parameter ς, it is possible to well determine the influence of the constraint during the optimization, so choosing a smaller threshold allows for more variability in the estimated components across subjects. Increasing the threshold increases the similarity between subject-specific components and common references. This means that if the similarity between every subject component and the shared reference increases, the similarity of components across subjects will also increase, which is quantified by the consistency measure in Equation 7. **Figure 5** illustrates the behavior of the consistency at different similarity thresholds ς in comparison to gICA. If the threshold parameter is set to a value of $\varsigma = 0.40$, the consistency is lower than that of gICA. By further increasing this threshold to a value of $\varsigma = 0.60$, the consistency of estimated resting-state networks with the proposed approach starts to exceed that of gICA.

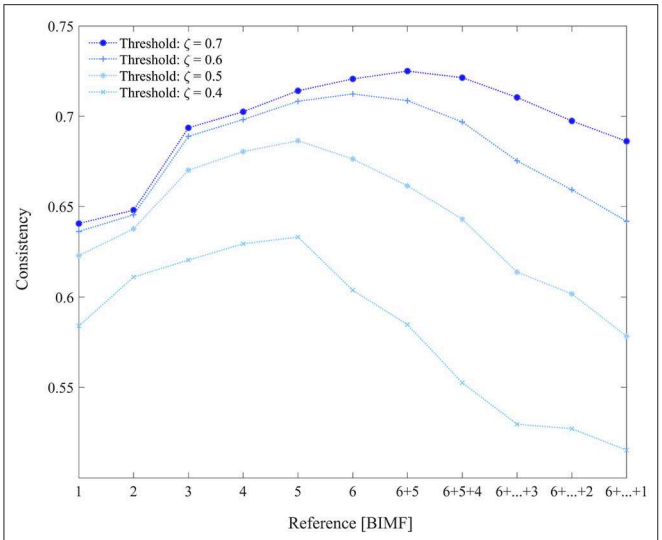


FIGURE 4 | The consistency values of obtained resting-state networks with constrained ICA depending on the different BIMFs used as references. Also BIMFs were gradually summed up again, and combinations like BIMF $6 + 5$ up to $6 + 5 + \ldots + 1$ were evaluated as potential reference signals.

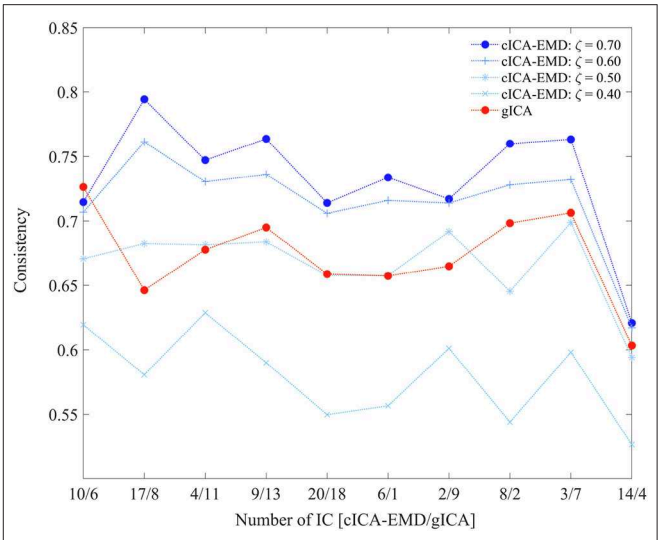


FIGURE 5 | Consistency values of the respective ICs. The numbers on the x-axis refer to the IC number of the cICA-EMD vs. the gICA approach.

TABLE 1 | Consistency values of ICs obtained by the proposed cICA-EMD approach for different settings of the similarity threshold ς, as well as the values of ICs obtained by gICA.

cICA-EMD										
Threshold	**IC #**									
	10	**17**	**4**	**9**	**20**	**6**	**2**	**8**	**3**	**14**
0.40	0.62	0.58	0.63	0.59	0.55	0.56	0.60	0.54	0.60	0.53
0.50	0.67	0.68	0.68	0.68	0.66	0.66	0.69	0.65	0.70	0.59
0.60	0.71	0.76	0.73	0.74	0.71	0.72	0.71	0.73	0.73	0.62
0.70	0.71	0.79	0.75	0.76	0.71	0.73	0.72	0.76	0.76	0.62
gICA										
	IC #									
	6	**8**	**11**	**13**	**18**	**1**	**9**	**2**	**7**	**4**
	0.73	0.65	0.68	0.70	0.66	0.66	0.66	0.70	0.71	0.60

*ICs are sorted as in **Figure 3** so that each column shows consistency values of comparable extracted networks.*

4. DISCUSSION

The motivation of this paper was to propose a novel workflow for extracting resting-state networks that are consistent across a group of subjects. First, the dimensionality of the fMRI dataset was reduced at subject level with PCA. Intrinsic modes were extracted from the data by employing the GiT-BEEMD algorithm (Al-Baddai et al., 2016b). These subject-specific intrinsic modes reflect spatial activity patterns at different spatial frequencies. Hence, for each underlying spatial frequency, a common reference mode can be formed. It turned out that low-frequency modes concentrated the activity into spatially contiguous patterns and were especially well-suited to serve as reference modes for the extraction of independent components with a cICA algorithm. Note that if intrinsic spatial modes, which are naturally ordered according to their dominant local spatial frequency, are chosen as reference signals within a cICA, the resulting independent modes are also ordered in correspondence to their assigned intrinsic modes. Thus, the natural ordering of the intrinsic modes with respect to their spatial frequencies helps to overcome the permutation ambiguity of ICA in extracting consistent resting-state networks across subjects. Previously, references for constrained ICA had to be predefined, in the form of temporal activation profiles or anatomical regions of interest derived from an atlas (Lu and Rajapakse, 2006; Lin et al., 2009; Rodriguez et al., 2014). In the absence of any stimulus, like in resting-state fMRI, such temporal profiles might not be available. Also, in the case of patients with neurological disorders, spatial priors can be inappropriate. Atlases are typically defined based on a cohort of healthy subjects, meaning that reference brain networks defined by the latter might unduly bias the outcome of the analysis. Therefore it was our goal to establish a purely data-driven workflow by hybridizing cICA with EMD and obtaining references from the same data as used in the study. We demonstrated that our fMRI data processing pipeline produces commonly observed resting-state patterns. These functional networks were then compared to those obtained by the widely used gICA, which is based on a temporal concatenation of individual datasets (Calhoun et al., 2001b). It was shown that with the constrained extended Infomax algorithm (Rodriguez et al., 2014), the influence of the references upon the estimation of the related ICs could be controlled well. Based on the mathematically well-described augmented Lagrangian framework in our workflow,

it is transparent to the user with respect to how homologous resting-state networks across subjects are deduced. In the processing pipeline presented, the cICA-EMD approach also allowed the optimization procedure to be shaped by adjusting the threshold parameter, which determines the impact of the reference on the IC extraction. By choosing a lower threshold, e.g., allowing for a lower similarity to the references, more freedom could be given to the exploratory character of ICA and the formation of subject-specific features. Defining a high threshold resulted, across subjects, in a higher consistency of the extracted resting-state networks. These RSNs were even more consistent than those obtained by the conventional gICA. Although there is no exact ground truth on how resting-state networks should ideally look, the threshold can be chosen in a way such that the obtained networks optimally fulfill the requirements of a particular study. For example, when performing a classification task, the threshold can be chosen to maximize the accuracy of the classifier. Thus, the good interpretability and high flexibility of the proposed processing pipeline can offer beneficial properties for application in resting-state studies. Besides applications of EMD for time series analysis in functional MRI (Qian et al., 2015; Goldhacker et al., 2018; Zhang et al., 2018), we showed that spatial EMD can also be used to extract useful intrinsic patterns from functional MRI data, representing characteristic resting-state activations. This could further motivate other researchers to consider the spatial variant of this technique and to investigate other applications of this method in the field of MRI.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://db.humanconnectome.org/.

AUTHOR CONTRIBUTIONS

SW and EL designed the study. SW and MG performed the simulations. SW, EL, AT, and MG wrote the manuscript. MWG and EL supervised the study and revised earlier versions of this manuscript.

ACKNOWLEDGMENTS

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657), funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University. The authors also thank Saad Al-Baddai for helpful discussions. SW and MWG were supported by the Deutsche Forschungsgemeinschaft (grant: GR988/25-1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00221/full#supplementary-material

REFERENCES

Al-Baddai, S., Al-Subari, K., Tomé, A. M., Ludwig, B., Salas-Gonzales, D., and Lang, E. W. (2016a). Analysis of fMRI images with bi-dimensional empirical mode decomposition based-on Green's functions. *Biomed. Signal Process. Control* 30, 53–63. doi: 10.1016/j.bspc.2016.06.019

Al-Baddai, S., Al-Subari, K., Tomé, A. M., Sol-Casals, J., and Lang, E. W. (2016b). A Green's function-based Bi-dimensional empirical mode decomposition. *Informat. Sci.* 348, 305–321. doi: 10.1016/j.ins.2016.01.089

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., and Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex (New York, N.Y.: 1991)* 24, 663–76. doi: 10.1093/cercor/bhs352

Allen, E. A., Erhardt, E. B., Damaraju, E., Gruner, W., Segall, J. M., Silva, R. F., et al. (2011). A baseline for the multivariate comparison of resting-state networks. *Front. Syst. Neurosci.* 5:2. doi: 10.3389/fnsys.2011.00002

Beckmann, C., DeLuca, M., Devlin, J., and Smith, S. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. Ser B Biol. Sci.* 360, 1001–1013. doi: 10.1098/rstb.2005.1634

Beckmann, C., and Smith, S. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage* 25, 294–311. doi: 10.1016/j.neuroimage.2004.10.043

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Biswal, B. B., and Ulmer, J. L. (1999). Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *J. Comput. Assist. Tomogr.* 23, 265–271. doi: 10.1097/00004728-199903000-00016

Biswal, B. B., Yetkin, F. Z., Haughton, V., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Calhoun, V. D., Adali, T., McGinty, V. B., Pekar, J. J., Watson, T. D., and Pearlson, G. D. (2001a). fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage* 14, 1080–1088. doi: 10.1006/nimg.2001.0921

Calhoun, V. D., Adalı, T., Pearlson, G. D., and Pekar, J. J. (2001b). A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* 14, 140–151. doi: 10.1002/hbm.1048

Calhoun, V. D., Liu, J., and Adalı, T. (2009). A review of group ICA for fMRI data and ica for joint inference of imaging, genetic, and ERP data. *NeuroImage* 45(1 Suppl.):S163–72. doi: 10.1016/j.neuroimage.2008.10.057

Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. IEEE Signal Process. Lett. 4, 112–114. doi: 10.1109/97.566704

Correa, N., Adalı, T., and Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magn. Reson. Imaging* 25, 684–694. doi: 10.1016/j.mri.2006.10.017

Erhardt, E. B., Rachakonda, S., Bedrick, E. J., Allen, E. A., Adalı, T., and Calhoun, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* 32, 2075–2095. doi: 10.1002/hbm.21170

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., et al. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage* 25, 193–205. doi: 10.1016/j.neuroimage.2004.10.042

Essen, D. V., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., et al. (2012). The human connectome project: a data acquisition perspective. *NeuroImage* 62, 2222–2231. doi: 10.1016/j.neuroimage.2012.02.018

Feinberg, D., Moeller, S., Smith, S., Auerbach, E., Ramanna, S., Günther, M., et al. (2010). Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PLoS ONE* 5:e15710. doi: 10.1371/journal.pone.0015710

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci*. 8, 700–711. doi: 10.1038/nrn2201

Glasser, M., Sotiropoulos, S., Anthony Wilson, J., Coalson, T., Fischl, B., Andersson, J., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80:105. doi: 10.1016/j.neuroimage.2013.04.127

Goldhacker, M., Tomè, A., Greenlee, M., and Lang, E. (2018). Frequency-resolved dynamic functional connectivity reveals scale-stable features of connectivity-states. *Front. Hum. Neurosci*. 12:253. doi: 10.3389/fnhum.2018.00253

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* 95, 232–247. doi: 10.1016/j.neuroimage.2014.03.034

Harshman, R., and Lundy, M. (1994). PARAFAC: parallel factor analysis. *Comput. Stat. Data Anal*. 18, 39–72. doi: 10.1016/0167-9473(94)90132-5

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage* 22, 1214–1222. doi: 10.1016/j.neuroimage.2004.03.027

Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., et al. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*. 454, 903–995. doi: 10.1098/rspa.1998.0193

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). *FSL. Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Jolliffe, I. (2014). *Principal Component Analysis*. New York, NY: Springer Verlag. doi: 10.1002/9781118445112.stat06472

Kiviniemi, V., Kantola, J.-H., Jauhiainen, J., Hyvärinen, A., and Tervonen, O. (2003). Independent component analysis of nondeterministic fMRI signal sources. *Neuroimage* 19, 253–260. doi: 10.1016/S1053-8119(03)00097-1

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Res. Logist. Q.* 2, 83–97. doi: 10.1002/nav.3800020109

Lee, T.-W., Girolami, M. A., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Comput*. 11, 417–441. doi: 10.1162/089976699300016719

Lin, Q.-H., Liu, J., Zheng, Y.-R., Liang, H., and Calhoun, V. (2009). Semiblind spatial ICA of fMRI using spatial constraints. *Hum. Brain Mapp*. 31, 1076–1088. doi: 10.1002/hbm.20919

Lin, Q.-H., Zheng, Y.-R., Yin, F., Liang, H., and Calhoun, V. D. (2007). A fast algorithm for one-unit ICA-R. *Inf. Sci*. 177, 1265–1275. doi: 10.1016/j.ins.2006.09.011

Lu, W., and Rajapakse, J. C. (2006). ICA with reference. *Neurocomputing* 69, 2244–2257. doi: 10.1016/j.neucom.2005.06.021

McKeown, M. J., Jung, T. P., Makeig, S., Brown, G. E., Kindermann, S., Lee, T. W., et al. (1998). Spatially independent activity patterns in functional MRI data during the stroop color-naming task. *Proc. Natl. Acad. Sci. U.S.A*. 95, 803–810. doi: 10.1073/pnas.95.3.803

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N. Y., et al. (2010). Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med*. 63, 1144–1153. doi: 10.1002/mrm.22361

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math*. 5, 32–38. doi: 10.1137/0105003

Nunes, J. C., Bouaoune, Y., Deléchelle, É., Niang, O., and Bunel, P. (2003). Image analysis by bidimensional empirical mode decomposition. *Image Vision Comput*. 21, 1019–1026. doi: 10.1016/S0262-8856(03)00094-5

Qian, L., Zhang, Y., Zheng, L., Shang, Y., Gao, J.-H., and Liu, Y. (2015). Frequency dependent topological patterns of resting-state brain networks. *PLoS ONE* 10:e0124681. doi: 10.1371/journal.pone.0124681

Remes, J. J., Starck, T., Nikkinen, J., Ollila, E., Beckmann, C. F., Tervonen, O., et al. (2011). Effects of repeatability measures on results of fMRI sICA: a study on simulated and real resting-state effects. *NeuroImage* 56, 554–569. doi: 10.1016/j.neuroimage.2010.04.268

Rodriguez, P. A., Anderson, M. Z., Li, X.-L., and Adalı, T. (2014). General non-orthogonal constrained ICA. *IEEE Trans. Sig. Process*. 62, 2778–2786. doi: 10.1109/TSP.2014.2318136

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468. doi: 10.1016/j.neuroimage.2013.11.046

Schmithorst, V. J., and Holland, S. K. (2004). Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. *J. Magn. Reson. Imaging* 19, 365–368. doi: 10.1002/jmri.20009

Setsompop, K., Gagoski, B., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magn. Reson. Med*. 67, 1210–1224. doi: 10.1002/mrm.23097

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., et al. (2013). Resting-state fMRI in the human connectome project. *NeuroImage* 80, 144–168. doi: 10.1016/j.neuroimage.2013.05.039

Svensén, M., Kruggel, F., and Benali, H. (2002). ICA of fMRI group study data. *NeuroImage* 16(3 Pt 1):551–563. doi: 10.1006/nimg.2002.1122

van den Heuvel, M. M. H. P., and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *Eur. Neuropsychopharmacol*. 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

Wein, S., Tomé, A. M., Goldhacker, M., Greenlee, M. W., and Lang, E. W. (2019). "Hybridizing EMD with cICA for fMRI analysis of patient groups," in *Proceedings of the EMBC 2019* (Berlin). doi: 10.1109/EMBC.2019.8856355

Xu, J., Moeller, S., Strupp, J., Auerbach, E., Chen, L., Feinberg, A., et al. (2012). "Highly accelerated whole brain imaging using aligned-blipped-controlled-aliasing multiband EPI," in *Proceedings of the 20th Annual Meeting of ISMRM* (Melbourne, VIC), 2036.

Zhang, Z., Liu, G., Yao, Z., Zheng, W., Xie, Y., Hu, T., et al. (2018). Changes in dynamics within and between resting-state subnetworks in juvenile myoclonic epilepsy occur at multiple frequency bands. *Front. Neurol*. 9:448. doi: 10.3389/fneur.2018.00448

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wein, Tomé, Goldhacker, Greenlee and Lang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

METHODS
published: 04 June 2020
doi: 10.3389/fnins.2020.00289

MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data

*Matthias S. Treder**

School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

MVPA-Light is a MATLAB toolbox for multivariate pattern analysis (MVPA). It provides native implementations of a range of classifiers and regression models, using modern optimization algorithms. High-level functions allow for the multivariate analysis of multi-dimensional data, including generalization (e.g., time x time) and searchlight analysis. The toolbox performs cross-validation, hyperparameter tuning, and nested preprocessing. It computes various classification and regression metrics and establishes their statistical significance, is modular and easily extendable. Furthermore, it offers interfaces for LIBSVM and LIBLINEAR as well as an integration into the FieldTrip neuroimaging toolbox. After introducing MVPA-Light, example analyses of MEG and fMRI datasets, and benchmarking results on the classifiers and regression models are presented.

Keywords: machine learning, classification, decoding, regression, MVPA, regularization, cross-validation, toolbox

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Veena A. Nair,
University of Wisconsin-Madison, United States
Stefan Haufe,
Charité–Universittsmedizin Berlin, Germany

***Correspondence:**
Matthias S. Treder
trederm@cardiff.ac.uk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *15 August 2019*
Accepted: *12 March 2020*
Published: *04 June 2020*

Citation:
Treder MS (2020) MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data. Front. Neurosci. 14:289. doi: 10.3389/fnins.2020.00289

1. INTRODUCTION

Multivariate pattern analysis (MVPA) refers to a set of multivariate tools for the analysis of brain activity or structure. It draws on supervised learning, a branch of machine learning mainly dealing with classification and regression problems. Multivariate classification has been used in EEG-based brain-computer interfaces since at least the 1980s (Farwell and Donchin, 1988), but it did not become a mainstream tool in cognitive neuroscience until the late 2000s (Mur et al., 2009; Pereira et al., 2009; Blankertz et al., 2011; Lemm et al., 2011). MVPA was first popularized by the seminal work of Haxby et al. (Haxby et al., 2001; Norman et al., 2006; Haxby, 2012). In an fMRI study, the authors provided evidence that visual categories (such as faces and houses) are associated with distributed representations across multiple brain regions. MVPA is designed to exploit such multivariate patterns by taking into account multiple voxels or channels simultaneously. This constitutes a major difference between MVPA and traditional statistical methods such as *t*-test and analysis of variance (ANOVA). Traditional statistical tests are often univariate i.e., a test is performed for each dependent variable, for instance voxel or EEG channel, separately. In contrast to MVPA, such tests are blind to the distributed information encoded in the correlations between different spatial locations.

To highlight this difference with an example, consider a hypothetical visual experiment: In each trial, subjects are presented an image of either a face or a house and their brain activity is recorded using fMRI. To make sure that they maintain attention, subjects are instructed to indicate via a button press whether the image represents a face or a house. This experiment will be referred to as "faces vs. houses" throughout this paper. To investigate the difference between the brain responses to faces vs. houses, a *t*-test can be applied to answer the question "Is the activity *at a specific voxel* different for faces vs. houses?." In contrast, MVPA addresses the more general

question "Is the *pattern of brain activity* different for faces vs. houses?." This example illustrates that univariate statistics and MVPA inhabit opposite ends of a spectrum between *sensitivity* ("Is there an effect?") and *localizability* ("Where is the effect?"). A classical univariate test might be unable to detect a specific effect because it is blind to multivariate dependencies (low sensitivity) but any effect it does detect is perfectly localized to a single voxel. In contrast, MVPA gains statistical power by capitalizing on correlations between different locations (high sensitivity) but it is difficult to attribute an effect to a specific brain location (low localizability). A MVPA technique called *searchlight analysis* (see glossary) attempts to cover the middle ground between these two extremes. As this comparison illustrates, MVPA should be considered as a complement, rather than a competitor, to traditional statistical methods. Finally, there are other ways in which MVPA and traditional statistics differ. For instance, MVPA includes kernel methods that are sensitive to non-linear relationships and it makes extensive use of techniques such as *cross-validation* that control for *overfitting*.

To use MVPA as part of a neuroimaging analysis pipeline, numerous excellent MATLAB toolboxes have been developed over the years, including the Amsterdam Decoding and Modeling Toolbox (ADAM) (Fahrenfort et al., 2018), BCILAB (Kothe and Makeig, 2013), Berlin BCI toolbox (Blankertz et al., 2016), CoSMoMVPA (Oosterhof et al., 2016), Decision Decoding ToolBOX (DDTBOX) (Bode et al., 2019), Donders Machine Learning Toolbox (DMLT) (`github.com/distrep/DMLT`), Pattern Recognition for Neuroimaging Toolbox (PRoNTo) (Schrouff et al., 2013), and The Decoding Toolbox (TDT) (Hebart et al., 2015). Beyond MATLAB, the currently most popular computer languages for machine learning are Python and R, with outstanding toolboxes such as Scikit Learn (Pedregosa et al., 2011) for Python and Caret (Kuhn, 2008) and MLR (Bischl et al., 2000) for R. A comprehensive comparison of MVPA-Light with all of these toolboxes is beyond the scope of this paper, but what sets it apart is the adherence to all of the following design principles:

- *Self-contained*: unlike many toolboxes that provide wrappers for existing classifiers, the backbone of MVPA-Light is native implementations of various classifiers, regression models, and their corresponding optimization algorithms (Trust-Region Newton, Dual Coordinate Descent). As a result, MVPA-Light works out-of-the-box, without the need for additional toolboxes or code compilation.
- *Transparent*: the toolbox has a shallow code base with well-documented functions. In many cases, the function call stack has a depth of two within the toolbox. For instance, a call to `mv_classify` using an LDA classifier triggers calls to functions such as `mv_check_inputs`, `train_lda`, and `test_lda`. Although the train/test functions might call additional optimization functions, most of the work is done at these two shallowest levels. To preserve the shallowness, high-level functions replicate some code that might be shared otherwise. Object orientation and encapsulation is avoided in favor of the more transparent MATLAB structs.
- *Fast*: all models and high-level functions are written with speed as a prime concern. In some cases, the need for speed conflicts with the out-of-the-box requirement. For instance, Logistic Regression and SVM use iterative optimization algorithms written in MATLAB. However, these algorithms potentially run faster using compiled code. To this end, an interface is provided for LIBSVM (Chang et al., 2011) and LIBLINEAR (Fan et al., 2008), two C implementations of Logistic Regression and SVM for users who do not shy away from compiling the code on their platform.
- *Modular and pluggable*: it is possible, and intended, to harvest parts of the code such as the classifiers for other purposes. It is also easy to plug the toolbox into a larger neuroimaging analysis framework. An interface for FieldTrip (Oostenveld et al., 2011) is described in the Methods section.
- *High-level interface*: common MVPA tasks such as searchlight analysis and time generalization including cross-validation can be performed with a few lines of MATLAB code. Many of the hyperparameters required by classifiers and regression models are automatically selected by MVPA-Light, taking the burden of hyperparameter selection off the user.

It is worth noting that MVPA-Light is a purely statistical toolbox. That is, it assumes that data has been preprocessed with a neuroimaging toolbox and comes in the shape of MATLAB arrays. Many neuroimaging toolboxes (e.g., FieldTrip, SPM, EEGLAB) store the imaging data in such arrays, so that MVPA-Light can easily be used as a plugin tool. This comes with the perk that adaptation to different imaging modalities is straightforward.

1.1. MVPA Glossary

MVPA comes with its own set of commonly used terms, many of which are borrowed from machine learning. Since they are used extensively throughout the paper, a glossary is provided here. Fully understanding these concepts can be challenging so unfamiliar readers are referred to review papers on MVPA (Mur et al., 2009; Pereira et al., 2009; Misaki et al., 2010; Grootswagers et al., 2017; Varoquaux et al., 2017). For an in-depth introduction to machine learning refer to standard textbooks (Bishop, 2007; Hastie et al., 2009; James et al., 2013).

- *Binary classifier*. A classifier trained on data that contains two classes, such as in the "faces vs. houses" experiment. If there is more than two classes, the classifier is called a multi-class classifier.
- *Classification*. One of the primary applications of MVPA. In classification, a classifier takes a multivariate pattern of brain activity (referred to as *feature vector*) as input and maps it onto a categorical brain state or experimental condition (referred to as *class label*). In the "faces vs. houses" experiment, the classifier is used to investigate whether patterns of brain activity can discriminate between faces and houses.
- *Classifier*. An algorithm that performs classification, for instance Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).
- *Classifier output*. If a classifier receives a pattern of brain activity (feature vector) as input, its output is a predicted class

label e.g., "face." Many classifiers are also able to produce class probabilities (representing the probability that a brain pattern belongs to a specific class) or decision values.

- *Class label*. Categorical variable that represents a label for each sample/trial. In the "faces vs. houses" experiment, the class labels are "face" and "house." Class labels are often encoded by numbers, e.g., "face" = 1 and "house" = 2, and arranged as a vector. For instance, the class label vector [1, 2, 1] indicates that a subject viewed a face in trial 1, a house in trial 2, and another face in trial 3.
- *Cross-validation*. To obtain a realistic estimate of classification or regression performance and control for overfitting, a model should be tested on an independent dataset that has not been used for training. In most neuroimaging experiments, there is only one dataset with a restricted number of trials. K-fold cross-validation makes efficient use of such data by splitting it into k different folds. In every iteration, one of the k folds is held out and used as test set, whereas all other folds are used for training. This is repeated until every fold served as test set once. Since cross-validation itself is stochastic due to the random assignment of samples to folds, it can be useful to repeat the cross-validation several times and average the results. See Lemm et al. (2011) and Varoquaux et al. (2017) for a discussion of cross-validation and potential pitfalls.
- *Data*. From the perspective of a classifier or regression model, a dataset is a collection of samples (e.g., trials in an experiment). Each sample consists of a brain pattern and a corresponding class label or response. In formal notation, each sample consists of a pair $(\mathbf{x}, y)$ where $\mathbf{x}$ is a feature vector and y is the corresponding class label or response.
- *Decision boundary*. Classifiers partition feature space into separate regions. Each region is assigned to a specific class. Classifiers make predictions for a test sample by looking up into which region it falls. The boundary between regions is known as decision boundary. For linear classifiers, the decision boundary is also known as a hyperplane.
- *Decision value*. Classifiers such as LDA and SVM produce decision values which can be thresholded to produce class labels. For linear classifiers and kernel classifiers, a decision value represents the distance to the decision boundary. The further away a test sample is from the decision boundary, the more confident the classifier is about it belonging to a particular class. Decision values are unitless.
- *Decoder*. An alternative term for a *classifier* or *regression model* that is popular in the neuroimaging literature. The term nicely captures the fact that it tries to invert the encoding process. In encoding e.g., a sensory experience such as viewing a face is translated into a pattern of brain activity. In decoding, one starts from a pattern of brain activity and tries to infer whether it was caused by a face or a house stimulus.
- *Feature*. A feature is a variable that is part of the input to a model. If the dataset is tabular with rows representing samples, it typically corresponds to one of the columns. In the "faces vs. houses" experiment, each voxel represents a feature.
- *Feature space*. Usually a real vector space that contains the feature vectors. The dimensionality of the feature space is equal to the number of features.
- *Feature vector*. For each sample, features are stored in a vector. For example, consider a EEG measurement with three electrodes Fz, Cz, and Oz and corresponding voltages 40, 65, and 97 μV. The voltage at each EEG sensor represents a feature, so the corresponding feature vector is the vector $[40, 65, 97] \in \mathbb{R}^3$.
- *Fitting (a model)*. Same as *training*.
- *Hyperparameter*. A parameter of a model that needs to be specified by the user, such as the type and amount of regularization applied, the type of kernel, and the kernel width γ for Gaussian kernels. From the user's perspective, hyperparameters can be nuisance parameters: it is sometimes not clear a priori how to set them, but their exact value can have a substantial effect on the performance of the model.
- *Hyperparameter tuning*. If it is unclear how a hyperparameter should be set, multiple candidate values can be tested. Typically, this is done via nested cross-validation: the training set is again split into separate folds. A model is trained for each of the candidate values and its performance is evaluated on the held-out fold, called validation set. Only the model with the best performance is then taken forward to the test set.
- *Hyperplane*. For linear classifiers, the decision boundary is a hyperplane. In the special case of a two-dimensional feature space, a hyperplane corresponds to a straight line. In three dimensions, it corresponds to a plane.
- *Loss function*. A function that is used for training. The model parameters are optimized such that the loss function attains a minimum value. For instance, in Linear Regression the sum of squares of the residuals serves as a loss function.
- *Metric*. A quantitative measure of the performance of a model on a test set. For example, precision/recall for classification or mean squared error for regression.
- *Model*. In the context of this paper, a model is a classifier or regression model.
- *Multi-class classifier*. A classifier trained on data that contains three or more classes. For instance, assume that in the "faces vs. houses" experiment additional images have been presented depicting "animals" and "tools." This would define four classes in total, hence classification would require a multi-class classifier.
- *Overfitting*. Occurs when a model over-adapts to the training data. As a consequence, it will perform well on the training set but badly on the test set. Generally speaking, overfitting is more likely to occur if the number of features is larger than the number of samples, and more likely for complex non-linear models than for linear models. Regularization can serve as an antidote to overfitting.
- *Parameters*. Models are governed by parameters e.g., beta coefficients in Linear Regression or the weight vector $\mathbf{w}$ and bias b in a linear classifier.
- *Regression*. One of the primary applications of MVPA (together with classification). Regression is very similar to classification, but it aims to predict a continuous variable rather than a class label. For instance, in the 'faces vs. houses' experiment, assume that the reaction time of the button press has been recorded, too. To investigate the question "Does the pattern of brain activity in each trial predict reaction

time?," regression can be performed using reaction time as responses.
- *Regression model*. An algorithm that performs regression, for instance Ridge Regression and Support Vector Regression (SVR).
- *Regularization*. A set of techniques that aim to reduce overfitting. Regularization is often directly incorporated into training by adding a penalty term to the loss function. For instance, L1 and L2 penalty terms are popular regularization techniques. They reduce overfitting by preventing coefficients from taking on too large values.
- *Response*. In regression, responses act as the target values that a model tries to predict. They play the same role that class labels play in classification. Unlike class labels, responses are continuous e.g., reaction time.
- *Searchlight analysis*. In neuroimaging analysis, a question such as "Does brain activity differentiate between faces and houses?" is usually less interesting than the question "Which brain regions differentiate between faces and houses?." In other words, the goal of MVPA is to establish the presence of an effect *and* localize it in space or time. Searchlight analysis intends to marry statistical sensitivity with localizability. It is a well-established technique in the fMRI literature, where a searchlight is defined e.g., as a sphere of 1 cm radius, centered on a voxel in the brain (Kriegeskorte et al., 2006). All voxels within the radius serve as features for a classification or regression analysis. The result of the analysis is assigned to the central voxel. If the analysis is repeated for all voxel positions, the resultant 3D map of classification accuracies can be overlayed on a brain image. Brain regions that have discriminative information then light up as peaks in the map. Searchlight analysis is not limited to spatial coordinates. The same idea can be applied to other dimensions such as time points and frequencies.
- *Testing*. The process of applying a trained model to the test set. The performance of the model can then be quantified using a metric.
- *Test set*. Part of the data designated for testing. Like with training sets, test sets are automatically defined in cross-validation, or they can arise naturally in multi-site studies or in experiments with different phases.
- *Training*. The process of optimizing the parameters of a model using a training set.
- *Training set*. Part of the data designated for training. In cross-validation, a dataset is automatically split into training and test sets. In other cases, a training set may arise naturally. For instance, in experiments with different phases (e.g., memory encoding and memory retrieval) one phase may serve as training set and the other phase as test set. Another example is multi-site studies, where a model can be trained on data from one site and tested on data from another site.
- *Underfitting*. Occurs when a classifier or regression model is too simple to explain the data. For example, imagine a dataset wherein the optimal decision boundary is a circle, with samples of class 1 being inside the circle and samples of class 2 outside. A linear classifier is not able to represent a circular decision boundary, hence it will be unable to adequately solve the task. Underfitting can be checked by fitting a complex model (e.g., kernel SVM) to data. If the complex model performs much better than a more simple linear model (e.g., LDA) then it is likely that the simple model underfits the data. In most neuroimaging datasets, overfitting is more of a concern than underfitting.

The rest of the paper is structured as follows. The high-level functions of the toolbox are described, followed by an introduction of the classifiers and regression models. Then, example analyses are presented using a publicly available Wakeman and Henson (2014, 2015) MEEG dataset and the Haxby et al. (2001) fMRI dataset. Finally, a benchmarking analysis is conducted wherein the computational efficiency of the classifiers and regression models in MVPA-Light is compared to models in other toolboxes in MATLAB, Python, and R.

2. MATERIALS AND METHODS

2.1. Requirements

A standard desktop computer is sufficient to run MVPA-Light. The RAM requirement is dictated by the memory footprint of the dataset. Since some functions operate on a copy of the data, it is recommended that the available RAM exceeds the size of the dataset by at least a factor of two (e.g., 4+ GB RAM for a 2 GB dataset). MVPA-Light is supported by MATLAB 2012a and more recent versions. The Statistics toolbox is required at some points in the toolbox (e.g., for calculating t-values). The cluster permutation test in `mv_statistics` uses the Image Processing toolbox to extract the clusters.

2.2. Getting Started

MVPA-Light is shipped with a set of example scripts (in the `/examples` subfolder) and an example EEG dataset. These scripts cover both the high-level functions in MVPA-Light and calling the train/test functions manually. The best starting point is to work through the example scripts and then adapt them to one's purpose. An up-to-date introduction to the toolbox with relevant hyperlinks is provided on the GitHub page (`github.com/treder/mvpa-light`).

The EEG data has been taken from the BNCI-Horizon-2020 repository (http://bnci-horizon-2020.eu/database). It consists of three `mat` files corresponding to three subjects (subject codes `VPaak`, `VPaan`, and `VPgcc`) from the auditory oddball paradigm introduced in Treder et al. (2014). Out of the experimental conditions, the "SynthPop" condition has been selected. Attended and unattended deviants are coded as class 1 and 2. The 64 EEG channels in the original dataset have been reduced to 32 channels.

To give a concrete code example, consider the "faces vs. houses" experiment. For each trial, the BOLD response has been recorded for all voxels. This yields a [samples x voxels] data matrix for one subject, where the samples correspond to trials and the voxels serve as features. The matrix is denoted as `X`. Each trial corresponds to either a "face" or a "house" stimulus. This is encoded in a vector of class labels, denoted as `clabel`, that contains 1's and 2's ("face" = 1, "house" = 2). Then the

following piece of code performs 10-fold cross-validation with 2 repetitions. LDA is used as classifier and area under the ROC curve (AUC) is calculated as a classification metric.

```
cfg = [];
cfg.model      = 'lda';
cfg.metric     = 'auc';
cfg.cv         = 'kfold';
cfg.k          = 10;
cfg.repeat     = 2;

auc = mv_classify(cfg, X, clabel);
```

The output value `auc` contains the classifier performance measure, in this case a single AUC value averaged across test folds and repetitions. `mv_classify` is part of the high-level interface that will be discussed next.

2.3. High-level Interface

The structure of MVPA-Light is depicted in **Figure 1**. The toolbox can be interacted with through high-level functions that cover common classification tasks. `mv_classify` is a general-purpose function that works on data of arbitrary dimension (e.g., time-frequency data). It performs any combination of cross-validation, searchlight analysis, generalization, and other tasks. Two more specialized functions are provided for convenience: `mv_classify_across_time` and `mv_classify_timextime`, assume that the data has a time dimension i.e., it is a 3-D [samples × features × time points] array. `mv_classify_across_time` performs classification for every time point, resulting in a vector of cross-validated metrics, the length of the vector being the number of time points. `mv_classify_timextime` expects the same 3-D input. It implements time generalization (King and Dehaene, 2014) i.e., classification for every combination of training and test time points, resulting in a 2-D matrix of cross-validated metrics. For regression tasks, the equivalent to `mv_classify` is the function `mv_regress`. It also works with data of arbitrary dimension and supports both searchlight and generalization.

All high-level functions take three input arguments. First, `cfg`, a configuration structure wherein parameters for the analysis can be set. Second, `X`, the data acting as input to the model. Third, `clabel` or `y`, a vector of class labels or responses. Some of the parameters in the `cfg` struct are common to all high-level functions:

- `cfg.model`: name of the classifier or regression model, e.g., `'lda.'`
- `cfg.hyperparameter`: a struct that specifies the hyperparameters for the model. For instance, `cfg.hyperparameter.lambda = 0.1` sets the magnitude of shrinkage regularization in LDA. LDA's hyperparameters are introduced in section 2.4.3.
- `cfg.metric`: specifies the metric to be calculated from the model predictions e.g., classification accuracy or mean-squared error for regression. Metrics are introduced in section 2.6.
- `cfg.preprocess`: a struct that specifies a nested preprocessing pipeline. The pipeline consists of preprocessing operations that are applied on train and test data separately. Preprocessing is discussed in section 2.3.3.

2.3.1. Cross-Validation

Cross-validation is implemented in all high-level functions. It is controlled by the following parameters that are part of the `cfg` struct defined in the previous section:

- `cfg.cv`: cross-validation type, either `'kfold,'` `'leaveout,'` `'predefined,'` `'holdout,'` or `'none'`.
- `cfg.k`: number of folds in k-fold cross-validation.
- `cfg.repeat`: number of times the cross-validation is repeated with new randomly assigned folds.
- `cfg.p`: if `cfg.cv = 'holdout,'` `p` is the fraction of test samples.
- `cfg.fold`: if `cfg.cv = 'predefined,'` fold is a vector of integers that specifies which fold a sample belongs to.
- `cfg.stratify`: if 1, for classification, the class proportions are approximately preserved in each test fold.

See the function `mv_get_crossvalidation_folds` for more details.

2.3.2. Hyperparameter Tuning

MVPA-Light tries to automate hyperparameter selection as much as possible. This is done using either reasonable default values, hyperparameter estimators [Ledoit and Wolf (2004) for LDA] or hyperparameter-free regularizers (log-F(1,1) for Logistic Regression). If this is not possible, automated grid search using nested cross-validation can be used for testing out different hyperparameter combinations essentially by brute force. For better performance, bespoke hyperparameter tuning functions are implemented for some classifiers. Otherwise, the generic tuning function `mv_tune_hyperparameter` is used.

2.3.3. Preprocessing

Preprocessing refers to operations applied to the data prior to training the classifier. To not bias the result, some preprocessing operations (such as Common Spatial Patterns) should be performed in a "nested" fashion. That is, they are performed on the training data first and subsequently applied to the test data using parameters estimated from the training data (Lemm et al., 2011; Varoquaux et al., 2017). Currently implemented functions include PCA, sample averaging (Cichy and Pantazis, 2017), kernel averaging (Treder, 2018), and under-/oversampling for unbalanced data. Preprocessing pipelines are defined by adding the `cfg.preprocess` parameter. For instance,

```
cfg.preprocess = {'undersample,' 'zscore,'
 'average_kernel'}
```

adds a preprocessing pipeline that performs undersampling of the data followed by z-scoring and kernel averaging.

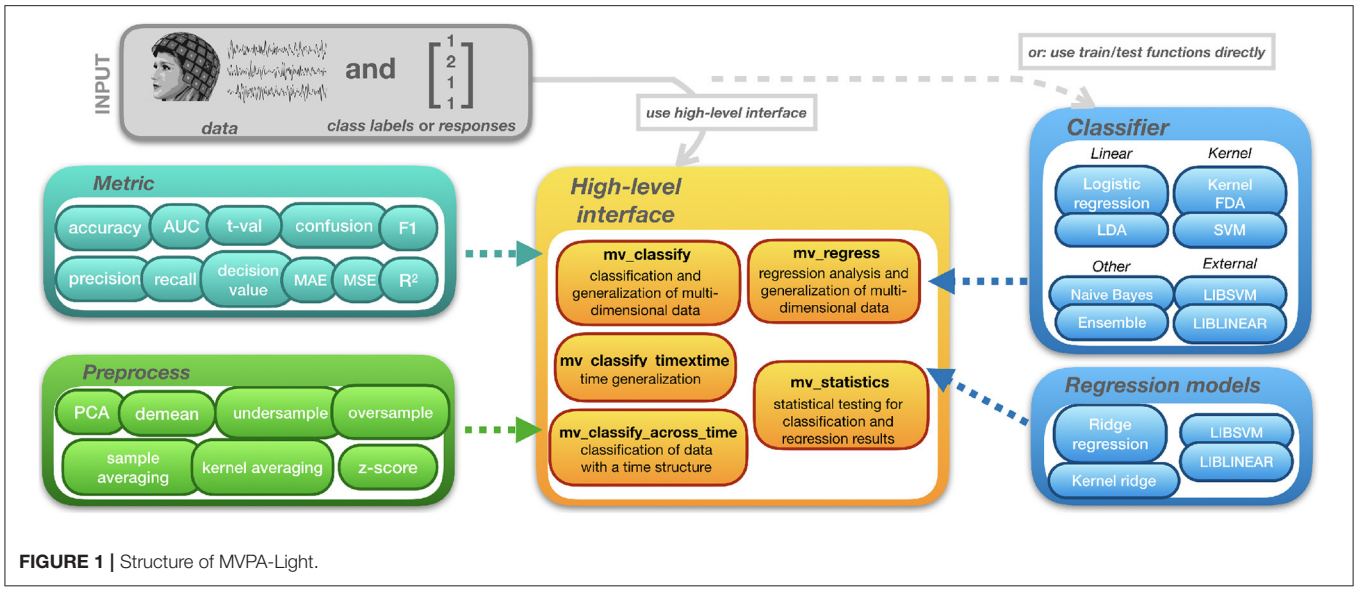


FIGURE 1 | Structure of MVPA-Light.

2.3.4. Searchlight Analysis

In MVPA-Light, `mv_classify_across_time` performs searchlight analysis across the time axis. More bespoke searchlight analyses can be conducted using `mv_classify` and `mv_regress` by setting the parameter `cfg.neighbours`.

2.4. Classifiers

The main workhorses of MVPA are classifiers and regression models. **Figure 2** provides a pictorial description of the classifiers. They are implemented using pairs of train/test functions. In the high-level interface, a classifier and its hyperparameters can be specified using `cfg.model` and `cfg.hyperparameter`. For instance,

```
cfg.model = 'lda';
cfg.hyperparameter.lambda = 0.1;
```

specifies an LDA classifier and sets the hyperparameter `lambda = 0.1`. The `cfg` struct can then be used in a high-level function call, e.g., `acc = mv_classify_across_time(cfg, X, clabel)`. Alternatively, as a low-level interface, the train/test functions can be called directly. For instance, an LDA classifier can be trained directly using

```
model = train_lda(param, X, clabel)
```

where `X` is the training data and `clabel` are the corresponding class labels. `param` is a MATLAB struct that contains hyperparameters (same as `cfg.hyperparameter`). It can be initialized by calling `param = mv_get_hyperparameter('lda')`. An explanation of the hyperparameters for LDA is given when typing `help('train_lda')` in MATLAB. The output `model` is a struct that contains the classifier's parameters after training. The classifier can be applied to test data, denoted as `Xtest`, by calling

```
[clabel, dval, prob] = test_lda(model,
 Xtest)
```

The first output argument `clabel` is the predicted class labels. They can be compared against the true class labels to calculate a classification performance metric. `test_lda` provides two additional outputs, but not all classifiers have this capability. `dval` is the decision value, a dimensionless quantity that measures the distance to the hyperplane. `prob` contains the probability for a given sample to belong to class 1.

To introduce some mathematical notation needed in the following, data is denoted as a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of n samples and p predictors/features. The i-th row of $\mathbf{X}$ is denoted as the column vector $\mathbf{x}_i \in \mathbb{R}^p$. Class labels are stored in a vector $\mathbf{y} \in \mathbb{R}^n$ with y_i referring to the i-th class label. When the index is not relevant, the feature vector and class label are simply referred to as $\mathbf{x}$ and y. Before describing the classifiers, two conceptual perspectives are introduced that highlight some of their similarities.

2.4.1. Perspective 1: Linear Classifiers

For two classes, linear classifiers such as LDA, Logistic Regression, and linear SVM act on the data in a unified way. The decision value for a test sample $\mathbf{x}$ is given by

$$\text{dval} = \mathbf{w}^\top \mathbf{x} + b \tag{1}$$

where $\mathbf{w}$ is the weight vector or normal to the hyperplane specifying the linear combination of features, and b is the threshold/bias term. A sample is assigned to the first class if `dval` > 0 and to the second class if `dval` < 0. If we encode class 1 as +1 and class 2 as −1, this can be expressed concisely as

$$\text{predicted class} = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$$

where $\text{sign} : \mathbb{R} \rightarrow \{-1, +1\}$ is the sign function. Linear classifiers differ only in the way that $\mathbf{w}$ and b are derived.

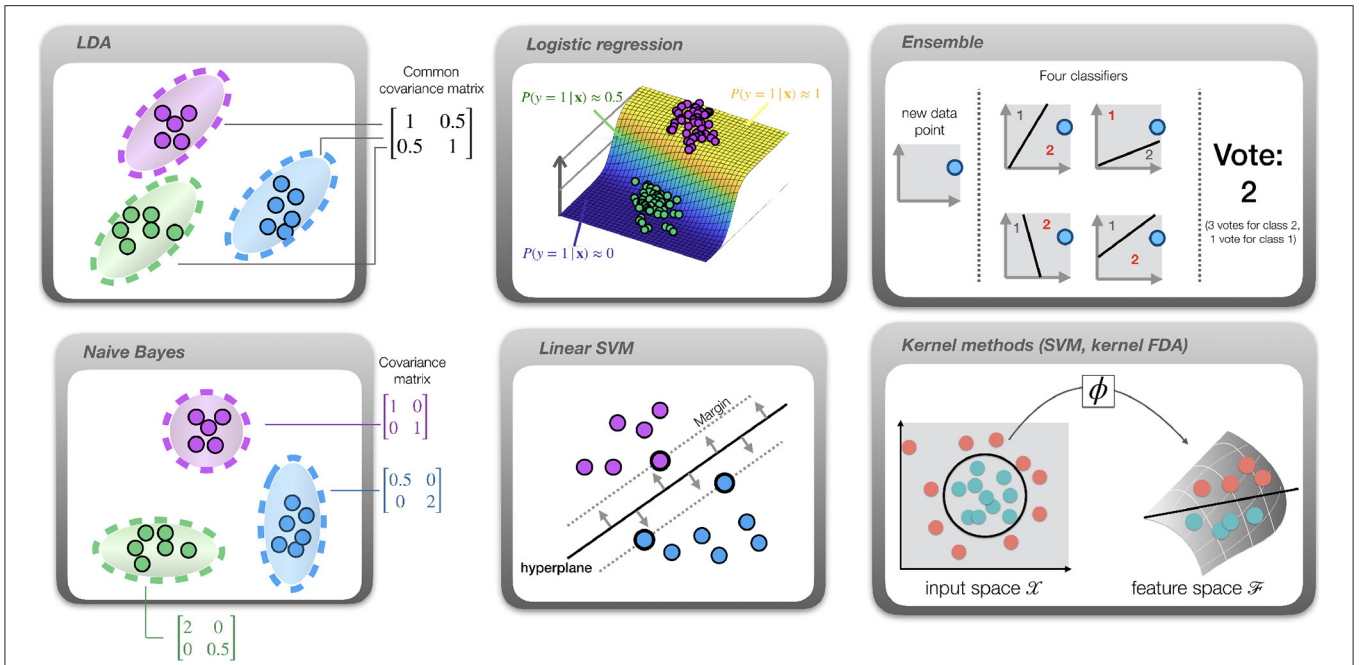


FIGURE 2 | Overview of the available classifiers. Dots represent samples, color indicates the class. *LDA*: different classes are assumed to have the same covariance matrix, indicated by the ellipsoids. *Gaussian Naive Bayes*: features are conditionally independent, yielding diagonal covariance matrices. *Logistic regression*: a sigmoid function (curved plane) is fit to directly model class probabilities. *SVM*: a hyperplane (solid line) is fit such that the margin (distance from hyperplane to closest sample; indicated by dotted lines) is maximized. *Ensemble*: multiple classifiers are trained on subsets of the data. In this example, their hyperplanes partition the data into spaces belonging to classes 1 and 2. After applying all classifiers to a new data point and collecting their "votes," the class receiving most votes is selected. *Kernel methods*: in this example the optimal decision boundary is circular (circle), hence the data is not linearly separable. After projection into a high-dimensional feature space using a map ϕ, the data becomes linearly separable (solid line) and a linear classifier such as SVM or LDA can be successfully applied in this space.

2.4.2. Perspective 2: Probabilistic Classifiers

Another useful perspective is given by the Bayesian framework (Bishop, 2007). Probabilistic classifiers such as LDA, Naive Bayes, and Logistic Regression are able to directly model class probabilities for individual samples. Let us denote the (posterior) probability for class i given test sample $\mathbf{x}$ as $P(y = i \,|\, \mathbf{x})$. A possible approach for calculating this quantity is Bayes' theorem:

$$P(y = i \,|\, \mathbf{x}) = \frac{P(\mathbf{x} \,|\, y = i)\, P(y = i)}{P(\mathbf{x})} \tag{2}$$

Here, $P(\mathbf{x} \,|\, y = i)$ is the likelihood function which quantifies the relative probability of observing $\mathbf{x}$ given the class label, and $P(y = i)$ is the prior probability for a sample to belong to class i. The denominator, called evidence, can be calculated by marginalizing across the classes: $P(\mathbf{x}) = \sum_i P(\mathbf{x} \,|\, y = i)\, P(y = i)$.

2.4.3. Linear Discriminant Analysis (LDA)

If the classes follow a multivariate Gaussian distribution with a common covariance matrix for all classes, LDA yields the theoretically optimal classifier (Duda et al., 2001). In the context of EEG/MEG analysis, LDA is discussed in detail in Blankertz et al. (2011). The likelihood function takes the form

$$P(\mathbf{x} \,|\, y = i) \sim \mathcal{N}(\mathbf{m}_i, \Sigma) \tag{3}$$

i.e., it is multivariate Gaussian distributed with a class-specific mean $\mathbf{m}_i$ and common covariance matrix Σ. Both need to be estimated from the training data. Equation (2) can then be evaluated to calculate class probabilities. A prediction can be done by selecting the most likely class out of all candidate classes,

$$\text{predicted class} = \arg\max_i P(y = i \,|\, \mathbf{x})$$

which is known as the maximum a posteriori (MAP) rule. LDA is closely related to other statistical models. For two classes, LDA is equivalent to Linear Regression using the class labels as targets. It is also equivalent to Linearly Constrained Minimum Variance (LCMV) beamforming when applied to ERP data (Treder et al., 2016). The latter equivalence relationship also applies to other methods based on generalized eigenvalue decomposition of covariance matrices (De Cheveigné and Parra, 2014).

In MVPA-Light, multi-class LDA is implemented as the classifier `'multiclass_lda.'` For two classes, a more efficient implementation denoted as `'lda'` is available. In practice, the covariance matrix is often ill-conditioned and needs to be regularized (Blankertz et al., 2011). The hyperparameter `lambda` controls the amount of regularization. In shrinkage regularization, `lambda` $\in [0, 1]$ blends between the empirical covariance matrix (`lambda` $= 0$) and a scaled identity matrix (`lambda` $= 1$). By default, `lambda` is estimated automatically

using the Ledoit-Wolf formula (Ledoit and Wolf, 2004). Section 4.1 (**Appendix** in Supplementary Material) discusses the implementation of LDA in detail.

2.4.4. Naive Bayes

In Naive Bayes, the features are assumed to be conditionally independent of each other given the class label (Bishop, 2007). While this is indeed naive and often wrong, Naive Bayes has nevertheless been remarkably successful in classification problems. The independence assumption leads to a straightforward formula for the likelihood function since only univariate densities need to be estimated. Let $x^{(j)}$ be the j-th feature and $\mathbf{x} = [x^{(1)}, x^{(2)}, ..., x^{(p)}]^\top$ be a feature vector then the likelihood function is given by

$$P(\mathbf{x} \mid y = i) = \prod_{j=1}^{p} P(x^{(j)} \mid y = i)$$

Like in LDA, the predicted class can be obtained using the MAP rule. In MVPA-Light, Naive Bayes is implemented as `'naive_bayes.'` Additionally, MVPA-Light assumes that these densities are univariate Gaussian i.e., $P(x^{(j)} \mid y = i) \sim \mathcal{N}(m_{ij}, \sigma_{ij}^2)$. For Gaussian densities, the independence assumption is equivalent to assuming that the covariance matrix is diagonal. As indicated in **Figure 2**, there is a close relationship between LDA and Gaussian Naive Bayes: LDA allows for a dense covariance matrix, but it requires that it is the same for all classes. In contrast, Naive Bayes allows each class to have a different covariance matrix, but it requires each matrix to be diagonal. Additional details on the implementation are given in section 4.2 (**Appendix** in Supplementary Material).

2.4.5. Logistic Regression

In Logistic Regression for two classes, the posterior probability is modeled directly by fitting a logistic function to the data (Hastie et al., 2009). If the two classes are coded as +1 and −1, it is given by

$$P(y = \pm 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-y\,(\mathbf{w}^\top \mathbf{x} + b))} \tag{4}$$

The weights $\mathbf{w}$ are found by minimizing the logistic loss function

$$\mathcal{L}_{\mathrm{LR}}(\mathbf{w}) = \sum_{i=1}^{n} \log[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))] \tag{5}$$

In MVPA-Light, Logistic Regression is implemented as `'logreg.'` By default, log-F(1,1) regularization (`reg = 'logf'`) is used by imposing Jeffrey's prior on the weights (Firth, 1993; King and Zeng, 2001; Rahman and Sultana, 2017). Alternatively, L2-regularization can be used to impose a Gaussian prior on the weights (`reg = 'l2'`). In this case, an additional hyperparameter `lambda` $\in [0, \infty)$ that controls the amount of regularization needs to be specified by the user. It can be set to a fixed value. Alternatively, a range of candidates can be specified (e.g., `lambda = [0.001, 0.01, 0.1, 1]`). A nested cross-validation is then performed to select the optimal value. Additional details on the implementation are given in section 4.3 (**Appendix** in Supplementary Material). An alternative implementation using LIBLINEAR is also available, see Section 2.9.

2.4.6. Linear Support Vector Machine (SVM)

A SVM has no underlying probabilistic model. Instead, it is based on the idea of maximizing the margin (Hearst et al., 1998; Schölkopf and Smola, 2001). For linearly separable data, the margin is the distance from the hyperplane to the closest data point (dotted line in **Figure 2**). This distance is given by $1/||\mathbf{w}||$. Minimizing $||\mathbf{w}||$ is then equal to maximizing the margin. At the same time, one needs to make sure that the training samples are correctly classified at a distance from the hyperplane. This is achieved by requiring $\mathbf{w}^\top \mathbf{x}_i + b \geq 1$ for class 1 and $\mathbf{w}^\top \mathbf{x}_i + b \leq -1$ for class 2. Encoding the classes as +1 and −1, both terms can be combined into $y_i\,(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$. This constraint cannot be satisfied for every training sample $i \in \{1, ..., n\}$ if the data cannot be perfectly separated. Therefore, positive slack variables ξ_i are introduced that allow for misclassifications. Now the goal becomes to maximize the margin while simultaneously minimizing the amount of constraint violations given by $\sum_i \xi_i$. Put together, this leads to the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + c\sum_i \xi_i \\ &\text{subject to } \forall i : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\ &\forall i : \xi_i \geq 0 \end{aligned} \tag{6}$$

The resultant classifier, called two-class L1-Support Vector Machine (SVM) is implemented as `'svm.'` The hyperparameter `c` controls the amount of regularization and needs to be set by the user. Despite the lack of a probabilistic model, a Platt approximation using an external function (`http://www.work.caltech.edu/htlin/program/libsvm/`) is used to estimate class probabilities if required. Additional details on the implementation are given in section 4.4 (**Appendix** in Supplementary Material). Alternative implementations using LIBSVM and LIBLINEAR are also available, see section 2.9.

2.4.7. Kernel Classifiers

In kernel methods such as SVM and kernel FDA, a sample is implicitly mapped from the input space $\mathcal{X}$ into a high-dimensional feature space $\mathcal{F}$ using a map $\phi : \mathcal{X} \rightarrow \mathcal{F}$. As illustrated in **Figure 2**, such a map can translate a non-linear classification problem into a linear problem in feature space (Schölkopf and Smola, 2001). For two classes, decision values are given by

$$\mathrm{dval} = \mathbf{w}_\phi^\top \phi(\mathbf{x}) + b \tag{7}$$

where $\mathbf{w}_\phi$ is the weight vector in feature space. If we compare this formula to Equation (1), it becomes evident that kernel classifiers are linear classifiers acting on non-linear transformations of the features. Often, it is infeasible to explicitly apply the map due to the high dimensionality of $\mathcal{F}$. However, for methods such as SVM

and LDA, an efficient workaround is available. The optimization problem can be rewritten into a form wherein only the inner products between pairs of samples are needed, i.e., $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ for samples $\mathbf{x}$ and $\mathbf{x}'$. Now, if ϕ maps to a Reproducing Kernel Hilbert Space (RKHS), these inner products can be efficiently calculated via a kernel function k that operates in input space, resulting in the identity $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$. This is known as the *kernel trick*.

To give a simple example, consider two samples with two-dimensional features, $\mathbf{x} = [x_1, x_2]$ and $\mathbf{x}' = [x_1', x_2']$. The homogeneous polynomial kernel of degree 2 has the kernel function $k(\mathbf{x}, \mathbf{x}') = (\sum_{i=1}^{2} x_i x_i')^2$ and the corresponding feature map $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ with $\phi(\mathbf{x}) = [x_1^2, \sqrt{2} x_1 x_2, x_2^2]$. It is now easily verified that $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$. For LDA, a kernelized version called Kernel Fisher Discriminant Analysis (KFDA) has been developed by Mika et al. (1999). It is available as `'kernel_fda.'` By default, the model is regularized using shrinkage regularization controlled by the hyperparameter `lambda`. Often, a small value (e.g., `lambda = 0.01`) is adequate. Additional details on the implementation are given in section 4.5 (**Appendix** in Supplementary Material). For kernel SVM, either `'svm'` or the LIBSVM interface can be used. For both SVM and KFDA, the kernel can be chosen by setting the `kernel` parameter. Further information on the kernels is provided in the train functions.

2.4.8. Ensemble Methods

An `'ensemble'` is a meta-classifier that trains dozens or even hundreds of classifiers. In ensembles, these individual classifiers are referred to as learners. The type of learner can be set using the `learner` hyperparameter. For instance, setting `learner = 'svm'` creates an ensemble of SVM classifiers. To encourage the learners to focus on different aspects of the data, every learner is presented just a subset of the training data. `nsamples` controls the number of training samples that is randomly selected for a given learner, whereas `nfeatures` controls the number of features. The final classifier output is determined via a voting strategy. If `strategy = 'vote,'` then the class label produced by each individual learner serves as a vote. The class that receives the maximum number of votes is then selected. If `strategy = 'dval'` then the raw decision values are averaged and the final decision is taken based on whether the average is positive or negative. The latter only works with classifiers that produce decision values.

2.4.9. Classifier Output Type

For every test sample, a classifier produces raw output. This output takes either a discrete form as a *class label* or a continuous one. If it is continuous, it comes either as a *decision value* or as a *probability*. A decision value is an unbounded number that can be positive or negative. Its absolute value corresponds to the distance to the hyperplane. For two classes, the probability is a number between 0 and 1 representing the probability that a sample belongs to class 1. In the high-level interface, the classifier output can be specified explicitly by setting `cfg.output_type` to `'clabel,'` `'dval,'` or `'prob.'` In most cases, however, it suffices to let MVPA-Light infer the output type.

2.5. Regression Models

Like classifiers, regression models are implemented using pairs of train/test functions. In the high-level function `mv_regress`, a regression model is specified using the `cfg.model` parameter. Low-level access is possible by directly calling the train/test functions. For instance, `model = train_ridge(param, X, y)` trains a ridge regression model. `X` is the training data and `y` are the corresponding responses. `param` is a MATLAB struct that contains hyperparameters. The output `model` is a struct that contains the model parameters after training. The model can be applied to test data by calling `yhat = test_ridge(model, Xtest)` where `Xtest` is test data. The output of the test function is the model predictions. In the following section, the individual regression models are introduced. It is assumed that the training data is contained in matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of n samples and p predictors. The i-th row of this matrix is denoted as the column vector $\mathbf{x}_i \in \mathbb{R}^p$. Responses are stored in a vector $\mathbf{y} \in \mathbb{R}^n$ with $\mathbf{y}_i$ referring to the i-th response.

2.5.1. Perspective: Linear Regression

Linear models such as Linear Regression, Ridge Regression, and linear Support Vector Regression, act on the data in a unified way by means of a vector of coefficients $\mathbf{w}$ (often represented by β's in the literature). Linear regression models differ only in the way that $\mathbf{w}$ is derived. To simplify the notation, it is assumed that the data matrix $\mathbf{X}$ contains a column of ones and hence the intercept term is contained in $\mathbf{w}$. For a test sample $\mathbf{x}$, the predicted response is given by $\widehat{y} = \mathbf{w}^\top \mathbf{x}$. The vector of predicted responses on the training data $\widehat{\mathbf{y}} \in \mathbb{R}^n$ can be written in matrix notation as

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w} \tag{8}$$

During training, the goal is to find a $\mathbf{w}$ such that $y_i \approx \widehat{y}_i$ for each training sample. A natural measure of closeness between the true response and the prediction is the squared distance $(y_i - \widehat{y}_i)^2$, which directly leads to the sum of squares measure $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$. In matrix notation, the sum of squares is denoted as

$$\mathcal{L}_{\mathrm{OLS}}(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 \tag{9}$$

The solution that minimizes this quantity, known as ordinary least squares (OLS) solution to linear regression, is given by $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. It is worth noting that if one divides the sum of squares by the number of samples n, one obtains the regression metric *mean squared error* (MSE).

2.5.2. Ridge Regression

Ridge regression is a regularized version of OLS regression. It is useful for data that suffers from multicollinearity. The model is regularized by adding a L2 penalty that shrinks the weights toward zero. For a given regularization parameter `lambda`$\in [0, \infty)$, denoted by the Greek symbol λ, the loss function is given by

$$\mathcal{L}_{\mathrm{ridge}}(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda ||\mathbf{w}||^2 \tag{10}$$

This convex optimization problem can be solved directly by calculating the gradient and setting it to zero. Alternatively, it can be rewritten into its dual Lagrangian form first (Bishop, 2007). The resultant primal and dual ridge solutions that minimize the loss function are given by

$$\begin{aligned} \mathbf{w} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1}\, \mathbf{X}^\top \mathbf{y} \quad \text{(primal solution)} \\ &= \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_n)^{-1}\, \mathbf{y} \quad \text{(dual solution)} \end{aligned} \tag{11}$$

where $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ and $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ are identity matrices. The equivalence between the primal and dual solution can be verified by left-multiplying both solutions with $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)$.

For `lambda` $= 0$ ridge regression reduces to OLS regression. By default (`form = 'auto'`), MVPA-Light dynamically switches between the primal and the dual form depending on whether n is larger or smaller than p.

2.5.3. Kernel Ridge Regression

Analogous to kernel classifiers (section 2.4.7), a non-linear version of ridge regression can be developed by applying a non-linear transformation to the features. Let this transformation be represented by $\phi : \mathcal{X} \rightarrow \mathcal{F}$, a map from input space to a Reproducing Kernel Hilbert Space, and $\Phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_n)]^\top$. The solution is given by replacing $\mathbf{X}$ by $\Phi(\mathbf{X})$ in Equation (11),

$$\begin{aligned} \mathbf{w}_\phi &= (\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda \mathbf{I})^{-1}\, \Phi(\mathbf{X})^\top \mathbf{y} \quad \text{(primal solution)} \\ &= \Phi(\mathbf{X})^\top (\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda \mathbf{I}_n)^{-1}\, \mathbf{y} \quad \text{(dual solution)} \end{aligned} \tag{12}$$

Unfortunately, this solution is of limited practical use, since generally speaking the feature space is too high-dimensional to represent $\mathbf{w}_\phi$ and $\Phi(\mathbf{X})$. However, the dual solution can be rewritten as follows. Let $\mathbf{K} = \Phi(\mathbf{X})\, \Phi(\mathbf{X})^\top$ be the kernel matrix with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for a kernel function k. Define the vector of dual weights $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\, \mathbf{y}. \tag{13}$$

Then the predicted response to a test sample $\mathbf{x}$ can be rewritten in terms of kernel evaluations:

$$f(\mathbf{x}) = \mathbf{w}_\phi^\top\, \phi(\mathbf{x}) = \boldsymbol{\alpha}^\top \Phi(\mathbf{X})\, \phi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i\, k(\mathbf{x}_i, \mathbf{x}). \tag{14}$$

2.6. Performance Metrics

In most cases, the quantity of interest is not the raw model output but rather a metric that summarizes the performance of the classifier or regression model on test data. The desired metric can be specified by e.g., setting `cfg.metric = 'accuracy'` in any high-level function. Multiple metrics can be requested by providing a cell array, e.g., `cfg.metric = {'accuracy','auc'}`. **Table 1** lists the metrics implemented in MVPA-Light. For a thorough discussion of classification metrics, refer to Sokolova and Lapalme (2009).

If cross-validation is used then the metric is initially calculated for each test set in each repetition separately. It is then averaged across test sets and repetitions. Since the number of samples in a test set can vary across different folds, a proportionally weighted average is used whereby larger test sets get a larger weight.

2.7. Statistical Analysis

In neuroimaging experiments, establishing the statistical significance of a metric is often more important than maximizing the metric *per se*. Neuroimaging data is typically hierarchical: a study comprises many subjects, and each subject comprises many trials. To perform group analysis, a common approach is then to start with a level 1 (single-subject) analysis and calculate a classification or regression metric. At this stage, the samples consist of single trials for a particular subject. The metrics are then taken on to level 2 (group level). At this stage, each subject constitutes one sample (Mumford and Poldrack, 2007). The function `mv_statistics` implements both level 1 (single-subject) and level 2 (group level) statistical analysis. For level 1 analysis, the following tests are available:

- Binomial test: uses a binomial distribution to calculate the p-value under the null hypothesis that classification accuracy is at chance. Requires classification accuracy as metric.
- Permutation test: non-parametric significance test. Creates a null distribution by shuffling the class labels or responses and repeating the multivariate analysis e.g., 1,000 times.
- Cluster permutation test: an elegant solution to the multiple comparisons problems arising when MVPA is performed along multiple dimensions (e.g., for each time-frequency point). Uses the cluster statistic introduced in Maris and Oostenveld (2007).

For level 2 analysis, a permutation test (with and without cluster correction) is available for within-subject and between-subjects designs. Note that no classification/regression is performed. The metrics that have been obtained in the level 1 analysis for each subject are simply subjected to a standard statistical test. In the *within-subject* design, two different cases are considered. If pairs of values have been observed (e.g., mean decision values for class 1 and 2) they are tested for a significant difference across subjects. If only one value has been observed (e.g., AUC) it is tested against a given null value (e.g., 0.5). As test statistics, mean, t-test, or Wilcoxon signed-rank test can be used. To create a null distribution, data is permuted by randomly swapping the pairs of values or swapping the value and its null value. In *between-subjects* design, subjects are partitioned into two different groups. The test statistic quantifies whether the metric differs between two groups. A null distribution is created by randomly assigning subjects to groups.

To illustrate this with an example, consider the "faces vs. houses" experiment. For the within-subject design, assume the mean decision values for houses and faces have been determined for each subject using cross-validation. A paired-samples t-test across subjects comparing the decision value for faces vs houses is used to calculate a t-statistic. A null distribution is created by

TABLE 1 | Metrics in MVPA-Light.

Task	Metric	Range	Description
Classification	`'accuracy'`	[0,1]	Fraction correctly predicted class labels.
	`'auc'`	[0,1]	For two classes only. An alternative to classification accuracy that is more robust to imbalanced classes. Requires continuous classifier output (decision values or probabilities). 0.5 means chance-level performance and 1 means perfect separation of the classes.
	`'confusion'`	[0,1]	Confusion matrix. Rows corresponds to true class, columns to predicted class. The (i,j)-th element gives the proportion of samples of class i that have been classified as class j.
	`'dval'`	$(-\infty, +\infty)$	For two classes only. Average decision value, for each class separately.
	`'f1'`	[0,1]	Combines precision (PR) and recall (R) into a single score using the harmonic average 2*PR*R / (PR+R).
	`'kappa'`	[-1, 1]	Cohen's kappa, a measure of inter-rater reliability.
	`'precision'`	[0,1]	TP / (TP + FP). Fraction of samples labeled as positive that actually belong to the positive class. For multi-class, it is calculated per class from the confusion matrix.
	`'recall'`	[0,1]	TP / (TP + FN). Fraction of positive samples that have been detected. For multi-class, it is calculated per class from the confusion matrix.
	`'tval'`	$(-\infty, +\infty)$	For two classes only. T-test statistic for the unequal sample size, equal variance case, based on decision values.
	`'none'`	$(-\infty, +\infty)$	Returns a cell array with the raw classifier outputs for all test sets.
Regression	`'mae'`	$[0, \infty)$	Mean absolute error: $1/n \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$.
	`'mse'`	$[0, \infty)$	Mean squared error: $1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.
	`'r_squared'`	$(-\infty, 1]$	R^2 coefficient representing the fraction of variance explained by the model.

TP, true positives; FP, false positives; FN, false negatives. Regression: y = responses, $\hat{y}$ = model predictions.

randomly swapping face and house values for each subject and recomputing the statistic. For a between-subjects design, assume the experiment has also been carried out with a clinical group of Parkinson's patients and AUC values have been recorded for both groups. A Wilcoxon rank sum test is used to compare the AUC for the two groups at each voxel. A null distribution is created by randomly assigning subjects to either the clinical or the control group.

2.8. Custom Classifiers and Regression Models

MVPA-Light can be extended with custom models. To this end, the appropriate train and test functions need to be implemented. Additionally, default hyperparameters need to be added to the function `mv_get_hyperparameter`. In the **Appendix**, it is shown how to implement a prototype classifier that assigns a sample to the closest class centroid.

2.9. LIBSVM and LIBLINEAR

LIBSVM (Chang et al., 2011) and LIBLINEAR (Fan et al., 2008) are two high-performance libraries for SVM, Support Vector Regression (SVR), and Logistic Regression. In order to use the libraries with MVPA-Light, the user needs to follow the installation instructions on the respective websites. In particular, the C-code needs to be compiled and added to the MATLAB path. In MVPA-Light, the models are denoted as `'libsvm'` and `'liblinear.'`

2.10. FieldTrip Integration

The FieldTrip (Oostenveld et al., 2011) function `ft_statistics_mvpa` provides a direct interface between FieldTrip and MVPA-Light. In brief, the function calls MVPA-Light functions to carry out multivariate analysis, and then stores the results back into FieldTrip structs. To use MVPA-Light from high-level FieldTrip functions such as `ft_timelockstatistics`, one has to set the parameter `cfg.method = 'mvpa.'` The interface is introduced in detail in a tutorial on the FieldTrip website [1].

2.11. Development

To maintain the integrity of the toolbox, the `unittests/` subfolder features a unit testing framework for all models, optimization algorithms, high-level functions and some of the important utility functions. The unit tests make use of both the example EEG data, random noise, and simulated data. Unit testing can be triggered by executing the `run_all_unittests` function.

2.12. Analysis of a MEEG Dataset

To illustrate MVPA-Light on a real dataset, a multivariate analysis was conducted on a multi-subject, multi-modal face processing dataset wherein subjects viewed images of famous faces, familiar faces, or scrambled faces. See Wakeman and Henson (2014, 2015) for a detailed description of the data. The dataset contains 16 subjects with EEG and MEG simultaneously recorded. The MEEG data was preprocessed using FieldTrip. It was low-pass filtered with a cut-off of 100 Hz and high-pass filtered using a FIR one-pass zero-phase filter with a cut-off of 0.1 Hz. A bandstop filter was applied at 50 Hz to suppress line noise. Subsequently, data was downsampled to 220 Hz and for each subject, the 6 separate runs were combined into a single dataset, yielding 880–889 trials per subject with roughly equal proportions for the three classes. All trials displaying famous faces were coded as class 1, familiar faces as class 2, and scrambled faces as class 3. MVPA was performed to investigate the following questions:

[1] `http://www.fieldtriptoolbox.org/tutorial/mvpa_light/`

1. ERP classification: Wakeman and Henson (2015) found two prominent event-related components, a N170 and a sustained component roughly starting at 400 ms post-stimulus. Cross-validation with a multi-class classifier was used to investigate whether these components discriminate between the three classes.
2. Time classification: Is there more discriminative information in MEG than in EEG? To answer this, classification across time was performed for three different channel sets, namely EEG only, MEG only, and EEG+MEG combined.
3. Time-frequency classification: Is the discriminative information for famous vs scrambled faces confined to specific oscillatory frequencies and times? To answer this, time-frequency spectra were calculated for single trials and classification was performed at each time-frequency bin separately.
4. Generalization: Are representations shared across time (King and Dehaene, 2014) or frequency? To answer this, time generalization (time x time classification) was applied to the ERF data, and frequency generalization (frequency x frequency classification) was applied to the time-frequency data.

MVPA was performed at the sensor level using a LDA classifier. All analyses were cross-validated using 5- or 10-fold cross-validation. Only the MEG channels were used as features except for analysis 2, where different sets of channels were compared. To assess statistical significance, the following tests were carried out:

- Level 1 statistics. For each subject, the statistical significance of the time generalization (famous vs. scrambled faces) was investigated. For illustrative purposes, the three statistical tests contained in MVPA-Light were compared: binomial, permutation, and cluster permutation tests. Permutation tests were based on 500 random permutations of the class labels. The cluster permutation test was corrected for multiple comparisons by using a cluster statistic, the other tests were uncorrected. For the cluster statistic, a critical value of 0.6 was chosen for classification accuracy. This analysis is reported only for the first subject.
- Level 2 statistics (across subjects). The AUC values obtained in the time-frequency classification analyses were statistically compared to a null value of 0.5 using cluster permutation tests based on a within-subject design.

2.13. Analysis of a fMRI Dataset

To illustrate the application of MVPA-Light to fMRI data, another analysis was conducted using a block-design fMRI study. See Haxby et al. (2001) for a detailed description. The dataset was downloaded from `http://www.pymvpa.org/datadb/haxby2001.html`. The study investigates face and object representations in human ventral temporal cortex. It comprises 6 subjects with 12 runs per subject. In each run, subjects viewed grayscale images of 8 living and non-living object categories, grouped in 24 s blocks separated by rest periods. Images were shown for 500 ms followed by a 1,500 ms inter-stimulus interval. Full-brain fMRI data were recorded with a volume repetition time of 2.5 s. Hence, a stimulus block was covered by roughly 9 volumes. A zero-phase Butterworth high-pass filter with a cut-off frequency of 0.01 Hz was applied in order to remove slow drifts. No other preprocessing was performed. The following questions were addressed:

1. Confusion matrix: Which image categories lead to similar brain activation patterns?
2. Time classification: How does classification performance evolve across time following stimulus onset?
3. Searchlight analysis: Which of the brain regions contain discriminative information that discerns between faces and houses?

Leave-one-run-out cross-validation was used to calculate classification performance. Multi-class LDA with 8 classes served as a classifier. For the searchlight analysis, binary LDA contrasting faces vs. houses was used with AUC serving as metric. The searchlight consisted of a 3x3x3 cube of voxels that was centered on each target voxel. A level 2 cluster permutation test was computed on the AUC values against the null hypothesis that AUC equals 0.5.

2.14. Benchmarking

Multivariate analyses can involve hundreds or even thousands of train/test iterations. Therefore, training time (the amount of time required to train a single model on data) is a relevant quantity when evaluating different model implementations. To benchmark MVPA-Light's models, their training time was compared to models in the MATLAB Statistics Toolbox as well as models in Python (Scikit Learn package) and R (different packages). The comparison to other MVPA toolboxes is of less relevance since they often rely on external packages such as LIBSVM and LIBLINEAR which are also available in MVPA-Light (this applies to e.g., DDTBOX, PRoNTo, TDT). The following three datasets were considered:

- MEG single-subjects. The Wakeman and Henson (2015) dataset was used with the famous vs. scrambled faces conditions, epoched in the range [−0.2, 1] s. Data dimensions were 585–592 trials per subject, 306 channels, and 265 time points. MVPA was performed for every subject and every time point separately, using channels as features.
- MEG super-subject. Trials of all subjects in the MEG single-subjects data were concatenated to form a single "super-subject" comprising 9,421 trials, 306 channels, and 265 time points. MVPA was performed for every time point separately, using channels as features.
- fMRI. For each subject in the Haxby et al. (2001) data, all voxels with a non-zero signal were concatenated to a single feature vector. The time dimension was dropped, different time points within a trial were simply considered as different samples. The two classes "face" and "house" were considered, yielding a data matrix of 216 samples (198 samples for subject 5) and between 163,665 and 163,839 voxels per subject. MVPA was performed for every subject separately, using voxels as features.

The MEG single-subjects dataset is of standard size for neuroimaging data and thus serves as a benchmark for ordinary

operation. The other two datasets are intended to test the computational limits of the models by using either a large number of trials (MEG super-subject) or a large number of features (fMRI). For the single-subjects dataset, classification performance was measured in addition to training time. To be as unbiased as possible, hyperparameters were mostly unchanged except when a change made the models more comparable across toolboxes (e.g., setting the same regularization value). No hyperparameter tuning was performed in order to quantify pure training time.

The MVPA-Light models were compared to LIBSVM, LIBLINEAR, and MATLAB 2019a's fitcdiscr (LDA), lassoglm (LogReg), fitcnb (Naive Bayes), fitcsvm (SVM), ridge, and fitrsvm (SVR). Python and R-based toolboxes were installed in virtual environments using Anaconda 4.7.12. Scikit Learn 0.21.2 was used together with Python 3.7.3. R version 3.6.1 was used with packages MASS (LDA), glmnet (LogReg and Ridge), e1071 (Naive Bayes, SVM, SVR), and listdtr (Kernel Ridge).

For the single-subject data, the timing results were averaged across subjects. Then for both the single-subject and the super-subject, mean and standard deviation was calculated across time points. For the fMRI data, mean and standard deviation was calculated across subjects. All analyses were conducted after a fresh restart of a desktop computer with networking disabled. The computer had an Intel Core i7-6700 @ 3.40 GHz x 8 CPU with 64 GB RAM running on Ubuntu 18.04. All scripts are available in the accompanying GitHub repository[2].

2.15. Results

2.15.1. MEEG Data

Figure 3 depicts the results of the MVPA, averaged across subjects. Errorbars depict standard error across subjects.

ERP classification (**Figure 3A**). The bar graph shows that for both the N170 and the sustained ERP component classification accuracy is significantly above the chance level of 33%. Accuracy can be broken down into confusion matrices that show which combinations of classes get misclassified ("confused"). For both N170 and the sustained ERP component, the highest accuracy is obtained for the scrambled images (0.63 and 0.78). Moreover, misclassification (off-diagonal elements) is most prominent for the famous and unfamiliar faces. This is not surprising since both types of images are identical in terms of low-level features and both show actual faces, in contrast to the scrambled images.

Time classification (**Figure 3B**). The classes are not discriminable prior to the occurrence of the N170. A classification peak at the time of the N170 can be seen for all channel sets. At this stage, the AUC values diverge, with EEG yielding a significantly lower AUC. Combining EEG+MEG seems to yield a slightly higher performance than MEG alone.

Time-frequency classification (**Figure 3C**). For famous vs scrambled faces, peak performance is reached in the delta frequency band at a latency between 0.2 and 0.4 s. For famous vs unfamiliar faces, peak performance is attained in the latter half of the trial (0.5–1 s) in the theta and alpha frequency bands.

Generalization (**Figure 3D**). The first plot depicts AUC (color-coded) as a function of training time (y-axis) and testing time (x-axis). There is evidence for widespread time generalization for famous vs scrambled faces starting about at the time of the N170 peak and covering most of the remaining trial. In particular, there is generalization between the N170 and the later sustained component (horizontal and vertical lines emanating at 0.17 s), suggesting some correlation between the spatial pattern of the N170 and the sustained component. The second plot depicts AUC as a function of frequency. There is some generalization in the theta band (lower-left corner), the alpha band, and the lower beta band (16–22 Hz). Also, when the classifier is trained in the beta band, classification performance partially generalizes to the alpha band. However, the overall performance is low when compared to the time-locked data.

Level 2 statistics (**Figure 3E**). Group statistical analysis based on the time-frequency classification data in the panel above. Images depict AUC values masked by significance (deep blue = not significant). For the famous vs. scrambled faces classification, a large cluster spanning the whole trial and especially the low frequency bands is evident. For the famous vs. unfamiliar faces condition, there is a significant cluster corresponding to large AUC values evident after 0.5 s and confined to the lower frequency range.

Level 1 statistics (**Figure 3F**). Level 1 statistical analysis based on the time generalization data in the panel above, shown exemplarily for subject 1. Images depict the AUC values masked by significance. Both uncorrected tests (binomial and permutation test) exhibit spurious effects even at pre-stimulus time. Most of these spurious effects disappear under the cluster permutation test.

2.15.2. fMRI Data

Figure 4 depicts the results of the MVPA on the fMRI data, averaged across subjects.

Confusion matrix (**Figure 4A**). A mask provided with the data was applied to select voxels from ventral temporal areas. A high overall performance is observed for LDA with 8 classes. Misclassifications tend to be confined to general semantic categories. For instance, misclassified faces tend to be labeled as cats (both living objects), whereas misclassified non-living objects tend to be labeled as other non-living objects. This indicates that there are shared representations for images from the same general category.

Time classification (**Figure 4B**). Although all ROIs and time points yield performances above the chance level of 12.5%, the ventral temporal area (which comprises both face and house responsive voxels) yields the best performance. For the latter, classification performance peaks at about 5 s after stimulus onset.

Searchlight analysis (**Figure 4C**). AUC values averaged across subjects are depicted. The AUCs are masked by the significant cluster ($p < 0.01$) and overlayed on an averaged anatomical MRI. Although the cluster is large, high values

[2]https://github.com/treder/MVPA-Light-Paper

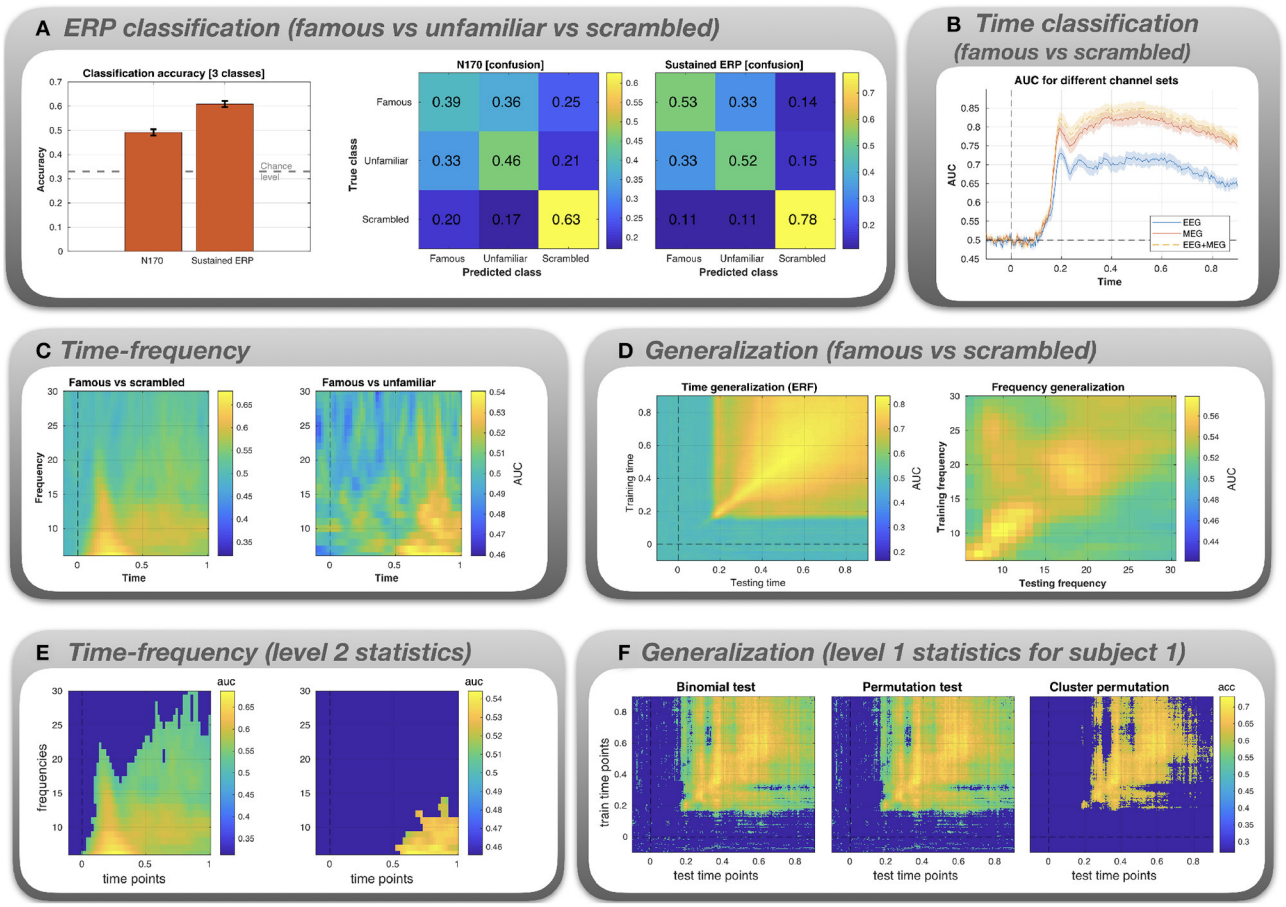


FIGURE 3 | Results for the classification analysis of the Wakeman and Henson (2015) MEEG data. **(A)** Multi-class classification (famous vs. unfamiliar vs. scrambled faces) of N170 and sustained ERP component. **(B)** AUC is plotted as a function of time for famous vs. scrambled images. The classification was performed using three different channel sets: EEG only, MEG only, and EEG+MEG combined. **(C)** Binary classification (famous vs. scrambled and famous vs unfamiliar) for time-frequency data. AUC is plotted as a function of both time and frequency. The AUC values are color-coded. **(D)** Time x time generalization and frequency x frequency generalization using a binary classifier (famous vs. scrambled). **(E)** Level 2 statistical analysis of the time-frequency classification. **(F)** Level 1 statistical analysis of the time x time generalization, shown exemplarily for subject 1.

$>$ 0.8 are predominantly found in dorsal and ventral visual areas including the paraphippocampal place area and the fusiform area, nicely dovetailing with the original findings of Haxby et al. (2001).

2.15.3. Benchmarking

Figure 5 depicts ERP classification accuracy across time on the MEG single-subjects data for different classifiers and different toolboxes, averaged across subjects. Except for the MATLAB classifiers, results are nearly identical for all implementations of LDA, LogReg, and linear SVM, with a peak performance of about 75%. Lower performance is evident for Naive Bayes, but consistently so across different implementations. For SVM with a RBF kernel, the best performance is obtained in R, followed MATLAB, with both MVPA-Light and Scikit Learn performing worse. Since no hyperparameter tuning was performed, the latter result is most likely due to differences in the default hyperparameters.

Tables 2, **3** show the timing results for different classifiers and regression models. These results are discussed model by model:

LDA. The MVPA-Light implementation consistently outperforms other implementations in terms of training time, in some cases by orders of magnitude. For the fMRI dataset, it is almost 100 times faster than Scikit Learn, whereas MATLAB and R both run out of memory. It is worth noting that a shrinkage value of 0.01 was applied for the MVPA-Light and MATLAB implementations. For R, low performance was achieved with rda (regularized LDA), so the standard unregularized LDA was used. For Scikit Learn, the default solver does not allow for shrinkage so no shrinkage was applied.

LogReg. The MVPA-Light implementation of Logistic Regression outperforms the competitors for the MEG single-subjects data. It is outperformed by the R implementation for the MEG super-subject. For the fMRI data, it causes an out of memory error and the best performing model is LIBLINEAR.

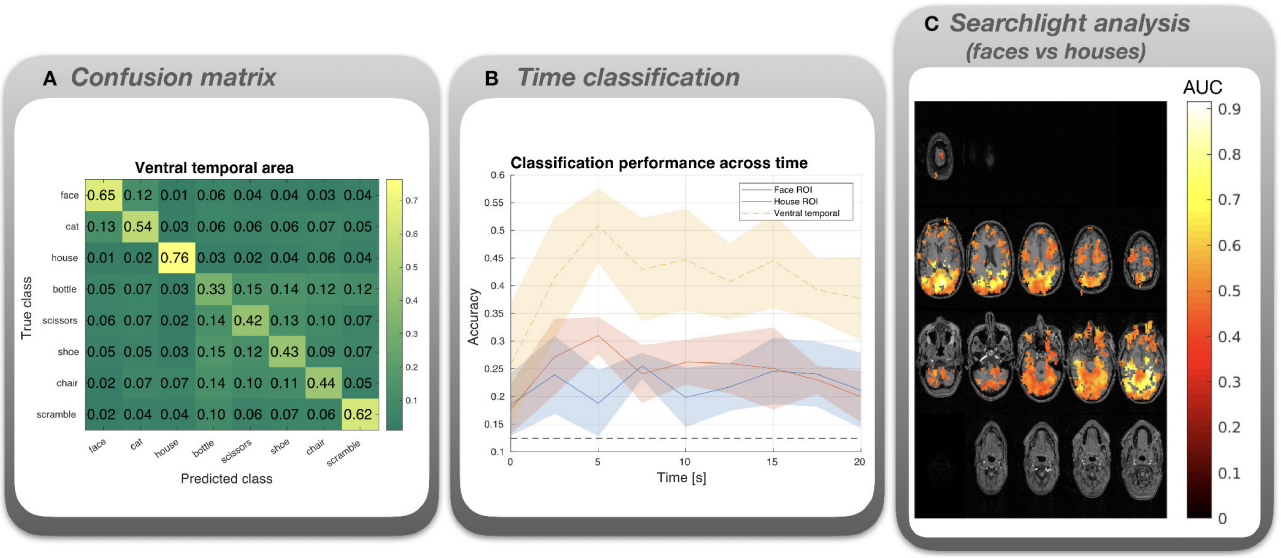


FIGURE 4 | Results for the classification analysis of the Haxby et al. (2001) fMRI data. **(A)** Confusion matrix for multi-class (8 classes) classification based on voxels in the ventral temporal area, averaged across subjects. **(B)** Multi-class (8 classes) classification accuracy was calculated for each time point following stimulus onset. Lines depict means across subjects, shaded areas correspond to standard error. Masks were used to select voxels in the ventral temporal area (yellow line), voxels responsive to faces (blue), or voxels responsive to houses (red). **(C)** Cluster permutation test results based on a searchlight analysis using a binary classifier (faces vs houses). Red spots represent AUC values superimposed on axial slices of the averaged structural MRI. All depicted AUC values correspond to the significant cluster; other AUC values have been masked out.

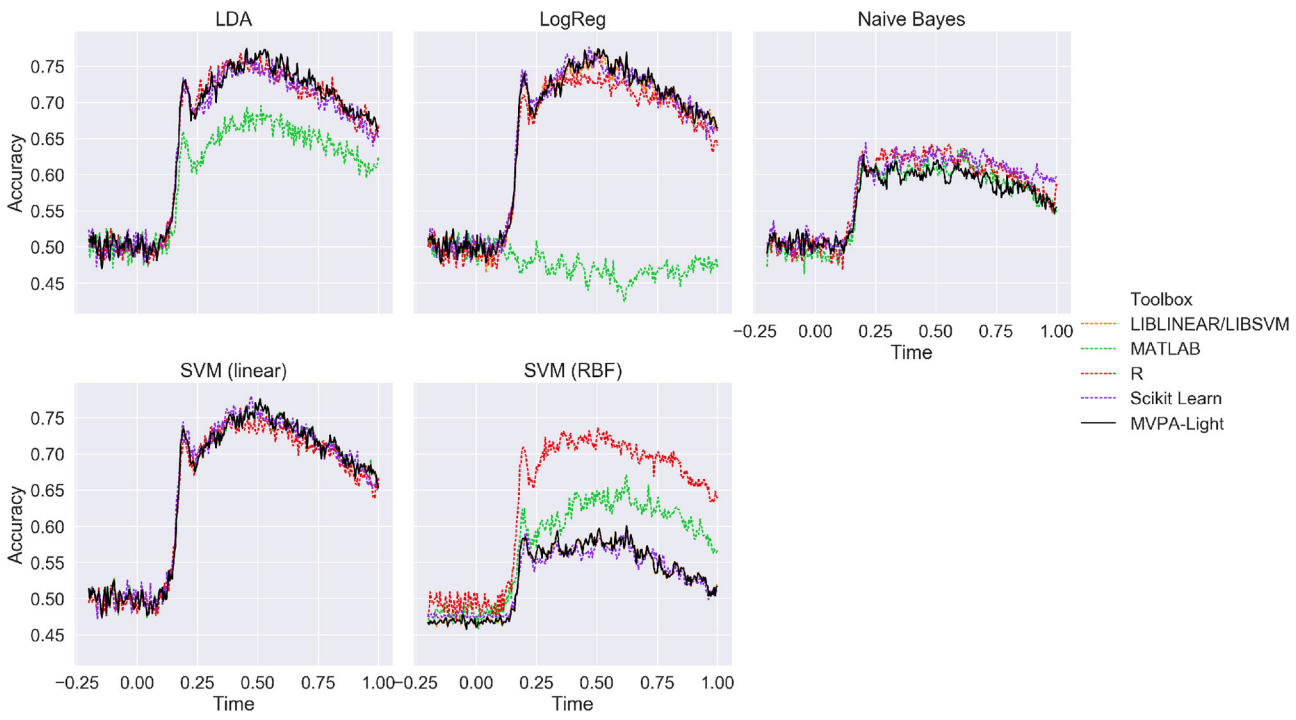


FIGURE 5 | Mean ERP classification accuracy for the benchmarking analysis using the MEG single-subjects data (averaged across subjects). MVPA-Light is depicted as a solid black line.

Naive Bayes. The MVPA-Light implementation consistently outperforms other implementations, in some cases by orders of magnitude. Scikit Learn is consistently second best, followed by R and MATLAB.

SVM. For linear SVM, LIBLINEAR yields the best training speed except for the fMRI data, where MVPA-Light performs best. For RBF kernels, MVPA-Light's SVM consistently outperforms the competitors, closely followed by MATLAB's

TABLE 2 | Benchmarking results: mean training time and standard deviation in seconds for different classifiers.

Dataset	Toolbox	Classifier				
		LDA	**LogReg**	**Naive Bayes**	**SVM (linear)**	**SVM (RBF)**
MEG single-subjects	MVPA-Light	**0.003** ± 0.0001	**0.0097** ± 0.0005	**0.001** ± 0.00004	0.07 ± 0.002	**0.02** ± 0.0001
	LIBLINEAR	–	0.014 ± 0.0009(*p*) 0.035 ± 0.001(*d*)	–	**0.023** ± 0.002(*p*) 0.231 ± 0.02(*d*)	–
	LIBSVM	–	–	–	0.098 ± 0.01	0.125 ± 0.001
	MATLAB	0.026 ± 0.0008	0.03 ± 0.006	0.05 ± 0.0001	0.041 ± 0.004	0.023 ± 0.0004
	Scikit Learn	0.097 ± 0.0006	0.1 ± 0.005	0.007 ± 0.0001	0.37 ± 0.052	0.45 ± 0.032
	R	0.084 ± 0.0003	0.013 ± 0.002	0.04 ± 0.0001	0.71 ± 0.113	0.41 ± 0.026
MEG super-subject	MVPA-Light	**0.026** ± 0.0028	0.437 ± 0.0062	**0.015** ± 0.0001	10.122 ± 1.05	**5.369** ± 0.033
	LIBLINEAR	–	0.732 ± 0.068(*p*) 0.998 ± 0.063(*d*)	–	**1.338** ± 0.168(*p*) 6.29 ± 0.519(*d*)	–
	LIBSVM	–	–	–	42.089 ± 4.188	37.941 ± 0.404
	MATLAB	0.149 ± 0.002	0.279 ± 0.137	0.231 ± 0.027	20.98 ± 1.78	11.65 ± 0.217
	Scikit Learn	0.596 ± 0.017	2.065 ± 0.109	0.09 ± 0.001	32.19 ± 2.07	34.56 ± 0.38
	R	0.84 ± .004	**0.159** ± 0.018	0.144 ± .0006	1123.16 ± 27.39	123.31 ± 9.38
fMRI	MVPA-Light	**0.293** ± 0.0078	OOM	**0.309** ± 0.011	**0.182** ± 0.0086	**2.064** ± 0.235
	LIBLINEAR	–	**4.008** ± 0.627(*p*) 6.689 ± 1.018(*d*)	-	2.235 ± 0.218(*p*) 6.125 ± 0.995(*d*)	-
	LIBSVM	–	–	–	11.79 ± 0.787	11.88 ± 0.822
	MATLAB	OOM	23.79 ± 4.008	357.49 ± 2.205	5.053 ± 0.325	4.845 ± 0.308
	Scikit Learn	24.45 ± 1.1	20.68 ± 4.24	2.86 ± 0.06	10.46 ± 0.59	9.15 ± 0.59
	R	OOM	7.1 ± 1.13	18.48 ± 0.35	39.67 ± 1.98	43.3 ± 2.18

For each combination of dataset and classifier, the fastest model is marked in bold. OOM, out of memory error; (p), primal form; (d), dual form.

TABLE 3 | Benchmarking results: mean training time and standard deviation in seconds for different regression models.

Dataset	Toolbox	Regression model			
		Ridge	**Kernel Ridge**	**SVR (linear)**	**SVR (RBF)**
MEG single-subjects	MVPA-Light	**0.0016** ± 0.00006	**0.019** ± 0.0001	–	–
	LIBSVM	–	–	0.02 ± 0.001	**0.0041** ± 0.0002
	MATLAB	0.0061 ± 0.0002	–	**0.018** ± 0.037	0.023 ± 0.0005
	Scikit Learn	0.0069 ± 0.0003	0.023 ± 0.003	0.654 ± 0.0647	0.481 ± 0.02
	R	0.055 ± 0.0027	–	1.59 ± 0.094	0.43 ± 0.002
MEG super-subject	MVPA-Light	**0.015** ± 0.001	**7.38** ± 0.023	–	–
	LIBSVM	–	–	**0.653** ± 0.038	**0.121** ± 0.014
	MATLAB	0.186 ± 0.007	–	6.931 ± 0.237	9.9798 ± 0.239
	Scikit Learn	0.062 ± 0.005	14.51 ± 0.21	3.213 ± 0.394	31.61 ± 1.51
	R	0.547 ± 0.0079	–	465.08 ± 49.83	151.66 ± 26.76
fMRI	MVPA-Light	**0.165** ± 0.0042	2.026 ± 0.256	–	–
	LIBSVM	–	–	**4.334** ± 1.48	**2.819** ± 0.0412
	MATLAB	OOM	–	4.545 ± 0.353	4.563 ± 0.284
	Scikit Learn	0.638 ± 0.022	**0.476** ± 0.01	16.138 ± 3.64	9.999 ± 0.59
	R	7.503 ± 0.593	–	37.211 ± 2.056	41.037 ± 2.298

For each combination of dataset and model, the fastest model is marked in bold. OOM, out of memory error; (p), primal form; (d), dual form.

fitcsvm. Significant differences are obtained for different toolboxes, with R being the slowest in many cases. The good performance of MVPA-Light's SVM may appear surprising at first glance, given some of its contenders run using C code. First, MVPA-Light uses a large tolerance value; this implies that its algorithm might perform fewer iterations than LIBSVM, although this has not been investigated. If this is the case, it does not seem to be detrimental to classification performance, as **Figure 5** illustrates. Second, the advantages of LIBSVM might not play out during a single training iteration. It has an integrated

cross-validation procedure, which is likely to be substantially faster than cross-validation using MVPA-Light, although this has not been investigated either.

Ridge and Kernel Ridge. MVPA-Light's models lead the field except for the fMRI data, where Scikit Learn's kernel ridge outperforms MVPA-Light. No results are available for R's krr model; it does not appear to have an interface for fixing hyperparameters and instead performs an expensive search using leave-one-out cross-validation, so it was omitted.

SVR. MVPA-Light exclusively relies on LIBSVM for SVR, which leads the field except for one case, in which it closely trails the MATLAB implementation. Overall, R yields the slowest implementation.

3. DISCUSSION

MVPA-Light offers a suite of classifiers, regression models and metrics for multivariate pattern analysis. A high-level interface facilitates common MVPA tasks such as cross-validated classification across time, generalization, and searchlight analysis. The toolbox supports hyperparameter tuning, pre-computed kernels, and statistical significance testing of the MVPA results.

MVPA-Light also provides a nested preprocessing pipeline that applies operations to training and test sets separately. Among others, it features over- and undersampling, PCA, and scaling operations. It also includes an averaging approach wherein samples are assigned to groups and then averaged in order to increase signal-to-noise ratio. For linear classifiers, this approach has been explored by (Cichy et al., 2015; Cichy and Pantazis, 2017). Recently, it has been generalized to non-linear kernel methods (Treder, 2018). Either approach can be used in the toolbox by adding the operation `average_samples` or `average_kernel` to the preprocessing pipeline. To showcase some of its features, analyses of an MEEG (Wakeman and Henson, 2015) and an fMRI (Haxby et al., 2001) dataset are reported. The results illustrate some ways in which the toolbox can aid in quantifying the similarity of representations, measuring the information content, localizing discriminative information in the time-frequency plane, highlighting shared representations across different time points or frequencies, and establishing statistical significance.

A benchmarking analysis was conducted in order to compare MVPA-Light (including LIBSVM and LIBLINEAR) to models provided in the MATLAB Statistics Toolbox, various R packages, and Scikit Learn for Python. While classification performance is largely consistent across different platforms, training time varies considerably. The MVPA-Light implementations of LDA, Naive Bayes, and Ridge Regression consistently outperform their competitors, in some cases by orders of magnitude. For Logistic Regression and SVM, the MVPA-Light implementations and LIBLINEAR lead the field. In all but one case, MVPA-Light's classifiers are faster than the contenders in MATLAB, R, and Scikit Learn. Overall, the fastest classifier is MVPA-Light's LDA and the fastest regression model is MVPA-Light's Ridge Regression. Partially, the success of MVPA-Light is due to specialization: MVPA-Light models tend to have fewer hyperparameters than other models, and MVPA-Light features separate optimized implementations for binary LDA and multi-class LDA, whereas the other toolboxes have a single implementation. Furthermore, MVPA-Light's LDA and Ridge Regression dynamically switch between primal and dual form. This can increase computational efficiency especially when dealing with a large dataset.

The benchmarking results should not be interpreted as final verdicts on the respective toolboxes. Undoubtedly, training speed can be improved by finding an optimal set of hyperparameters for a model. For instance, increasing regularization tends to lead to smoother loss surfaces and often faster convergence for gradient descent algorithms. The strategy for the present analysis was to change default parameters minimally and, if so, only in order to increase comparability e.g., by setting a regularization parameter to a common value. Although MVPA-Light will likely perform well in other situations, too, the present results are mostly indicative of *default performance*, obtained with minimal user interference. This is a relevant measure since it is our belief that the burden of hyperparameter selection should be taken off the user as much as possible.

3.1. Setting Up a MVPA Pipeline

If one is spoilt for choice, selecting a model, metrics, and preprocessing steps can be challenging. This section offers practical advice in this regard. Such recommendations tend to be subjective to some extent, hence users are encouraged to perform their own MVPA experiments and compare different models, hyperparameter settings etc. To prevent a statistical bias, extensive experiments should not be performed on the dataset at hand. Instead, a similar dataset e.g., recorded using the same hardware with a similar paradigm can be used for experimentation.

3.2. Preprocessing the Data

Although MVPA can be applied to raw data, this may negatively affect performance, so data has ideally been cleaned and corrupted trials have been rejected. It is useful to normalize the data for numerical stability by e.g., z-scoring across trials such that each feature has mean $= 0$ and standard deviation $= 1$. This is particularly important for Logistic Regression which uses the exponential function. It also applies to LDA and kernel methods because lack of normalization can lead to results being dominated by the features with the largest scaling. Generally speaking, preprocessing operations should be nested in the cross-validation loop i.e., performed on the training set first and then applied to the test set. The `cfg.preprocess` option serves this purpose. In some cases such as demeaning, it may be admissible to perform the operation globally on the whole dataset, but one then needs to assure that there is no information leakage from the test set that could bias the results. The same argumentation applies to unsupervised techniques such as PCA. Any preprocessing steps involving the class labels, such as CSP (Blankertz et al., 2008), also need to be nested. Furthermore, for kernel methods, computation can be speeded up by precomputing the kernel matrix using `compute_kernel_matrix`, although this approach does not work when generalization is required.

3.3. Choosing a Classifier

Linear classifiers perform well in a large variety of tasks. LDA is a good default model, since it is fast and robust thanks to regularization (Blankertz et al., 2011). Logistic Regression and linear SVM are more resilient to outliers than LDA, so may be preferred for noisy or strongly non-Gaussian data. Logistic Regression has a hyperparameter-free regularization by default, hence it is more user-friendly than SVM which requires setting the hyperparameter *c*. Naive Bayes should only be used after the features have been decorrelated using PCA or ICA. For non-linear problems, kernel FDA or SVM can be used. Again, SVM requires *c* to be set, whereas for kernel FDA the default regularization often works well. Regarding the choice of a kernel, the RBF kernel is adequate for most classification tasks, but its hyperparameter `gamma` determining the kernel width might require tuning. If maximizing classification accuracy is vital, it is worth to try an ensemble of classifiers.

3.4. Choosing a Regression Model

Ridge regression tends to perform well on a variety of tasks. If the data is noisy, linear Support Vector Regression (SVR) using LIBLINEAR can be applied. If the problem is non-linear, either kernel ridge or kernel SVR using LIBSVM with a RBF kernel is recommended.

3.5. Metrics

The most common classification metric is accuracy. For multi-class problems, it is useful to complement it with a confusion matrix. For two classes, AUC is a good alternative to accuracy since it is more robust to class imbalances and invariant to shifts of the classifier threshold. When the roles of the classes are asymmetric (e.g., patients vs. controls), it is useful to report precision and recall along with their harmonic mean (F1 score). If in doubt, report multiple metrics.

3.6. Cross-Validation

Classification and regression metrics should be cross-validated. Unless the number of samples is very small, leave-one-out cross-validation should be avoided because it suffers from a large bias; instead, use 5- or 10-fold cross-validation (James et al., 2013). Since samples are randomly assigned to folds, repeating the cross-validation is recommended to get a more stable estimate.

3.7. Conclusion

MVPA-Light is a comprehensive toolbox for multivariate pattern analysis. Its models perform competitively compared to other implementations. Future development of MVPA-Light will include additional feature extraction techniques for oscillations, such as Common Spatial Patterns (Blankertz et al., 2008) and the Riemannian geometry approach (Barachant et al., 2013), and further computational improvements, such as efficient permutation testing for LDA/KFDA (Treder, 2019) and faster calculation of the regularization path for SVM (Hastie et al., 2004).

DATA AVAILABILITY STATEMENT

The MEEG dataset can be found in the OpenNeuro repository (https://openneuro.org/datasets/ds000117/versions/1.0.3).
The fMRI dataset can be found on the PyMVPA website (http://www.pymvpa.org/datadb/haxby2001.html). Scripts and figures used in this paper are available in the accompanying GitHub repository (github.com/treder/MVPA-Light-Paper).

AUTHOR CONTRIBUTIONS

MT developed the toolbox, performed all analyses and authored the manuscript.

ACKNOWLEDGMENTS

I would like to thank colleagues from the Psychology department at University of Birmingham for advice and early adaptation of the toolbox, Jan-Mathijs Schoffelen and Sophie Arana for their efforts toward integrating it into FieldTrip, and Hong-Viet Ngo and the reviewers for insightful comments on the manuscript. Many thanks to all contributors to the GitHub repository[3].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00289/full#supplementary-material

[3] `https://github.com/treder/MVPA-Light/graphs/contributors`

REFERENCES

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing* 112, 172–178. doi: 10.1016/j.neucom.2012.12.039

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2000). mlr: Machine Learning in R. *J. Mach. Learn. Res.* 17, 1–5.

Bishop, C. M. (2007). Pattern recognition and machine learning. *J. Electron. Imaging* 16:049901. doi: 10.1117/1.2819119

Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., et al. (2016). The Berlin brain-computer interface: progress beyond communication and control. *Front. Neurosci.* 10:530. doi: 10.3389/fnins.2016.00530

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K. R. (2011). Single-trial analysis and classification of ERP components - A tutorial. *Neuroimage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Magaz.* 25, 41–56. doi: 10.1109/MSP.2008.4408441

Bode, S., Feuerriegel, D., Bennett, D., and Alday, P. M. (2019). The Decision Decoding ToolBOX (DDTBOX) – A multivariate pattern analysis toolbox for event-related potentials. *Neuroinformatics* 17, 27–42. doi: 10.1007/s12021-018-9375-z

Chang, C.-C., Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Cichy, R. M., and Pantazis, D. (2017). Multivariate pattern analysis of MEG and EEG: a comparison of representational structure in time and space. *Neuroimage* 158, 441–454. doi: 10.1016/j.neuroimage.2017.07.023

Cichy, R. M., Ramirez, F. M., and Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *Neuroimage* 121, 193–204. doi: 10.1016/j.neuroimage.2015.07.011

De Cheveigné, A., and Parra, L. C. (2014). Joint decorrelation, a versatile tool for multichannel data analysis. *Neuroimage* 98, 487–505. doi: 10.1016/j.neuroimage.2014.05.068

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. New York, NY: Wiley.

Fahrenfort, J. J., van Driel, J., van Gaal, S., and Olivers, C. N. L. (2018). From ERPs to MVPA using the Amsterdam decoding and modeling toolbox (ADAM). *Front. Neurosci.* 12:368. doi: 10.3389/fnins.2018.00368

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874. doi: 10.5555/1390681.1442794

Farwell, L., and Donchin, E. (1988). Talking Off the Top of Your Head. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80:27. doi: 10.1093/bio-met/80.1.27

Grootswagers, T., Wardle, S. G., and Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: a tutorial on Multivariate pattern analysis applied to time series neuroimaging data. *J. Cogn. Neurosci.* 29, 677–697. doi: 10.1162/jocna01068

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* 5, 1391–1415. doi: 10.5555/1005332.1044706

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York Inc.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* 62, 852–855. doi: 10.1016/j.neuroimage.2012.03.016

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Hearst, M., Dumais, S., Osuna, E., Platt, J., and Schölkopf, B. (1998). Support vector machines. *IEEE Intell. Syst. Appl.* 13, 18–28. doi: 10.1109/5254.708428

Hebart, M. N., Görgen, K., and Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front. Neuroinform.* 8:88. doi: 10.3389/fninf.2014.00088

James, G., Witten, D., Hastie, T., and Tibishirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

King, G., and Zeng, L. (2001). Logistic regression in rare events data. *Polit. Anal.* 9, 137–163. doi: 10.1093/oxfordjournals.pan.a004868

King, J.-R., and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* 18, 203–210. doi: 10.1016/j.tics.2014.01.002

Kothe, C. A., and Makeig, S. (2013). BCILAB: a platform for brain–computer interface development. *J. Neural Eng.* 10:056014. doi: 10.1088/1741-2560/10/5/056014

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Kuhn, M. (2008). Building predictive models in R using the caret Package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Ledoit, O., and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *J. Portfolio Manage.* 30, 110–119. doi: 10.3905/jpm.2004.110

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K. R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Mika, S., Ratsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999). "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)* (IEEE), 41–48.

Misaki, M., Kim, Y., Bandettini, P. A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* 53, 103–118. doi: 10.1016/j.neuroimage.2010.05.051

Mumford, J. A., and Poldrack, R. A. (2007). Modeling group fMRI data. *Soc. Cogn. Affect. Neurosci.* 2, 251–257. doi: 10.1093/scan/nsm019

Mur, M., Bandettini, P. A., and Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI - An introductory guide. *Soc. Cogn. Affect. Neurosci.* 4, 101–109. doi: 10.1093/scan/nsn044

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 1–9. doi: 10.1155/2011/156869

Oosterhof, N. N., Connolly, A. C., and Haxby, J. V. (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front. Neuroinform.* 10:27. doi: 10.3389/fninf.2016.00027

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.5555/1953048.2078195

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45(1 Suppl.), S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Rahman, M. S., and Sultana, M. (2017). Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC Med. Res. Methodol.* 17:33. doi: 10.1186/s12874-017-0313-9

Schölkopf, B., and Smola, A. (2001). *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., et al. (2013). PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics* 11, 319–337. doi: 10.1007/s12021-013-9178-1

Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inform. Process. Manage.* 45, 427–437. doi: 10.1016/j.ipm.2009.03.002

Treder, M. (2019). "Direct calculation of out-of-sample predictions in multi-class kernel FDA," in *27th European Symposium on Artificial Neural Networks (ESANN)* (Bruges), 245–250.

Treder, M. S. (2018). Improving SNR and reducing training time of classifiers in large datasets via kernel averaging. *Lecture Notes Comput. Sci.* 11309, 239–248. doi: 10.1007/978-3-030-05587-523

Treder, M. S., Porbadnigk, A. K., Shahbazi Avarvand, F., Müller, K.-R., and Blankertz, B. (2016). The LDA beamformer: optimal estimation of ERP source time series using linear discriminant analysis. *Neuroimage* 129, 279–291. doi: 10.1016/j.neuroimage.2016.01.019

Treder, M. S., Purwins, H., Miklody, D., Sturm, I., and Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *J. Neural Eng.* 11:026009. doi: 10.1088/1741-2560/11/2/026009

Varoquaux, G., Raamana, P., Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145B, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Wakeman, D. G., and Henson, R. N. (2014). OpenfMRI.

Wakeman, D. G., and Henson, R. N. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Sci. Data* 2:150001. doi: 10.1038/sdata.2015.1

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Treder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fnins.2020.00599

Automatic Reconstruction of Mitochondria and Endoplasmic Reticulum in Electron Microscopy Volumes by Deep Learning

Jing Liu[1,2], *Linlin Li*[1], *Yang Yang*[3], *Bei Hong*[1,2], *Xi Chen*[1], *Qiwei Xie*[1,4]* *and Hua Han*[1,2,5]*

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Artificial Intelligence, School of Future Technology, University of Chinese Academy of Sciences, Beijing, China, [3] School of Life Science and Technology, ShanghaiTech University, Shanghai, China, [4] Data Mining Lab, Beijing University of Technology, Beijing, China, [5] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China

OPEN ACCESS

Edited by:
Hamid R. Rabiee,
Sharif University of Technology, Iran

Reviewed by:
Tito Cali',
University of Padova, Italy
Marwan Abdellah,
École Polytechnique Fédérale de Lausanne, Switzerland
Marco Agus,
King Abdullah University of Science and Technology, Saudi Arabia

***Correspondence:**
Qiwei Xie
qiwei.xie@ia.ac.cn
Hua Han
hua.han@ia.ac.cn

Specialty section:
This article was submitted to Neural Technology, a section of the journal Frontiers in Neuroscience

Received: 07 August 2019
Accepted: 15 May 2020
Published: 21 July 2020

Citation:
Liu J, Li L, Yang Y, Hong B, Chen X, Xie Q and Han H (2020) Automatic Reconstruction of Mitochondria and Endoplasmic Reticulum in Electron Microscopy Volumes by Deep Learning. Front. Neurosci. 14:599. doi: 10.3389/fnins.2020.00599

Together, mitochondria and the endoplasmic reticulum (ER) occupy more than 20% of a cell's volume, and morphological abnormality may lead to cellular function disorders. With the rapid development of large-scale electron microscopy (EM), manual contouring and three-dimensional (3D) reconstruction of these organelles has previously been accomplished in biological studies. However, manual segmentation of mitochondria and ER from EM images is time consuming and thus unable to meet the demands of large data analysis. Here, we propose an automated pipeline for mitochondrial and ER reconstruction, including the mitochondrial and ER contact sites (MAMs). We propose a novel recurrent neural network to detect and segment mitochondria and a fully residual convolutional network to reconstruct the ER. Based on the sparse distribution of synapses, we use mitochondrial context information to rectify the local misleading results and obtain 3D mitochondrial reconstructions. The experimental results demonstrate that the proposed method achieves state-of-the-art performance.

Keywords: mitochondria, endoplasmic reticulum, electron microscopes, segmentation, 3D reconstruction

1. INTRODUCTION

In eukaryotic cells, mitochondria and the endoplasmic reticulum (ER) together occupy more than 20% of the cell volume. Evidence suggests that mitochondrial and ER morphology changes can have severe consequences for cell physiological or pathological functions, such as apoptosis, Ca^{2+} homeostasis, and metabolite processing (Karbowski and Youle, 2003; Twig et al., 2008; Bhatti et al., 2017). Mitochondria–ER plasma membrane (MAM) contacts are particularly abundant in cell bodies and are also important for multiple physiological functions (Marchi et al., 2014). Recently, emerging evidence has documented that mitochondria, ER, and MAM abnormalities or dysfunctions can result in various neurodegenerative disorders, such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis (Manfredi and Kawamata, 2016; Liu and Zhu, 2017; Rodriguez-Arribas et al., 2017).

The three-dimensional (3D) ultrastructure of mitochondria and ER at high resolutions requires large 3D reconstructions to be generated via electron microscopy (EM) (Vincent et al., 2016; Hirabayashi et al., 2017; Krols et al., 2018; Delgado et al., 2019). With the rapid development of

large-scale electron microscopy (EM) technology, the brain imaging methods have undergone tremendous changes. Manual contouring and reconstruction of the 3D ultrastructures of these organelles from high-resolution EM images has previously been accomplished in biological studies, such as glycogen granules (Agus et al., 2019) and mitochondria (Calì et al., 2019). However, manual annotation alone is insufficient to match the speed of data acquisition and meet the data analysis demands. Hence, developing an automatic algorithm to detect mitochondria and ER from EM images is both necessary and urgent.

Unlike ER segmentation, which has been less studied thus far, the detection and segmentation of mitochondria has been a hot topic in the neuroscience and computer vision fields. Nevertheless, the complex textures as well as similar ultrastructures in EM images have made the segmentation of mitochondria a challenging problem. Recently, variety of methods have been developed to automatically detect and segment mitochondria (Liu et al., 2018; Xiao et al., 2018; Xie et al., 2018b). GentleBoost classifier was trained for detecting mitochondria based on textural features (Vitaladevuni et al., 2008). Narasimha et al. (2009) utilized multiple classifiers to localize and segment mitochondria jointly in 3D images. Lucchi et al. (2012b) presented an automated approach with consideration of 3D shape cues for segmentation of mitochondria, and this method greatly reduced computational complexity by operating on super voxels instead of voxels. And then they improved classification accuracy by using context-based features and modeling the double membrane that encloses mitochondria (Lucchi et al., 2014). Jorstad and Fua (2015) iteratively refined the boundaries of mitochondria surfaces, starting from rough prediction provided by a machine learning-based method. In addition, there are some methods introducing graphical models into segmentation of mitochondria and achieving promising results, such as Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) (Lucchi et al., 2012a, 2013; Márquez-Neila et al., 2014).

However, all the aforementioned works require handcrafted mitochondrial features as the kernel of the algorithm. The wide success of the convolutional neural network (CNN) when applied to image processing tasks has proven that these models can learn powerful feature representations (Huang et al., 2018; Xie et al., 2018a). In the field of computer vision, instance segmentation, which involves the automatic detection of all objects appearing in an image and precise delineation of each instance, has undergone significant improvements based on deep learning models in recent years. We briefly review some of the most significant works below. Mask R-CNN (He et al., 2017) segments objects based on bounding boxes produced by CNNs. Concretely, Mask R-CNN extends Faster R-CNN by adding a parallel branch to predict a segmentation mask for each region of interest (RoI). In addition to the proposal-based methods, recurrent networks have gradually been introduced to solve the instance segmentation problem. Reversible Recursive Instance-level Object Segmentation (R2-IOS) (Liang et al., 2016) is composed primarily of two subnetworks. A reversible proposal refinement subnetwork is used to refine the object bounding boxes, while an instance-level segmentation subnetwork predicts the object mask for each proposal. The flooding filling network (FFN, Januszewski et al., 2018) utilizes a recurrent convolutional network to delineate a single object; it takes an object image channel and an object mask channel as inputs. The output mask channel of the FFN serves as the input mask channel for the next iteration, providing an explicit snapshot of the current segmentation state. One major challenge in FFNs is determining the initial state of the object mask, which is called a seed.

In this paper, we aim to develop an automated pipeline for detecting mitochondria and ER to seek morphological characteristics in different domains. The workflow of the entire pipeline is illustrated in **Figure 1**. The main contributions of this work are as follows:

1. We present an efficient, fully CNN to reconstruct the ER from EM images. To our knowledge, this is the first work that applies a deep network to ER segmentation.
2. We propose a novel recurrent network that iteratively refines mitochondrial segmentation. The detection subnetwork generates bounding boxes that are passed to the segmentation subnetwork as the initial seeds. Then, the segmentation subnetwork recursively identifies the mitochondrial boundaries.
3. For the highly anisotropic EM images, we introduce an efficient 2D-to-3D approach to reconstruct mitochondria that first obtains 2D segmentation results utilizing the recurrent network; then, morphological processing and mitochondrial context information are used to rectify local misleading results. Finally, a 3D connection algorithm is applied to obtain the 3D mitochondrial structures.
4. We reconstructed all the mitochondria and the ER on two EM image stacks with volumes up to $\sim$ 28,628 μm^3 and found 4 different structural features of mitochondria-ER contact sites.

2. MATERIALS AND METHODS

2.1. Materials

The mouse cortex sample used in this study was acquired by the Institute of Neuroscience, Chinese Academy of Sciences. The tissue block was automatically cut into serial sections with thicknesses of approximately 50 nm. Then, the sections were imaged by the Institute of Automation at the Chinese Academy of Sciences using a scanning electron microscope (SEM) (Zeiss Supra55) with a resolution of 3 nm $\times$ 3 nm and a dwell time of 1.5 μs. The SEM images were acquired using secondary electron detection (9 kV accelerating potential, and at a working distance of $\sim$ 6.0 mm). A total of 511 image planes with a size of 15,000 $\times$ 8,300 were acquired, yielding 59.25 GB of data.

2.2. Image Preprocessing

As mentioned above, the imaging data were obtained by SEM. The SEM technique is non-destructive, which means that the specimens can be imaged more than once. However, because of the section collection method, arbitrary angle rotations and some distortions are inevitable. Hence, image alignment is important to correct these errors and obtain a sequential image stack. We adopted the method in (Chen et al., 2018)

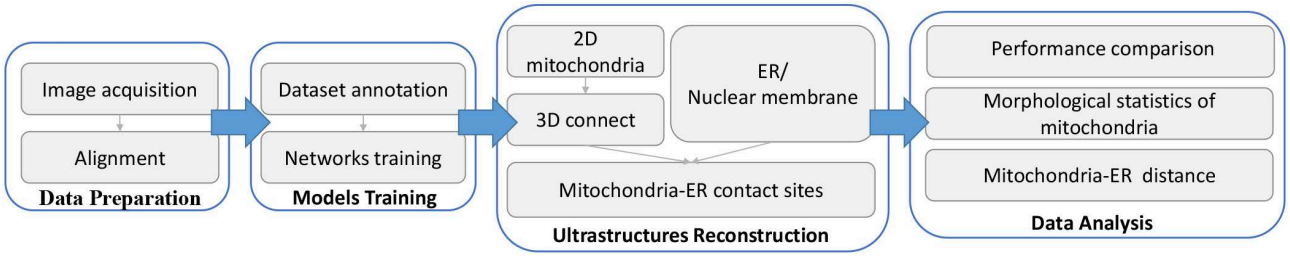


FIGURE 1 | Workflow of our proposed method. Briefly, there are four main steps in the complete method. (1) Data preparation: First, serial images are obtained by SEM and then aligned to create 3D image stacks using a non-linear registration method. (2) Model training: From the aligned data, we annotate some images to form a training dataset, and then separately train two networks to process mitochondria and ER. (3) Ultrastructure reconstruction: 2D mitochondria were predicted using the trained network and connected to form 3D mitochondria. The ER is segmented by the trained network. Then, we detect mitochondria-ER contact sites from the above results. (4) Data analysis: We measure the performance of our method and calculate biological measurements to conduct various analyses.

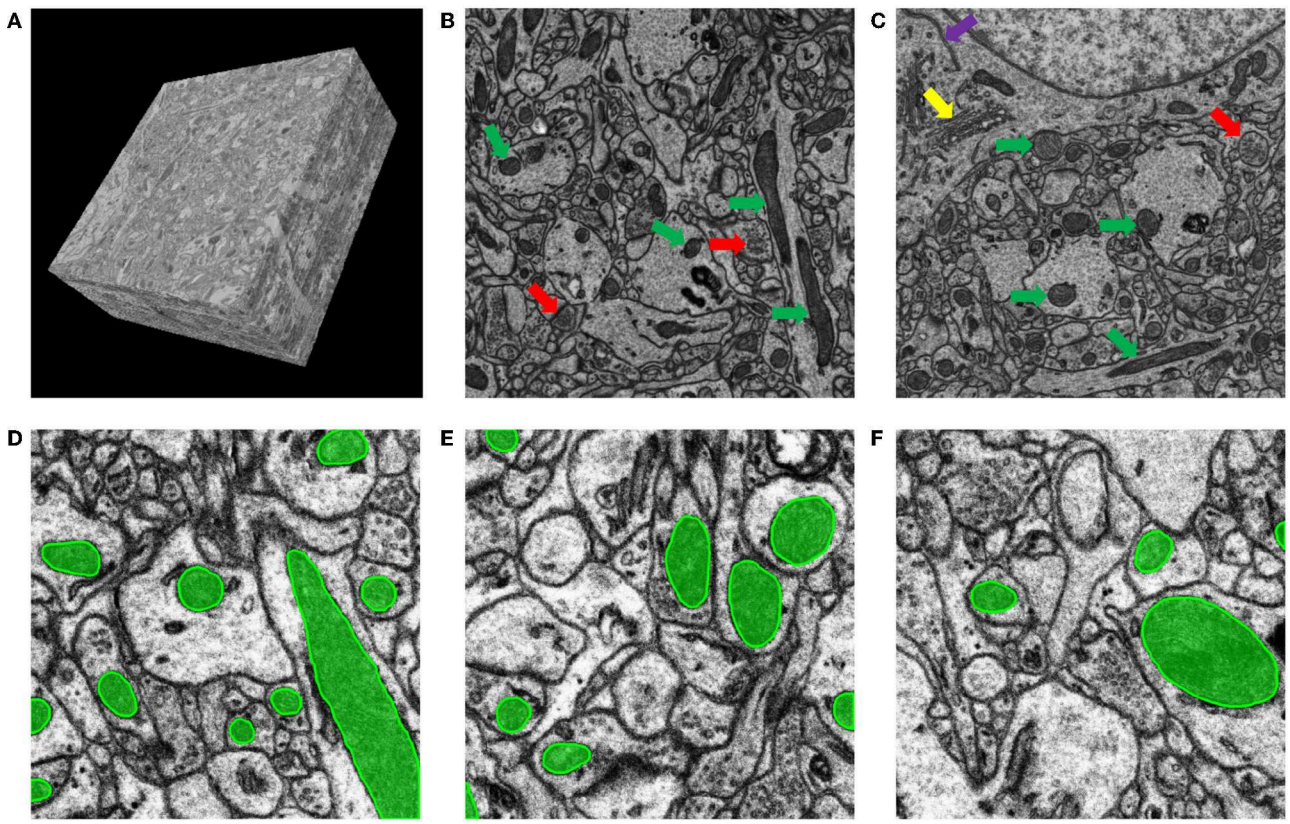


FIGURE 2 | Mouse cortex neural tissue acquired by ATUM-SEM. **(A)** An example of an aligned image stack that covers approximately 20 × 20 × 10 μm through the ATUM-SEM method. **(B,C)** Examples of mitochondria and other ultrastructures. The green arrows indicate mitochondria; the red arrows indicate vesicles; the yellow arrow indicates the Golgi body, and the purple arrow indicates the endoplasmic reticulum. **(D–F)** Examples of mitochondria segmentation in our training dataset.

for image registration. First, we used the scale invariant feature transform (SIFT, Liu et al., 2008) to detect corresponding landmarks across adjacent sections. Then, the wrinkled areas were annotated manually. Finally, we used a modified moving-least-squares (MLS, Schaefer et al., 2006) deformation algorithm to register adjacent sections with wrinkles. This algorithm reflects the discontinuity around wrinkle areas while maintaining smoothness in other regions. A registered image stack is shown in **Figure 2**.

In addition, some differences exist in the distribution of grayscale values in the EM images. To reduce the network training difficulty, we preprocessed the raw images using a histogram matching method to maintain grayscale consistency.

2.3. Proposed Network for Mitochondria

2.3.1. Network Architecture

Mask R-CNN (He et al., 2017) was proposed to solve the problem of instance segmentation based on segment proposals.

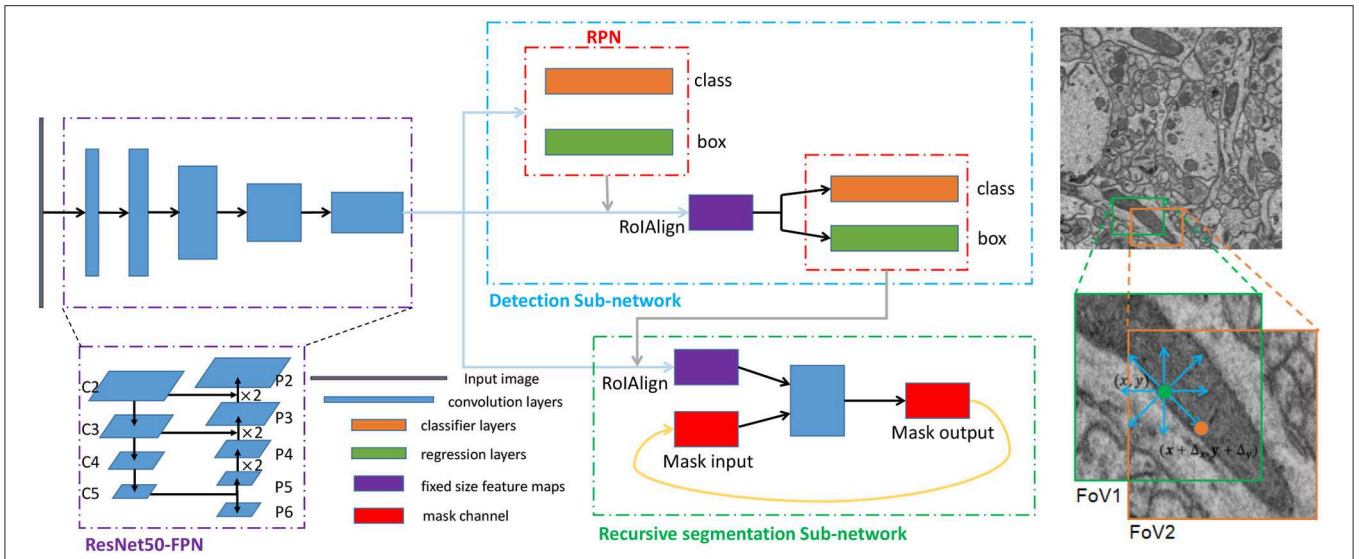


FIGURE 3 | The architecture and training scheme of the proposed network for segmenting mitochondria. **Left**: The three dotted boxes represent the backbone network (purple), the detection subnetwork (blue) and the recursive segmentation subnetwork (green). The black block, blue blocks, orange blocks, green blocks, purple blocks, and red blocks indicate the input image, convolution layers, classification layers, regression layers, and the fixed-size feature maps obtained from the RoIAlign and mask channels, respectively. **Right**: A simplified execution scheme for segmenting mitochondria. The green box indicates the outputs from the detection subnetwork, which may be smaller than the ground truth. After the first iteration of the segmentation subnetwork, the eight prediction directions are checked, and the direction is the input for the next iteration. In this example, the position in the orange circle is selected, and then the field of view (FoV) moves to the orange box.

It produces candidate object regions through a region proposal network (RPN). Based on the proposals from the RPN, R-CNN conducts further classification and regression. In contrast to Faster R-CNN (Ren et al., 2015), Mask R-CNN predicts object masks in end-to-end fashion by adding a mask branch. In combination with the feature pyramid network (FPN, Lin et al., 2017) in the backbone network, Mask R-CNN can extract features from different levels of the feature pyramid based on the scale of each RoI. This mechanism is able to exploit more detailed features to obtain finer segmentation results for small objects.

However, mitochondria vary widely in size and shape. In addition to the common elliptical mitochondria, some are long and narrow, and some are curly. A Mask R-CNN model with an FPN backbone is unsuitable for this situation. Specifically, the bounding boxes that R-CNN predicts are always smaller than the true bounding boxes for larger mitochondria, which leads to incomplete segmentation due to the series structure of the mask branch and R-CNN. Inspired by the scheme proposed in FFN (Januszewski et al., 2018), we propose to refine the segmentation results by moving the field of view (FoV) of the mask branch to extend the detection boxes, which is implemented by introducing an input mask that preserves the previous segmentation state. In the following subsections, we describe the key components of the proposed network (see **Figure 3**) in detail, including the detection subnetwork and the recursive segmentation subnetwork.

2.3.1.1. Detection subnetwork

Due to the relatively simple characteristics of mitochondria, a very deep network may cause overfitting; moreover, a simpler network will have considerably lower computational costs. Therefore, we adopt ResNet50 (He et al., 2016) as the backbone network. Considering the substantial variance in mitochondrial size, we utilize a feature pyramid network (Lin et al., 2017) to explore features at different scales. Concretely, an FPN uses a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input. The main task of the RPN is to produce the candidate object regions. At every feature level of FPN, we use 3 ratios, yielding 3 anchors at each location of the convolutional layer. The RPN predicts a score and a bounding-box regression for each anchor. Then, RoI features are extracted from different levels of the feature pyramid based on their scales. To avoid misalignments between the RoI and the extracted features, we use RoIAlign to avoid any quantization operations. Finally, R-CNN performs further classification and regression based on the RoI features from RPN. The outputs of the R-CNN are then used as the initial inputs of the recursive segmentation subnetwork.

2.3.1.2. Recursive segmentation subnetwork

The main role of the recursive segmentation subnetwork is to achieve a precise depiction of the mitochondria detected by the detection subnetwork. The recursive segmentation subnetwork takes an RoI feature and an object mask as inputs and predicts the probability map of the object with the focus. The input object mask is initialized by assigning the center position of the detection box an active value and assigning the other positions zero. This serves to show that the center pixel may be part of the target. Then, the subnetwork runs a forward pass to obtain a segmentation mask for the current FoV. Due to the existence

of incomplete detection boxes, we introduce a mechanism for moving the FoV to extend the object via the subnetwork iterations. Note that the input object mask in the next iteration uses the updated probability map based on the output of the preceding iteration. This approach provides information about the current focus target.

The specific architecture of the segmentation refinement subnetwork is shown in **Figure 3**. Because the detection boxes have different sizes, we still use RoIAlign to obtain fixed-size feature maps from the corresponding level of the feature pyramid; these function as the feature channel of the inputs. In this study, we set the fixed size to 33×33. To combine the two inputs, we simply concatenate them. The subsequent operation includes four convolution layers with a ReLU: each convolution has a kernel size of 3×3 and uses the same mode to guarantee an identical output size. Before each ReLU, we utilize batch normalization (BN) to accelerate convergence during the training process. In the final layer, we use a 1×1 convolution to reduce the number of channels to the desired number of classes, which is one in our case. We do not have many convolution operations in this subnetwork, and one of the reasons is that the input features are extracted from the shared feature pyramid, which already possesses high-level abstract information. Reusing these shared features greatly simplifies the training process.

2.3.2. Training

The two subnetworks are trained separately to learn mitochondrial feature representations to obtain the detection and segmentation results.

For the detection subnetwork, the implementation details are similar to those of Lin et al. (2017). We obtain the last residual block of each stage from ResNet50 as one pyramid level of the bottom-up pathway in FPN, denoted as C_2, C_3, C_4, and C_5. We upsample the feature maps and leverage lateral connections to form the top-down pathways P_2, P_3, P_4, and P_5, as shown in **Figure 3**. Notably, we also introduce pathway P_6 to enlarge the range of the scale of anchors by simply subsampling P_5. P_6 is used only in the RPN to predict candidate regions. From the high-level to the low-level feature maps, the corresponding anchor scales are $\{32^2, 64^2, 128^2, 256^2, 512^2\}$, and each scale has three ratios, $\{1:1, 1:2, 2:1\}$. The RPN generates RoIs of various sizes; we loop through all the candidate RoIs and extract features from the pyramid levels according to their scales. This strategy helps ensure that larger objects obtain stronger semantic features and that smaller objects obtain more detailed features.

This paper defines a multitask loss as

$$L = L_{cls}^{rpn} + L_{box}^{rpn} + L_{cls}^{r-cnn} + L_{box}^{r-cnn}, \quad (1)$$

Based on the aforementioned settings, we trained the detection subnetwork with a backpropagation algorithm and saved the weight parameters of the network. Then, we initialized the recursive segmentation subnetwork with the trained model to share the FPN backbone features.

To optimize the segmentation subnetwork, we use cross-entropy loss, which can be represented as follows:

$$L_{mask} = \sum_i -[y_i * log(p_i) + (1 - y_i) * log(1 - p_i)], \quad (2)$$

where y denotes the ground-truth label, and p denotes the prediction for each pixel, which is the activation value of the sigmoid. i is the pixel index.

During the second training stage, the training samples are generated randomly. Specifically, we first generate RoI proposals to simulate the detection results of the detection subnetwork. Then, the RoI features are extracted from the corresponding level of pyramid features according to the scale of the bounding box, and the RoI mask is extracted from a mask we maintain during training to update the probability map of the iterative procedure. To fit the size of the input channel, we resize the RoI mask using the nearest bilinear interpolation algorithm. Similarly, the ground-truth mask should be resized to the same size. Here, we use nearest neighbor interpolation to guarantee a binary value in the ground truth.

The movement of the FoV is restricted to one step away from the current position to simplify the training process. After one forward pass of a training sample, the eight directions relative to the current position are checked to seek the new positions that the network will move to next (see **Figure 3**). We set a movement step $(\Delta_x, \Delta_y) = (8, 8)$ in the fixed-size $(33, 33)$ object mask to inform the movement. That is, first, the center position of the current FoV is denoted as (x, y); then, we sequentially check the eight new positions $(x + \Delta_x, y), (x + \Delta_x, y + \Delta_y), (x, y + \Delta_y), (x - \Delta_x, y + \Delta_y), (x - \Delta_x, y), (x - \Delta_x, y - \Delta_y), (x, y - \Delta_y), (x + \Delta_x, y - \Delta_y)$. During the inspection, if the predicted probability is greater than or equal to the moving threshold ($T_{move} = 0.9$), that point becomes the new center of the FoV in the next iteration. This process is repeated until all possible positions have been traversed.

2.3.3. Inference

The inference process is also divided into two steps. First, the proposals (x_1, y_1, x_2, y_2) are obtained by running the detection subnetwork. Then, the recursive segmentation subnetwork takes these boxes as inputs and refines the segmentation results recursively.

Compared with training, the inference scheme of the segmentation subnetwork has the following characteristics. To extend to the complete mitochondria, the inference is not limited to only one step in the eight directions; instead, a queue is maintained to preserve the new positions. Before a new iteration, a position is popped; then, new positions are computed after a forward pass and pushed into the queue in descending order.

If we denote the width and height of the current box as (w, h) and the size of the input object mask as (s, s), then the movement vector is computed as follows:

$$(\Delta_x', \Delta_y') = (\frac{\Delta_x \cdot w}{s}, \frac{\Delta_y \cdot h}{s}). \quad (3)$$

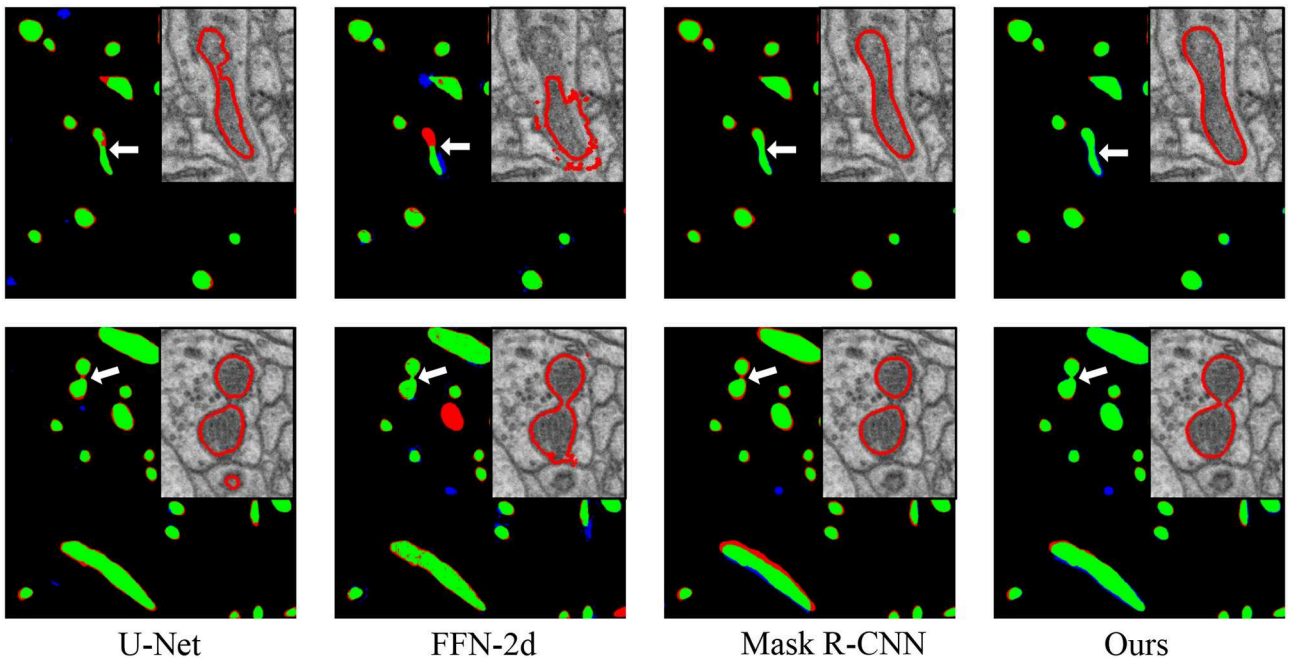


FIGURE 4 | Performance comparisons with the baseline approaches. From left to right: comparisons to the manual ground truth from U-Net, FFN-2d, Mask R-CNN and our proposed network. In each comparison image, green pixels represent true positives (TP), blue pixels represent false positives (FP), red pixels represent false negatives (FN), and black pixels represent true negatives (TN). The insert at the top right shows enlarged details pointed to by the white arrow. Qualitatively, the segmentation by the proposed network contains fewer false positives than that of FFN-2d and more true positives than those of U-Net and Mask R-CNN.

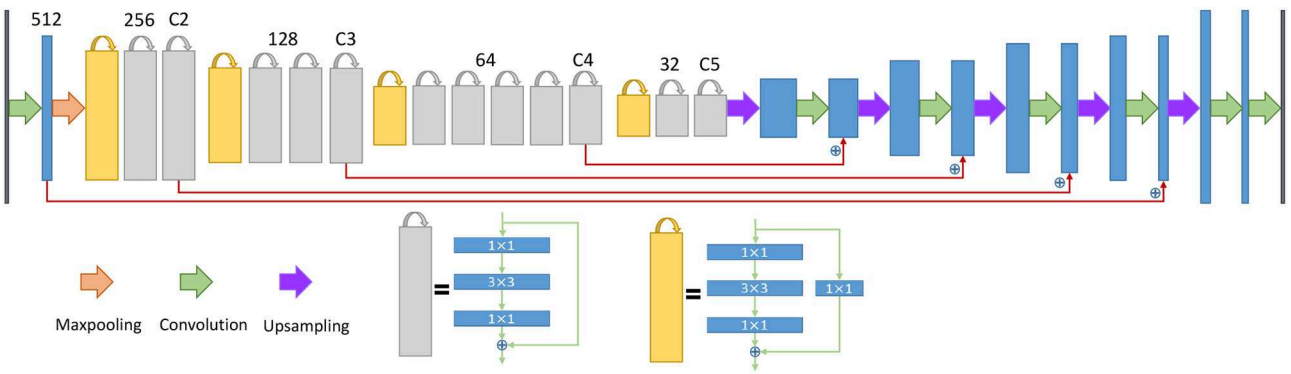


FIGURE 5 | The architecture of the neural network for the endoplasmic reticulum. The yellow blocks and gray blocks denote two different residual blocks, and the numbers above the blocks indicate the number of feature map channels. The blue plus signs indicate a sum operation rather than concatenation.

This simple strategy results in the process taking larger steps for larger objects and smaller steps for smaller objects, which helps to solve the inefficiency problem to a certain extent.

2.4. Proposed Network for ER

To explore the contact sites between the ER and mitochondria, we need to obtain the morphologies of the ER. We also observed that the nuclear membrane was incorrectly detected if we attempted to segment the ER only. Therefore, we decided to segment the ER and the nuclear membrane as an auxiliary task using one deep neural network, which promotes the ER segmentation performance. Additionally, the inner nuclear membrane is labeled in the training samples as a third category used to distinguish the ER and nuclear membrane in postprocessing. Then, we build a fully convolutional network based on ResNet50 (He et al., 2016) to classify all the pixels in one image. The network contains an analysis and a synthesis path similar to the architecture of U-Net (see **Figure 5**). The first six resolution steps in the analysis path are taken from the first six stages of ResNet50. In the synthesis path, each resolution step is upsampled first, followed by a 3 × 3 convolution layer and batch normalization (BN). Then, the layer is merged with the

equal resolution from the analysis path by a pixelwise addition operation, followed by a rectified linear unit (ReLU). Using this strategy, the localization and contextual information can be finely balanced. In the final layer, a 1×1 convolution layer is used to map the feature channel to the desired number of classes, which is 3 in our case.

During inference, each pixel is classified with the class that has the maximum value in the three-dimensional classification vector. To differentiate ER from the nuclear membrane, we mark the connected components that are adjacent to the interior of the nuclear membrane as the nuclear membrane; then, the remaining components are labeled as the ER. To illustrate this process of separating ER from the nuclear membrane, **Figure 6** shows an example that includes a prediction from the network and the corresponding relabeled result in.

2.5. Training and Testing Datasets

As noted previously, we acquired an SEM dataset with a voxel resolution size of 3 nm $\times$ 3 nm $\times$ 50 nm.

To create a dataset for ER segmentation, we manually labeled 60 serial sections with sizes of 4,096 $\times$ 6,344, used 49 slices as the training data, and used the remainder as testing data. The pixels were labeled with three categories: ER, nuclear membrane, and interior of nuclear membrane or background. **Figure 6** shows a manually annotated sample.

To create a dataset for segmentation of mitochondria, we extracted 15 slices as a training dataset, each of which has a size of 7,168 $\times$ 7,168; then, 5 slices of the same size were used as a test dataset to evaluate the algorithm performance. The ground truth of the mitochondria dataset was annotated by experienced students using Trakem2 software (Cardona et al., 2012). **Figure 2** shows some specific examples from our annotated ATUM-SEM dataset. In addition, to ensure a fair comparison with existing methods, we conducted experiments on the FIB-SEM dataset, which has been widely used for mitochondrial segmentation. The FIB-SEM dataset contains a training volume and a testing volume taken from the CA1 hippocampus region of the brain. Each volume consists of 165 slices with a resolution of 5 $\times$ 5 $\times$ 5 nm.

Deep learning requires a considerable amount of training data to avoid overfitting problems. Therefore, we enlarged the training dataset using data augmentation techniques, including random rotation, random flipping and adding random noise. The data augmentation process was conducted online.

2.6. Experimental Setup

All the proposed networks were implemented using the Keras open-source deep learning library (Chollet et al., 2015), with the TensorFlow library as the backend. Due to GPU memory constraints, the original images were cut into smaller images as inputs. We trained all the networks using stochastic gradient descent. However, the related training parameters were slightly different between the networks for mitochondria and ER. For the mitochondrial network, the momentum was set to 0.9, the weight decay was set to 0.0001, and the learning rate was initially set to 0.001 and then decreased by a factor of 10 whenever the learning process stagnated. The ER network was trained using a learning rate of 0.01 and a momentum of 0.9. All the training and testing tasks were conducted on a server equipped with an Intel i7 CPU, 512 GB of main memory, and a Tesla K40 GPU.

2.7. Method for Reconstructing Mitochondria

Because we employed a refinement subnetwork to obtain more precise segmentation results, many trivial false segmentations will appear in the 2D slices. Therefore, a suitable postprocessing method is necessary to improve accuracy. This section focuses on how the segmentation results are optimized to further improve the performance while obtain 3D mitochondria. Note that the mitochondrial sizes are far larger than the resolution in the z-direction. We utilize the multilayer information fusion algorithm proposed by Li et al. (2018) to reconstruct the mitochondria in 3D and discard mitochondria whose "length" (the number of occurrences in the z-direction) is less than L (e.g., 15).

For completeness, we note that there is no need to conduct 3D connections for the ER and nuclear membrane. We labeled three cell bodies manually and this analysis is restricted to the cell body only.

2.8. Measurement of Biological Statistics

To obtain the cross sections, the first task is to obtain the skeleton of the mitochondrion. We first obtain the main skeleton, excluding the tiny branches, using ImageJ software. However, the anisotropy of the data causes the main skeleton to be a rough line. We smooth this skeleton by applying polynomial fitting. Finally, we take some points on this skeleton line and determine the perpendicular planes. Thus, the cross sections of one mitochondrion are the intersections of the perpendicular planes and the 3D mitochondrion. Using this approach, we measured the area and perimeter of the cross section.

To obtain the minimum distances between mitochondria and the ER, we simply compute all the point pairs and find the smallest values. To reduce the computational complexity, we downsampled the original images by a factor of 4 in the X and Y directions and a factor of 2 in the Z direction. Additionally, we reduced the number of points by edge extraction.

3. RESULTS

3.1. Performance Comparisons

In this subsection, we focus primarily on the segmentation performance and the computational costs of the proposed method. For completeness, we do not measure the segmentation performance of ER because no method for doing exists at present.

3.1.1. Evaluation Metrics

To quantify the pixelwise segmentation quality, we select the following metrics: *Jaccard index*, *Accuracy*, *Precision*, *Recall*, *F-score*, and *Dice coefficient*; these are the common criteria used in image segmentation tasks. All these measures are calculated based on combinations of the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) of the predicted pixels. The values of all the criteria range

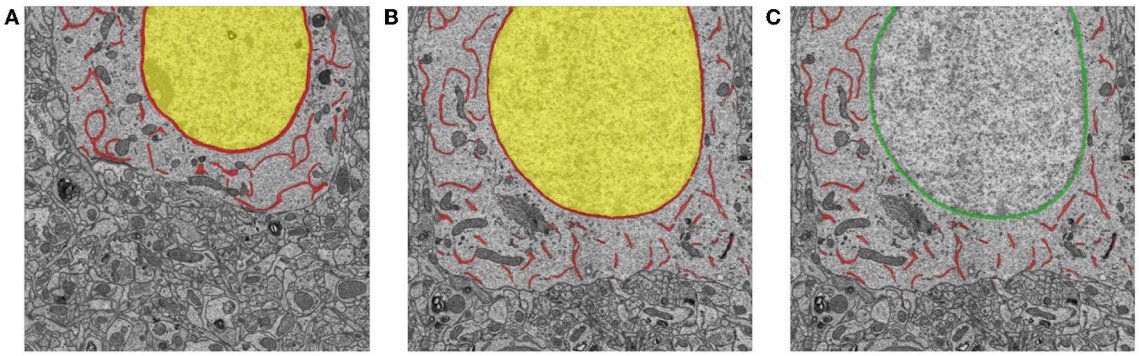


FIGURE 6 | Examples of training samples and predictions for ER segmentation. **(A)** An example of manual annotations of the nucleus (yellow), ER, and the nuclear membrane (red). **(B,C)** An inference from the trained neural network and the corresponding relabeled result. The red and green pixels indicate the ER and the nuclear membrane, respectively.

TABLE 1 | Different evaluation indicators.

Performance measure	Mathematical expression
Jaccard index	$\frac{TP}{TP+FP+FN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
Dice coefficient	$\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$

TABLE 2 | Segmentation performance comparison on the ATUM-SEM dataset.

Indicator \ Method	U-Net	FFN-2d	Mask R-CNN	Proposed method
Jaccard index	0.7972	0.6193	0.7799	**0.8021**
Accuracy	**0.9877**	0.9699	0.9863	0.9876
Precision	0.9403	0.7712	**0.9621**	0.9425
Recall	0.8397	0.7586	0.8051	**0.8442**
Dice coefficient	0.8845	0.7590	0.8750	**0.8891**

TABLE 3 | Segmentation performance comparison on the FIB-SEM dataset.

Method	Jaccard index
Kernelized SSVM/CRF	0.840
Lucchi2013	0.867
Non-parametric higher-order random fields	0.762
Improved KernelBoost	0.776
Ours proposed method	0.864

from 0 to 1, where 0 indicates no coincidence and 1 denotes total coincidence.

Jaccard index is defined as the ratio of the area of the intersection and the area of the union between the ground truth and the segmentation result. *Dice coefficient* computes the spatial overlap between the ground truth and the segmentation result, describing a measure of similarity. The definitions of *Accuracy*, *Precision*, *Recall*, and *F-score* are the same as those commonly used in machine learning. The explicit mathematical expressions of these metrics are shown in **Table 1**.

3.1.2. Segmentation Performance

Experiments on both datasets are conducted to compare with recent approaches.

On the ATUM-SEM dataset, we compared our method with U-Net (Ronneberger et al., 2015), Mask R-CNN and FFN. The same backbone network and feature pyramid network (FPN) are employed in Mask R-CNN. We note that we adopted a 2D version of FFN for this comparison for two main reasons: first, the ground truth is labeled on the 2D images in our dataset, which means the mitochondria cannot be discriminated from each other in 3D. However, the original FFN used a 3D reconstruction method that requires 3D segmentation for training. Second, our proposed network learns features only at the 2D level, but the original FFN trains a recurrent 3D convolutional network to extend objects; thus, it integrates 3D features. For a more consistent comparison, we trained a 2D CNN called FFN-2d, which has 9 convolution layers. The network architecture is generally the same as (Januszewski et al., 2018) except for the dimensions of the convolution operations. We set both the input and output sizes of FFN-2d to 33 × 33.

We present a quantitative comparison with state-of-the-art methods in **Table 1** in terms of segmentation performance indicators, where the highest values are marked in bold for distinction. For a fair performance comparison, all the evaluation values were obtained directly from the networks, without any postprocessing. All the values for the thresholding operation were set to 127. As shown in **Table 2**, our proposed approach achieves a Jaccard index of 0.8021, a recall of 0.8442, and a dice coefficient of 0.8891. Its accuracy value (0.9876) is nearly equal to the highest value. Clearly, the proposed method achieves a better segmentation performance than do the previous methods. For qualitative comparison, two specific examples are illustrated in **Figure 4**. As indicated by the purple arrows, our proposed method produces more complete and accurate

TABLE 4 | Computational costs comparisons between manual annotation and our proposed method.

	Segmentation of ER	Segmentation of Mitochondria
Manual annotation	4,044 h	5,055 h
Our proposed method	12.83 h (training) + 10.63 h (inference)	43.8 h (training) + 123 h (inference)

segmentation results than do the baseline approaches. From the detection viewpoint, our method obtains more accurate numbers of mitochondria.

On the FIB-SEM dataset, we compared our method with Non-parametric Higher-order Random Fields (Márquez-Neila et al., 2014), Improved KernelBoost (Rigamonti et al., 2014), Kernelized SSVM/CRF (Lucchi et al., 2012a), and Lucchi2013 (Lucchi et al., 2013). In our method, the trained segmentation subnetwork outputs mitochondria probability maps for each slice. Then a simple thresholding operation is performed to obtain binary results for computing the jaccard index. Additionally, the network architecture and parameters are the same as those of ATUM-SEM dataset. For other methods, we obtain the reported values on the FIB-SEM dataset from corresponding literatures. The performance comparisons are shown in **Table 3** in terms of Jaccard index. Our results yielded 0.864 Jaccard index, showing comparable performance with the other state-of-the-art methods.

3.1.3. Computational Costs

In this subsection, we present the time consumed by complete manual annotation and our proposed method to show the importance of having an automatic algorithm. **Table 4** illustrates the time consumed for each step: the steps mainly consist of the ER segmentation and the segmentation of mitochondria.

Due to the limited GPU memory of computers, we cut the original EM images into small patches used as network inputs. Therefore, all the computations are estimated based on the times required for these small patches. On average, it takes approximately 4 and 5 min to annotate all the ER and mitochondria in a 1,024 × 1,024 image, respectively, using the Trakem software. There are approximately 60,670 patches in the 28,628 μm^3 EM stack. In our pipeline, the training time required 12.83 h for ER and 43.8 h for mitochondria. Subsequently, the inference step takes 0.6308 s for ER and 8.2081 s for the mitochondria in a 1,024 × 1,024 image. All the costs were measured on a Tesla K40 GPU. These times would be further accelerated with the application of more powerful computing resources. From the comparison, we can see that applying our method improves the segmentation costs by orders of magnitude.

3.2. 3D Reconstruction of Mitochondria

After automatic segmentation and 3D reconstruction of the mitochondria of dendrites, axons and neurons in the mouse cortex, we found that mitochondrial morphology was varying in different cell positions (**Figure 7**). For example, in the axons and dendrites, the mitochondria were mainly long tubular structures (see **Figure 7A**). However, in the cell body, the mitochondria were typically adapted into reticular organizations (see **Figures 7B,C**). After quantification, the volume and surface area of the interconnected mitochondrial network in the cell body were approximately 0.217 and 2.093 μm^2, respectively, which was larger than the mitochondrial networks in the axons and dendrites (see **Figure 7D**).

Previous studies indicated that communication between mitochondria involves adjacency ("kissing") or a dynamic nanotubular tunnel between non-adjacent mitochondria ("nanotunneling") (Sun et al., 2012; Huang et al., 2013). After automatic reconstruction of mitochondria in the mouse cortex, we found that non-nanotunneling communication was distributed mainly in axons/dendrites (**Figures 7E,F**). Next, we applied the cross-sectional area to detect and quantify the non-nanotunneling ones. As shown in **Figure 7G**, the average cross-sectional area is 0.053 and 0.194 μm^2 in non-nanotunneling and nanotunneling mitochondria, respectively. We found that the cross-sectional area variance in nanotunneling mitochondria is 0.159 μm^2, and 0.0058 μm^2 in non-nanotunneling mitochondria. Our work revealed that automatic reconstruction of mitochondria provides a precise and unbiased method for detecting mitochondrial nanotunneling in neurons.

3.3. 3D Reconstruction of ER

Mitochondria–ER contact sites are highly conserved structures in eukaryotic cells (Marchi et al., 2014). Several approaches have been described in the literature for quantifying the mitochondria–ER contact site distance from EM images (Filadi et al., 2015; Naon et al., 2016; Faustini et al., 2019; Garrido-Maraver et al., 2020). This distance has also been analyzed using 2D imaging of fluorescently labeled mitochondria and ER or 3D reconstruction of fluorescently labeled mitochondria and ER (Friedman et al., 2010; Rowland and Voeltz, 2012). Due to low resolution and artifacts in 2D analysis, previous researchers applied EM to study the relationship between mitochondria and ER, but they mainly used manual annotation to reconstruct mitochondria and ER, which is time consuming (Hirabayashi et al., 2017; Krols et al., 2018). Therefore, individual algorithms are needed to exactly detect the distance between mitochondria and ER in 3D space. In our work, we present a novel method for quantifying the distance between mitochondria and ER of neurons after 3D reconstruction using the ATUM-SEM technique. After automatic segmentation and reconstruction of ER in neurons, we found that the mitochondria and the ER have different structural features. As shown in **Figure 8**, apart from non-contact between these two organelles, some contacts between mitochondria and ER are more extensive. For example, we found that the ER tubules circumscribed almost completely around the mitochondrial membrane, either with (**Figures 8A,B**) or without direct contact (**Figure 8C**). The average distance between these two organelles was 133.499 nm. The percentage of direct contact between mitochondria and the ER is 21.7%, while the percentage of distances less than 30 nm but without direct contact between these two organelles is 9.2% (**Figure 8F**). Our method provides a precise and unbiased

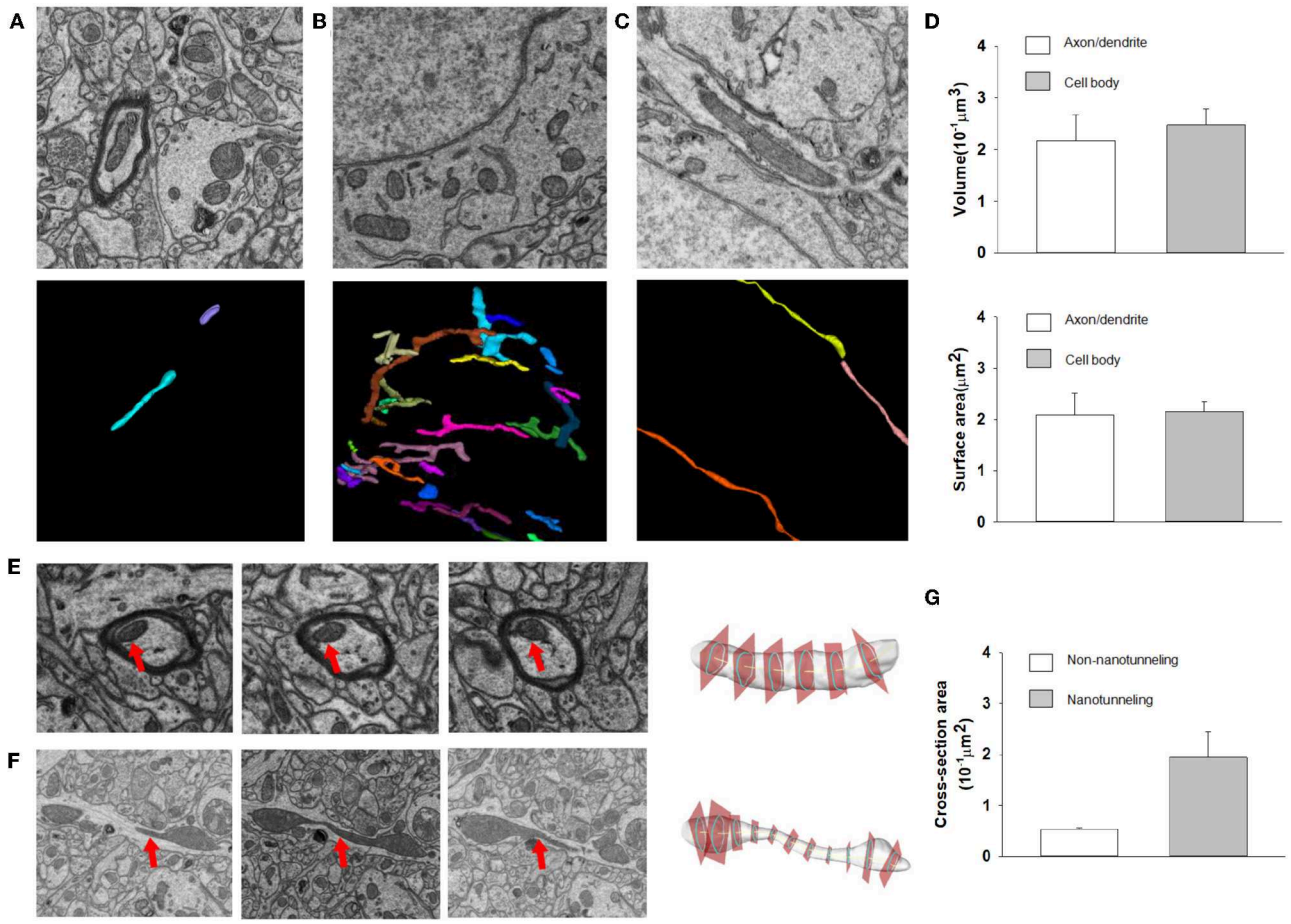


FIGURE 7 | The 3D morphology of mitochondria for different neuron locations. **(A–C)** Raw EM images and 3D reconstructed mitochondria in axons, neuron somas and dendrites. **(D)** Top: The volumes of mitochondria in the cell body ($n = 86$) are greater than those in axons/dendrites ($n = 81$) (one-sided Mann–Whitney test, $p < 0.0001$). Bottom: The surface area of mitochondria in the cell body ($n = 86$) is greater than that in axons/dendrites ($n = 81$) (one-sided Mann–Whitney test, $p = 0.0002$). **(E,F)** Example of cross section analysis in non-nanotunneling and nanotunneling mitochondria, respectively. The red arrows indicate mitochondria. **(G)** The cross-sectional areas of non-nanotunneling and nanotunneling mitochondria.

analysis of distances between mitochondria and ER with high resolution and no artifacts.

4. DISCUSSION

The main function of mitochondria is to serve as a local energy source that supplies ATP (van der Bliek et al., 2017). Evidence suggests that mitochondrial morphology changes can lead to defects in intermitochondrial signal propagation, Ca^{2+} uptake, and neuronal protection against apoptotic stimuli (Karbowski and Youle, 2003; Bhatti et al., 2017). In this study, we developed an unbiased and rapid method for mitochondrial morphology detection. In the axons and dendrites, the mitochondria were mainly organized into long tubular structures. However, in the cytoplasm, the mitochondria were typically adapted into reticular networks. The extremely varied morphological features of mitochondria for a structural basis for maintaining energy homeostasis in neurons. A previous study reported that mitochondria are dynamically transported between axons and dendrites, which is crucial for synaptic and neuronal function (Wang and Schwarz, 2009; Zinsmaier et al., 2009). The elongated tubules of mitochondria in axons and dendrites might contribute to highly efficient mitochondrial transport. The high cross-sectional area SD of nanotunneling mitochondria in the axons and dendrites indicates a high ratio of fission and fusion processes, which benefits mitochondrial delivery among neuron bodies, axons, dendrites and synapses. Therefore, regional differences in mitochondrial morphological features are likely accompanied by organelle transport differences.

The contacts between mitochondria and ER are important for many physiological functions, such as apoptosis, calcium hemostasis, bioenergetics and protein synthesis (Marchi et al., 2014; Phillips and Voeltz, 2016). Many researchers documented that the function and structure of MAMs defects can result in multiple neurodegenerative diseases, such as Alzheimer's disease, Amyotrophic lateral sclerosis motor neuron disease, and Parkinson's disease (Manfredi and Kawamata, 2016; Liu and Zhu, 2017; Rodriguez-Arribas et al., 2017). Previous studies

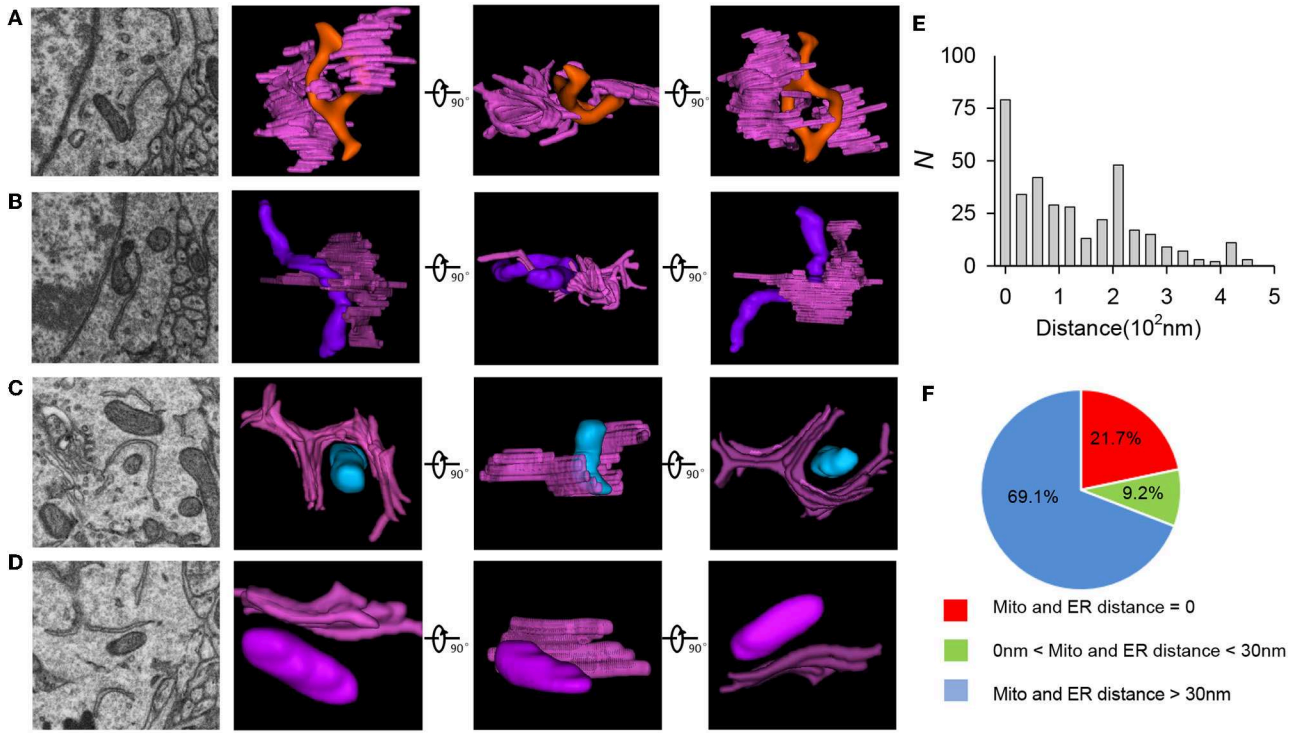


FIGURE 8 | The 3D morphology of ER in neurons. **(A–D)** Left: the original EM image; Right: ER 3D morphology (pink) under different directions. **(E)** Mitochondria and ER distance distribution in neurons. **(F)** Percentage of different mitochondria and ER distances in neurons.

have documented the ER's role in mitochondrial fission and fusion. ER is involved in mitochondrial fusion and fission process (Friedman et al., 2011; Lee et al., 2016; Abrisch et al., 2020). Our work provided a precise and unbiased method to detect the mitochondria and ER relationship in 3D space with high resolution. The 3D reconstruction of these two organelles can provide structural evidence. Due to low resolution of fluorescent microscopy and artifacts in 2D analysis, large-scale EM provides an excellent approach to study and quantify the contacts between these two organelles. Using high-resolution 3D reconstruction, ER tubules were shown to wrap around the mitochondria in neurons. Our data confirms that ER tubule may be involved in mitochondria division machinery recruitment in the early stage. Meanwhile, our work shows ER may play an unexpected role during the mitochondrial biogenesis process.

This paper proposes an automated pipeline that effectively obtains the 3D morphology of mitochondria, ER and mitochondria-ER contact sites in large-scale ATUM data. To enable the 3D segmentation of mitochondria, we first obtain the instance segmentations on each slice; then, we connect the 2D mitochondria to obtain the 3D reconstruction results based on contextual information. Instance segmentation is achieved by exploiting a recurrent network that outperforms most state-of-the-art methods. Compared with a semantic segmentation method, utilizing a region-based strategy can produce instance segmentation directly while preventing interruptions from noise and artifacts. We added a recursive segmentation subnetwork to refine the segmentation results, which reduces the effect of inaccurate detection results. In the case of the ER and nuclear membrane, 3D connections can be obtained simply in a single cell body. Therefore, we utilize a fully CNN to learn the ER and nuclear membrane as a single membrane system that are subsequently distinguished using a relabeling method.

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available because the datasets are being used for other studies. Requests to acess the datasets should be directed to yangyang2@shanghaitech.edu.cn.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Committee of the Institute of Neuroscience Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

JL and LL designed the experiments and wrote the paper. YY conducted the sample preparation. LL provided the scanning electron microscope images. BH conducted the annotation and visualization. XC registered the EM images. HH and QX

conceived the ideas for all the methods and provided suggestions based on the experimental results.

FUNDING

This research was supported by the National Natural Science Foundation of China (NOs. 61673381 and 31970960), the Special Program of Beijing Municipal Science & Technology Commission (NO. Z181100000118002), the Strategic Priority Research Program of Chinese Academy of Science (NO. XDB32030200), Bureau of International Cooperation, CAS (NO. 153D31KYSB20170059), and the Scientific Instrument Developing Project of Chinese Academy of Sciences (NO. YZ201671). We also acknowledge funding from the key program of the Ministry of Science and Technology of the People's Republic of China (NO. 2018YFC1005004).

ACKNOWLEDGMENTS

The authors would like to thank Mr. Lixin Wei and colleagues (Institute of Automation, CAS) for the use of the Zeiss Supra55 scanning electron microscope and for their technical support.

REFERENCES

Abrisch, R. G., Gumbin, S. C., Wisniewski, B. T., Lackner, L. L., and Voeltz, G. K. (2020). Fission and fusion machineries converge at ER contact sites to regulate mitochondrial morphology. *J. Cell Biol.* 219:e201911122. doi: 10.1083/jcb.201911122

Agus, M., Calì, C., Al-Awami, A., Gobbetti, E., Magistretti, P., and Hadwiger, M. (2019). Interactive volumetric visual analysis of glycogen-derived energy absorption in nanometric brain structures. *Comput. Graph. Forum* 38, 427–439. doi: 10.1111/cgf.13700

Bhatti, J. S., Bhatti, G. K., and Reddy, P. H. (2017). Mitochondrial dysfunction and oxidative stress in metabolic disorders–a step towards mitochondria based therapeutic strategies. *Biochim. Biophys. Acta Mol. Basis Dis.* 1863, 1066–1077. doi: 10.1016/j.bbadis.2016.11.010

Calì, C., Agus, M., Kare, K., Boges, D. J., Lehväslaiho, H., Hadwiger, M., et al. (2019). 3D cellular reconstruction of cortical glia and parenchymal morphometric analysis from serial block-face electron microscopy of juvenile rat. *Prog. Neurobiol.* 183:101696. doi: 10.1016/j.pneurobio.2019.101696

Cardona, A., Saalfeld, S., Schindelin, J., Arganda-Carreras, I., Preibisch, S., Longair, M., et al. (2012). TrakEM2 software for neural circuit reconstruction. *PLoS ONE* 7:e38011. doi: 10.1371/journal.pone.0038011

Chen, X., Xie, Q., Shen, L., and Han, H. (2018). "Morphology-retained non-linear image registration of serial electron microscopy sections," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens), 3833–3837. doi: 10.1109/ICIP.2018.8451280

Chollet, F. (2015). *Keras: Deep Learning Library for Theano and Tensorflow*. Available online at: https://keras.io/k

Delgado, T., Petralia, R. S., Freeman, D. W., Sedlacek, M., Wang, Y.-X., Brenowitz, S. D., et al. (2019). Comparing 3D ultrastructure of presynaptic and postsynaptic mitochondria. *Biol. Open* 8:bio044834. doi: 10.1242/bio.044834

Faustini, G., Marchesan, E., Zonta, L., Bono, F., Bottani, E., Longhena, F., et al. (2019). Alpha-synuclein preserves mitochondrial fusion and function in neuronal cells. *Oxidat. Med. Cell. Long.* 2019:4246350. doi: 10.1155/2019/4246350

Filadi, R., Greotti, E., Turacchio, G., Luini, A., Pozzan, T., and Pizzo, P. (2015). Mitofusin 2 ablation increases endoplasmic reticulum-mitochondria coupling. *Proc. Natl. Acad. Sci. U.S.A.* 112, E2174–E2181. doi: 10.1073/pnas.1504880112

Friedman, J. R., Lackner, L. L., West, M., DiBenedetto, J. R., Nunnari, J., and Voeltz, G. K. (2011). ER tubules mark sites of mitochondrial division. *Science* 334, 358–362. doi: 10.1126/science.1207385

Friedman, J. R., Webster, B. M., Mastronarde, D. N., Verhey, K. J., and Voeltz, G. K. (2010). ER sliding dynamics and ER-mitochondrial contacts occur on acetylated microtubules. *J. Cell Biol.* 190, 363–375. doi: 10.1083/jcb.200911024

Garrido-Maraver, J., Loh, S. H. Y., and Martins, L. M. (2020). Forcing contacts between mitochondria and the endoplasmic reticulum extends lifespan in a drosophila model of Alzheimer's disease. *Biol. Open* 9:bio047530. doi: 10.1242/bio.047530

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). "Mask R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)* (Venice). doi: 10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas). doi: 10.1109/CVPR.2016.90

Hirabayashi, Y., Kwon, S.-K., Paek, H., Pernice, W. M., Paul, M. A., Lee, J., et al. (2017). ER-mitochondria tethering by PDZD8 regulates Ca2+ dynamics in mammalian neurons. *Science* 358, 623–630. doi: 10.1126/science.aan6009

Huang, G. B., Scheffer, L. K., and Plaza, S. M. (2018). Fully-automatic synapse prediction and validation on a large data set. *Front. Neural Circ.* 12:87. doi: 10.3389/fncir.2018.00087

Huang, X., Sun, L., Ji, S., Zhao, T., Zhang, W., Xu, J., et al. (2013). Kissing and nanotunneling mediate intermitochondrial communication in the heart. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2846–2851. doi: 10.1073/pnas.1300741110

Januszewski, M., Kornfeld, J., Li, P. H., Pope, A., Blakely, T., Lindsey, L., et al. (2018). High-precision automated reconstruction of neurons with flood-filling networks. *Nat. Methods* 15:605. doi: 10.1038/s41592-018-0049-4

Jorstad, A., and Fua, P. (2015). "Refining mitochondria segmentation in electron microscopy imagery with active surfaces," in *Computer Vision - ECCV 2014 Workshops*, eds L. Agapito, M. M. Bronstein, and C. Rother (Cham: Springer International Publishing), 367–379. doi: 10.1007/978-3-319-16220-1_26

Karbowski, M., and Youle, R. (2003). Dynamics of mitochondrial morphology in healthy cells and during apoptosis. *Cell Death Differ.* 10:870. doi: 10.1038/sj.cdd.4401260

Krols, M., Detry, S., Asselbergh, B., Almeida-Souza, L., Kremer, A., Lippens, S., et al. (2018). Sensory-neuropathy-causing mutations in ATL3 cause aberrant ER membrane tethering. *Cell Rep.* 23, 2026–2038. doi: 10.1016/j.celrep.2018.04.071

Lee, J. E., Westrate, L. M., Wu, H., Page, C., and Voeltz, G. K. (2016). Multiple dynamin family members collaborate to drive mitochondrial division. *Nature* 540, 139–143. doi: 10.1038/nature20555

Li, W., Liu, J., Xiao, C., Deng, H., Xie, Q., and Han, H. (2018). A fast forward 3D connection algorithm for mitochondria and synapse segmentations from serial EM images. *Biodata Mining* 11:24. doi: 10.1186/s13040-018-0183-7

Liang, X., Wei, Y., Shen, X., Jie, Z., Feng, J., Lin, L., et al. (2016). "Reversible recursive instance-level object segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas). doi: 10.1109/CVPR.2016.75

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu). doi: 10.1109/CVPR.2017.106

Liu, C., Yuen, J., Torralba, A., Sivic, J., and Freeman, W. T. (2008). "Sift flow: dense correspondence across different scenes," in *European Conference on Computer Vision* (Marseille: Springer), 28–42. doi: 10.1007/978-3-540-88690-7_3

Liu, J., Li, W., Xiao, C., Hong, B., Xie, Q., and Han, H. (2018). "Automatic detection and segmentation of mitochondria from SEM images using deep neural network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 628–631. doi: 10.1109/EMBC.2018.8512393

Liu, Y., and Zhu, X. (2017). Endoplasmic reticulum-mitochondria tethering in neurodegenerative diseases. *Transl. Neurodegener.* 6:21. doi: 10.1186/s40035-017-0092-6

Lucchi, A., Becker, C., Márquez Neila, P., and Fua, P. (2014). "Exploiting enclosing membranes and contextual cues for mitochondria segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014*, eds P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe (Cham: Springer International Publishing), 65–72. doi: 10.1007/978-3-319-10404-1_9

Lucchi, A., Li, Y., and Fua, P. (2013). "Learning for structured prediction using approximate subgradient descent with working sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland), 1987–1994. doi: 10.1109/CVPR.2013.259

Lucchi, A., Li, Y., Smith, K., and Fua, P. (2012a). "Structured image segmentation using kernelized features," in *Computer Vision-ECCV 2012*, eds A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid (Berlin; Heidelberg: Springer), 400–413. doi: 10.1007/978-3-642-33709-3_29

Lucchi, A., Smith, K., Achanta, R., Knott, G., and Fua, P. (2012b). Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features. *IEEE Trans. Med. Imaging* 31, 474–486. doi: 10.1109/TMI.2011.2171705

Manfredi, G., and Kawamata, H. (2016). Mitochondria and endoplasmic reticulum crosstalk in amyotrophic lateral sclerosis. *Neurobiol. Dis*. 90, 35–42. doi: 10.1016/j.nbd.2015.08.004

Marchi, S., Patergnani, S., and Pinton, P. (2014). The endoplasmic reticulum-mitochondria connection: One touch, multiple functions. *Biochim. Biophys. Acta Bioenerget*. 1837, 461–469. doi: 10.1016/j.bbabio.2013.10.015

Márquez-Neila, P., Kohli, P., Rother, C., and Baumela, L. (2014). "Non-parametric higher-order random fields for image segmentation," in *Computer Vision-ECCV 2014*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 269–284. doi: 10.1007/978-3-319-10599-4_18

Naon, D., Zaninello, M., Giacomello, M., Varanita, T., Grespi, F., Lakshminaranayan, S., et al. (2016). Critical reappraisal confirms that mitofusin 2 is an endoplasmic reticulum-mitochondria tether. *Proc. Natl. Acad. Sci. U.S.A*. 113, 11249–11254. doi: 10.1073/pnas.1606786113

Narasimha, R., Ouyang, H., Gray, A., McLaughlin, S. W., and Subramaniam, S. (2009). Automatic joint classification and segmentation of whole cell 3D images. *Pattern Recogn*. 42, 1067–1079. doi: 10.1016/j.patcog.2008.08.009

Phillips, M. J., and Voeltz, G. K. (2016). Structure and function of ER membrane contact sites with other organelles. *Nat. Rev. Mol. Cell Biol*. 17, 69–82. doi: 10.1038/nrm.2015.8

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, CA: Curran Associates, Inc.), 91–99.

Rigamonti, R., Lepetit, V., and Fua, P. (2014). Beyond kernelboost. *CoRR abs/1407.8518*.

Rodriguez-Arribas, M., Yakhine-Diop, S., Bravo-San Pedro, J., Gomez-Suaga, P., Gomez-Sanchez, R., Martinez-Chacon, G., et al. (2017). Mitochondria-associated membranes (MAMs): overview and its role in Parkinson's disease. *Mol. Neurobiol*. 54, 6287–6303. doi: 10.1007/s12035-016-0140-8

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*, eds N. Navab, J. Hornegger, W. M. Wells, and A. Frangi (Cham: Springer International Publishing), 234–241. doi: 10.1007/978-3-319-24574-4_28

Rowland, A. A., and Voeltz, K. G. (2012). Endoplasmic reticulum-mitochondria contacts: function of the junction. *Nat. Rev. Mol. Cell Biol*. 13, 607–615. doi: 10.1038/nrm3440

Schaefer, S., McPhail, T., and Warren, J. (2006). Image deformation using moving least squares. *ACM Trans. Graph*. 25, 533–540. doi: 10.1145/1141911.1141920

Sun, X., Wang, Y., Zhang, J., Tu, J., Wang, X., Su, X., et al. (2012). Tunneling-nanotube direction determination in neurons and astrocytes. *Cell Death Dis*. 3:e438. doi: 10.1038/cddis.2012.177

Twig, G., Elorza, A., Molina, A. J. A., Mohamed, H., Wikstrom, J. D., Walzer, G., et al. (2008). Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. *EMBO J*. 27, 433–446. doi: 10.1038/sj.emboj.7601963

van der Bliek, A. M., Sedensky, M. M., and Morgan, P. G. (2017). Cell biology of the mitochondrion. *Genetics* 207, 843–871. doi: 10.1534/genetics.117.300262

Vincent, A. E., Ng, Y. S., White, K., Davey, T., Mannella, C., Falkous, G., et al. (2016). The spectrum of mitochondrial ultrastructural defects in mitochondrial myopathy. *Sci. Rep*. 6:30610. doi: 10.1038/srep30610

Vitaladevuni, S., Mishchenko, Y., Genkin, A., Chklovskii, D., and Harris, K. (2008). "Mitochondria detection in electron microscopy images," in *Workshop on Microscopic Image Analysis with Applications in Biology* (New York, NY).

Wang, X., and Schwarz, T. L. (2009). Imaging axonal transport of mitochondria. *Nat. Methods* 457, 319–333. doi: 10.1016/S0076-6879(09)05018-6

Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., et al. (2018). Automatic mitochondria segmentation for EM data using a 3D supervised convolutional network. *Front. Neuroanat*. 12:92. doi: 10.3389/fnana.2018.00092

Xie, H., Du, Y., Yu, H., Chang, Y., Xu, Z., and Tang, Y. (2018a). Open set face recognition with deep transfer learning and extreme value statistics. *Int. J. Wavel. Multiresol. Inform. Process*. 16:1850034. doi: 10.1142/S0219691318500340

Xie, Q., Chen, X., Shen, L., Li, G., Ma, H., and Han, H. (2018b). Brain microstructure reconstruction based on deep learning. *Syst. Eng. Theory Pract*. 38, 482–491. doi: 10.12011/1000-6788(2018)02-0482-10

Zinsmaier, K. E., Babic, M., and Russo, G. J. (2009). Mitochondrial *Transport Dynamics in Axons and Dendrites*. Berlin; Heidelberg: Springer. doi: 10.1007/400_2009_20

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Liu, Li, Yang, Hong, Chen, Xie and Han. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/fnins.2020.00676

The Importance of Anti-correlations in Graph Theory Based Classification of Autism Spectrum Disorder

Amirali Kazeminejad[1,2]* *and Roberto C. Sotero*[2,3]

[1] Department of Biomedical Engineering, University of Calgary, Calgary, AB, Canada, [2] Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada, [3] Department of Radiology, University of Calgary, Calgary, AB, Canada

With the release of the multi-site Autism Brain Imaging Data Exchange, many researchers have applied machine learning methods to distinguish between healthy subjects and autistic individuals by using features extracted from resting state functional MRI data. An important part of applying machine learning to this problem is extracting these features. Specifically, whether to include negative correlations between brain region activities as relevant features and how best to define these features. For the second question, the graph theoretical properties of the brain network may provide a reasonable answer. In this study, we investigated the first issue by comparing three different approaches. These included using the positive correlation matrix (comprising only the positive values of the original correlation matrix), the absolute value of the correlation matrix, or the anticorrelation matrix (comprising only the negative correlation values) as the starting point for extracting relevant features using graph theory. We then trained a multi-layer perceptron in a leave-one-site out manner in which the data from a single site was left out as testing data and the model was trained on the data from the other sites. Our results show that on average, using graph features extracted from the anti-correlation matrix led to the highest accuracy and AUC scores. This suggests that anti-correlations should not simply be discarded as they may include useful information that would aid the classification task. We also show that adding the PCA transformation of the original correlation matrix to the feature space leads to an increase in accuracy.

Keywords: autism spectrum disorder, fMRI, machine learning, graph theory, anti-correlations, neural networks, GSR

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Vahid Salari,
University of the Basque Country, Spain
Reza Khosrowabadi,
Shahid Beheshti University, Iran

***Correspondence:**
Amirali Kazeminejad
amirali.kazeminejad@ucalgary.ca

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *09 May 2019*
Accepted: *02 June 2020*
Published: *07 August 2020*

Citation:
Kazeminejad A and Sotero RC (2020) The Importance of Anti-correlations in Graph Theory Based Classification of Autism Spectrum Disorder. Front. Neurosci. 14:676. doi: 10.3389/fnins.2020.00676

INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that is growing in prevalence in recent years (Zablotsky et al., 2015). While it is usually diagnosed by carefully monitoring a child's behavioral development (Barbaresi et al., 2006), recent studies have shown that brain imaging can also be used to aid in that diagnosis by identifying underlying differences between the ASD and Healthy Control (HC) Brain (Dichter, 2012; Bos et al., 2014).

Functional Magnetic Resonance Imaging (fMRI) is a widely used tool for such studies due to its high spatial resolution. It monitors the changes in the Blood Oxygen Level Dependent (BOLD) signal which indirectly measures the neuronal activity (Sotero and Trujillo-Barreto, 2007). By computing functional connectivity measures between BOLD signals from different brain areas,

researchers can examine the human brain at a network level. A variant of the technique called Resting State fMRI (rs-fMRI) has been widely used to examine brain networks while subjects are at rest (Fox and Greicius, 2010; van den Heuvel and Hulshoff Pol, 2010). Graph theory is one of the more novel methods being used for the network-level analysis of the brain. It provides a mathematical framework for quantifying network characteristics and quantitatively analyzing the differences between different brain networks (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010). Machine learning is another relatively new technique that is being applied to rs-fMRI data to extract insights such as important biomarkers as well as to develop novel algorithms with the hope of automatically diagnosing brain disorders from medical imaging data (Pereira et al., 2009; Lee et al., 2013). The network characteristics that is provided by graph theory has been used as the input to machine learning algorithms to identify diseases such as Alzheimer's (Khazaee et al., 2016), Parkinson's (Kazeminejad et al., 2017), and Autism (Kazeminejad and Sotero, 2019b).

One of the variables that can affect the results of the mentioned analysis is how the graph construction step was performed. After the initial preprocessing of the rs-fMRI data, a correlation matrix is created by calculating the correlation between the activation time-series of different brain regions. The correlation matrix is then transformed into a sparse binary matrix representing the existence connections between different regions. This transformation is usually performed by a thresholding step in which only a percentage of the strongest correlations are kept (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010). However, this step ignores the anti-correlations which may be biologically relevant (Fox et al., 2005). This problem will be even more severe if the preprocessing pipeline includes the regression of the global mean signal (GSR) from the time-series, known to result in the removal of motion, cardiac and respiratory signals, which may result in the presence of more anticorrelations in the correlation matrix (Murphy and Fox, 2017). The use of GSR is a controversial matter in the field. Although there is evidence suggesting that the anticorrelations introduced through GSR have no biological basis (Murphy et al., 2009), a recent study has shown using GSR in the preprocessing pipeline for rs-fMRI leads to better prediction accuracies of behavioral measures (Li et al., 2019).

There has been evidence of network-level changes in the ASD brain compared to a HC brain (Redcay et al., 2013; Rudie et al., 2013; Bos et al., 2014). Therefore, using graph theory to extract features for classification of ASD is likely to provide good results. This claim was previously examined by training a SVM classifier on graph theoretical features to classify between ASD and HC (Kazeminejad and Sotero, 2019b). However, the methodology of that paper was limited to using only the positive correlations in order to construct the brain graph, potentially ignoring some informative connections.

The release of the Autism Brain Imaging Data Exchange I (ABIDE I) dataset (Di Martino et al., 2014) allowed researchers to examine ASD in large sample sizes. Nielsen et al. (2013) conducted one of the earliest classification studies on ABIDE and were able to achieve an accuracy of 60% over all samples. More recently, Heinsfeld et al. (2018) were able to achieve an accuracy of 70%, evaluated by 10 fold cross-validation, on the entire dataset by utilizing neural networks and transfer learning. They also reported the leave-one-site out performance of their model averaging at 65% accuracy. Plitt et al. (2015) achieved a higher accuracy, 69.71% when using only 178 subjects from the ABIDE I dataset. Another study by Khosla et al. (2018) used 3D convolutional neural networks to achieve cross validated accuracies as high as 71.7% on a subset of the ABIDE I dataset.

A recent study by Hallquist and Hillary (2019), examining 106 papers (with only 2 papers focusing on ASD) using graph theoretical measures, shows that 79.2% of graph theoretical studies either did not specify how they handled negative correlations or discarded them. Another 9.4% used absolute values of the correlation matrix. However, this is mostly an arbitrary choice.

In this paper, we study the effect of different approaches to handling anti-correlations for the classification accuracy of a machine learning model on the ABIDE dataset. We trained a regularized deep learning neural network using features extracted from transforming the correlation matrix using three pipelines: The positive correlation pipeline which does not change the matrix, the anti-correlation pipeline which prioritizes negative correlations, and the absolute value pipeline which disregards the sign of the correlations. These pipelines are explained further in **Figure 1**. Our model was evaluated using a leave-one-site out approach. Our results suggest that on average, the anti-correlation pipeline results in better accuracy and area under curve (AUC) score with a small loss in sensitivity. Interestingly, adding the PCA transformation of the original correlation matrix increased the accuracy, sensitivity, specificity and AUC score for all pipelines.

MATERIALS AND METHODS

Data and Preprocessing

This study used a publicly available dataset from the ABIDE Preprocessed Initiative (Cameron et al., 2013a). To ensure that our results were not affected by any custom preprocessing pipeline, we used the preprocessed data provided by ABIDE in the C-PAC (Cameron et al., 2013b) pipeline. The preprocessing included the following steps. The Analysis of Functional Neuro Images (AFNI) (Cox, 1996) software was used for removing the skull from the images. The brain was segmented into three tissues using FMRIB Software Library (FSL) (Smith et al., 2004). The images were then normalized to the MNI 152 stereotactic space (Mazziotta et al., 2001; Grabner et al., 2006) using Advanced Normalization Tools (ANTs) (Avants et al., 2011). Functional preprocessing included motion and slice-timing correction as well as the normalization of voxel intensity. Nuisance signal regression included 24 parameters for head motion, CompCor (Behzadi et al., 2007) with 5 principal components for tissue signal in Cerebrospinal fluid and white matter, linear and quadratic trends for Low-frequency drifts and a global bandpass filter (0.01–0.1 Hz). GSR was also applied to remove the global mean from the signals. These images where then co-registered to

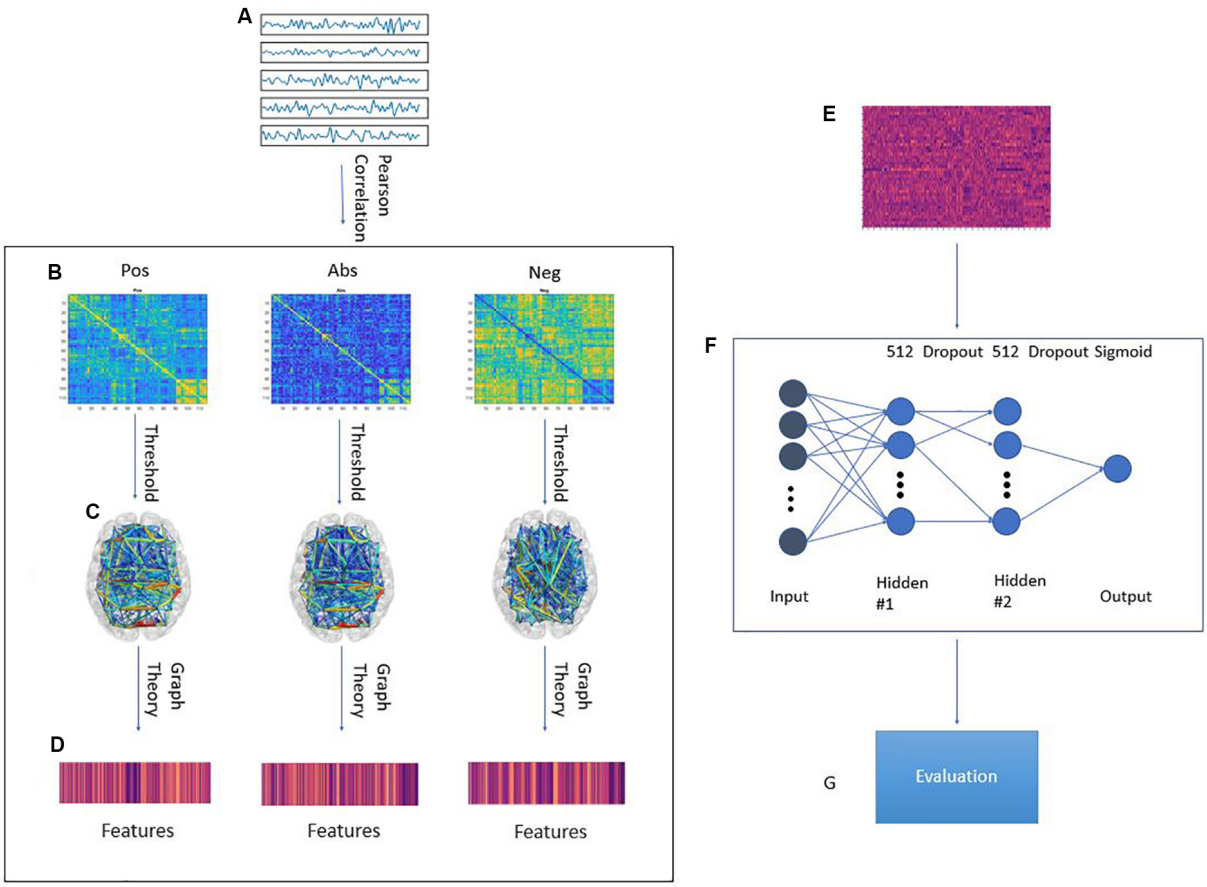


FIGURE 1 | Graphical framework of the experiment. **(A)** By averaging the BOLD activity in each ROI in the parcellation atlas, a time series is extracted representing brain activity in that region; **(B)** Using Pearson's correlation, a connectivity matrix is generated from the ROI time series quantifying the connectivity level between individual ROIs; The connectivity matrix is transformed using three pipelines: Pos. No change to the connectivity matrix, Neg: Multiplying the connectivity matrix by −1, Abs: calculating the absolute value of the matrix. Then, by treating the ROIs as graph nodes and the connectivity matrix as graph weights the brain network is expressed in graph form; **(C)** A threshold is applied to keep only the 20–50% strongest connections in 2% increments; **(D)** Graph theoretical analysis is applied to the resulting graph from to obtain a feature vector for each subject; **(E)** Feature matrix for all subjects in the training fold; **(F)** MLP architecture. First ReLU layer is l1-regularized and second ReLU layer is l2-regularized. A dropout of 0.7 was applied between the first and second ReLU as well as the second ReLU and output; **(G)** The model is tested on a previously unseen test set from a different site in the ABIDE dataset. The evaluation metrics are: Accuracy, Sensitivity, Specificity, AUC score.

their anatomical counterpart by FSL. They were then normalized to the MNI 152 space using ANTs. The average voxel activity in each Region of Interest (ROI) of the Craddock 200 atlas (Craddock et al., 2012) was then extracted as the time-series for that region. Furthermore, in order to replicate our results on the Craddock 200 atlas, we reran the pipelines on the anatomical automatic labeling (AAL) (Tzourio-Mazoyer et al., 2002) atlas. Overall, this study had 1,035 subjects, the demographics of whom are outlined in **Table 1**.

Network Construction, Graph Extraction and Feature Extraction

To construct the brain network, the timeseries for each atlas ROI where correlated, using Pearson correlation, with the other regions. The strengths of these correlations were used as the strengths of the connection between different ROIs. In graph terms, this represents a fully connected graph with each of the nodes being in the center of the corresponding ROIs and each edge weight is the correlation between the two nodes on the opposite ends of the edge. Three different pipelines were obtained based on selecting only the positive values, only the negative values, or taking the absolute value of the correlation matrix. These graphs were then subjected to a thresholding step in which only the 20–50% strongest connections were set to one and the other edges were discarded. This threshold was incremented in 2% steps. This resulted in sparse binary graphs. This step was done because binary graphs have been shown to have more easily defined null models and are more easily characterizable (Rubinov and Sporns, 2010). Several measures of integration (characteristic path length and efficiency), segregation (clustering coefficient and transitivity), centrality (betweenness centrality, eigenvector centrality, participation coefficient and within module z-score) and resilience (assortativity) (Bullmore and Sporns, 2009) were

TABLE 1 | Subject demographics.

	Age Mean (SD)				Count			
Group	**ASD**		**HC**		**ASD**		**HC**	
Sex	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
Site								
CALTECH	24.35 (6.91)	28.27 (11.08)	22.77 (4.31)	29.51 (11.83)	4	15	4	14
CMU	27.00 (6.93)	26.18 (5.88)	26.00 (5.29)	27.10 (6.12)	3	11	3	10
KKI	9.61 (1.85)	10.12 (1.38)	9.49 (0.81)	10.22 (1.23)	4	16	8	20
LEUVEN	14.00 (1.41)	18.18 (5.09)	13.56 (0.75)	19.01 (5.05)	3	26	5	29
MAX_MUN	35.33 (8.50)	24.76 (15.27)	32.00 (X)	24.37 (8.81)	3	21	1	27
NYU	18.62 (9.14)	14.14 (6.59)	15.37 (6.45)	15.76 (6.10)	10	65	26	74
OHSU	–	11.43 (2.18)	–	10.09 (1.12)	–	–	–	14
OLIN	19.00 (5.00)	16.06 (3.04)	17.50 (3.54)	16.54 (3.78)	3	16	2	13
PITT	12.60 (0.95)	20.03 (7.38)	16.07 (3.60)	19.36 (6.97)	4	25	4	23
SBL	–	35.00 (10.43)	–	33.73 (6.61)	–	15	–	15
SDSU	12.44 (X)	14.89 (1.70)	13.26 (2.69)	14.58 (1.46)	1	13	6	16
STANFORD	10.02 (1.63)	9.99 (1.68)	9.01 (1.01)	10.19 (1.66)	4	15	4	16
TRINITY	–	16.81 (3.17)	–	17.08 (3.77)	–	22	–	25
UCLA	12.46 (3.42)	13.08 (2.34)	12.82 (0.89)	13.03 (2.04)	6	48	6	38
UM	13.53 (3.12)	13.11 (2.30)	14.78 (2.82)	14.82 (3.85)	9	57	18	56
USM	–	23.46 (8.33)	–	21.29 (8.35)	–	46	–	25
YALE	12.88 (3.00)	12.70 (3.14)	13.52 (2.64)	12.34 (2.79)	8	20	8	20

M: Male, F: Female, ASD: Autism spectrum disorder, HC: Healthy control, SD: Standard deviation. Some sites did not have any female individuals, these have been identified with a "–." The term "X" for SD means that there were not enough samples to calculate SD.

then extracted from the binary brain network graph. These steps were done using the Python libraries brainconn ([1]FIU-Neuro/brainconn) and network[2]. This resulted in 3 feature sets of 1,404 features for each group. A breakdown of these features can be found in section 6 of the **Supplementary Material**. The main body of this paper focuses mainly on only using the graph features in the analysis. In the **Supplementary Material**, we report the results of adding the PCA transformation (Bartholomew, 2010) of the unaltered correlation matrix to the graph features. These principle components were calculated for each individual and we have 600 components per individual.

Leave-One-Site-Out Cross-Validation

The structure of the ABIDE I dataset allows for an interesting cross-validation approach that captures the multi-site nature of it. Our data was divided into 17 cross-validation sets. In each one, one site was used as the test set and the other 16 sites were used as the training set.

Model Training and Evaluation

In this study, we used a multilayer perceptron with two hidden layers. The model was implemented with Keras and using the Tensorflow 1.13 backend (Khosla et al., 2018). This model was chosen due to its ability to construct data driven features before using them for the final classification task. Both layers consisted of 512 rectified linear neurons. Due to the limited number of samples, the model was heavily regularized. The first hidden layer is subjected to L1 regularization in order to force the network to have some feature selection capabilities. A dropout layer is then applied before the second hidden layer. L2 regularization was used in the second hidden layer. Finally, a dropout layer was included before the output layer. The output layer is a single neuron which is activated by a sigmoid function.

The model training process is as follows. First, each feature was standardized by removing the mean and scaling to unit variance. Then the training data was ran through the neural network. The model used binary cross-entropy as its loss function and an Adam optimizer with $\beta_1 = 0.99$, $\beta_2 = 0.01$, and an initial learning rate of $l = 0.0002$. The learning rate was decayed by a factor of 0.5 base on the validation loss calculated on a random 10% of the training data withheld during the process. The model parameters were tuned on only one of the left out sites (PITT) to ensure low information leak. The trained model was then evaluated on the test set. We repeated the above steps 5 times for each Leave-One-Site-Out (LOSO) fold and report the average accuracy, sensitivity, specificity and AUC score over the 5 repetitions. It is worth noting that the healthy subjects were given a label of 1 and the ASD subjects were given label 0. Thus, sensitivity should be interpreted as the percentage of HCs correctly identified. Likewise, specificity is the percentage of ASDs correctly identified.

We used two approaches to aggregate the results of each left out site. First, we report the overall average metric for each of these sites as "average." We also report a weighted average by using the sample size of each left out site as the weight of that site.

[1]https://github.com

[2]https://networkx.github.io/

Statistical Evaluation

We performed two statistical analyses to compare model results for each left out site between different pipelines. First, we compared the average of the 5 trials performance of the different models for each pipeline for each site against each other using a Welch's *t*-test (Welch, 1947). The average was calculated by computing the mean of each metric for each trial over the different thresholds. This was corrected for multiple comparisons using the Holms-Sidak method (Cardillo, 2006). We also compared model performance in a threshold-wise manner, comparing each threshold results for each pipeline against the same threshold results in another pipeline using the same test. Furthermore, we also report the *p*-value of the mean of each performance metrics between the three pipelines by conducting a paired *t*-test based on the results of individual sites. The significance level is set to $p < 0.05$ and all tests are two-tailed.

RESULTS

In order to manage the different thresholds, we report the average performance measures and their standard deviation across the different thresholds. This applies to all following paragraphs unless explicitly specified. The effects of this choice are discussed further in section "Discussion."

Our results in **Table 2** show that, on average, the negative correlation pipeline can achieve higher accuracies and AUC score than the other two pipelines. The average sensitivity of the model remains comparable over the pipelines while being slightly lower for the absolute value pipeline. The model was able to achieve a higher specificity, increasing its ability to correctly identify ASDs. This is interesting because it suggests different feature extraction pipelines will allow some flexibility between interchanging sensitivity and specificity.

The same results held when adding the PCA transformation of the original correlation matrix as input features. However, the overall performance of the model was improved. Suggesting that information from the original correlation matrix supplements the information available in the graph features (**Supplementary Table 1**). To test whether these observations were dependent on the choice of the atlas, we applied the same methodology on a different atlas When using the AAL parcellation, the negative pipeline performed the best as well. Thus, strengthening the hypothesis that this result holds across different parcellation atlases (**Supplementary Table 2**). Another variable in the dataset was gender. The negative correlation pipeline was still the best performing pipeline on average when only modeling the male subjects. However, the difference between the absolute correlation and negative correlation pipelines where negligible (AUC of 0.59 vs 0.592). Interestingly, overall model performance was lowered when the training and test data only included from male subjects (**Supplementary Table 3**).

As evident in the performance tables, there is a high variability in model performance between sites. One possible explanation for this may be related to the age of the subjects in that site. To investigate this, we plotted the average AUC scores of the 5 trials for each site against their age means. A quadratic regression was applied on this data and is shown in **Figure 2**. In the case of added PCA features, the sites at opposing ends of the age mean scale performed worse than the sites closer to the mean. This phenomenon was less pronounced in the case of using only graph features with it being almost non-existent in the case of the positive correlation pipeline. Furthermore, we report the R-squared values and *F*-test *p*-values of these plots. While most regression curves show statistical significance, the quadratic regression is shown to be a poor fit for the results from the positive pipeline when using only graph features.

Our results show that using different pipelines may play a significant role in the classification outcome for certain sites. **Figure 3** shows how these pipelines compare across different site when the same threshold is applied. We report statistically significant differences between the specificity of the Yale site model between the absolute value and positive pipelines as well as between the anticorrelation and absolute value pipeline. Such significant differences can be spotted elsewhere for the PITT site in the case of absolute value vs anticorrelation pipeline as well as to a more limited scale across different sites and pipelines. This suggests that between site differences such as subject demographics and equipment differences could be used to inform the choice of which pipeline to use.

Supplementary Tables 4.1 and **4.2** show the *p*-values of between pipeline differences compared to the anti-correlation pipeline for the graph-features-only method. When comparing the best models, based on what threshold led to the highest AUC score, trained with only graph features for each site, the negative pipeline achieved a higher AUC than the other pipelines except in the following sites. OLIN, OHSU, NYU and SBL for the unchanged pipeline and OHSU and NYU for the absolute value pipeline. Of these sites, none were significantly different ($p < 0.05$) between the unchanged and anticorrelation pipelines. For accuracy, the same comparison holds between the anticorrelation and unchanged pipelines. OSHU ($p = 0.042$) and NYU ($p = 0.043$) results were the only significantly different sites. The absolute value pipeline was able to achieve better accuracies than the negative pipeline for the following site: OLIN, OHSU, NYU. Likewise, the AAL atlas results did not show many statistical significances in the site accuracies between different pipelines.

As it may not be fair to compare different thresholds for each site, we also performed threshold matched comparisons for the graph-only craddock-200 atlas results. A paired t-test was performed on the performance metrics of each pipeline pair for each site. **Figure 3** shows these comparisons in a heatmap style plot. Interestingly, no site showed consistent difference in performance (AUC) over the tested threshold range, suggesting that this value should be treated as a parameter than can affect the accuracy of machine learning predictions using graph theory and tuned for the task at hand.

Another paired *t*-test was performed comparing the out of site fold performance of the three pipelines with just graph features (**Table 3**). The negative correlation pipeline had statistically different accuracy and AUC metrics when compared to the absolute value pipeline. No other significant differences were found between the pipelines.

TABLE 2 | Results with only graph features.

	ACC			AUC			SEN			SPE		
	Abs	**Neg**	**Pos**	**Abs**	**Neg**	**Pos**	**Abs**	**Neg**	**Pos**	**Abs**	**Neg**	**Pos**
CALTECH	0.527 (0.032)	0.547 (0.034)	0.519 (0.029)	0.529 (0.031)	0.55 (0.034)	0.523 (0.029)	0.592 (0.04)	0.671 (0.055)	0.675 (0.07)	0.466 (0.058)	0.429 (0.06)	0.372 (0.065)
CMU	0.575 (0.044)	0.623 (0.052)	0.584 (0.044)	0.581 (0.043)	0.629 (0.053)	0.592 (0.043)	0.723 (0.052)	0.778 (0.083)	0.823 (0.076)	0.438 (0.064)	0.479 (0.054)	0.362 (0.101)
KKI	0.55 (0.037)	0.549 (0.035)	0.491 (0.035)	0.564 (0.036)	0.586 (0.034)	0.512 (0.041)	0.481 (0.052)	0.364 (0.056)	0.389 (0.061)	0.646 (0.052)	0.807 (0.055)	0.635 (0.107)
LEUVEN	0.551 (0.026)	0.608 (0.038)	0.581 (0.04)	0.535 (0.025)	0.596 (0.037)	0.563 (0.042)	0.734 (0.05)	0.744 (0.051)	0.785 (0.038)	0.337 (0.035)	0.448 (0.043)	0.342 (0.066)
MAX_MUN	0.556 (0.024)	0.562 (0.033)	0.557 (0.033)	0.554 (0.024)	0.558 (0.034)	0.554 (0.033)	0.582 (0.063)	0.612 (0.034)	0.602 (0.047)	0.526 (0.059)	0.505 (0.061)	0.505 (0.053)
NYU	0.635 (0.021)	0.615 (0.019)	0.644 (0.019)	0.636 (0.02)	0.618 (0.018)	0.645 (0.019)	0.626 (0.033)	0.602 (0.032)	0.641 (0.028)	0.646 (0.026)	0.633 (0.02)	0.648 (0.025)
OHSU	0.44 (0.038)	0.403 (0.047)	0.516 (0.045)	0.442 (0.039)	0.4 (0.048)	0.515 (0.041)	0.412 (0.047)	0.442 (0.051)	0.532 (0.141)	0.473 (0.074)	0.357 (0.06)	0.497 (0.111)
OLIN	0.619 (0.034)	0.625 (0.042)	0.643 (0.043)	0.603 (0.035)	0.623 (0.043)	0.633 (0.043)	0.463 (0.05)	0.608 (0.074)	0.542 (0.057)	0.743 (0.039)	0.638 (0.051)	0.723 (0.054)
PITT	0.586 (0.029)	0.592 (0.034)	0.562 (0.018)	0.586 (0.029)	0.596 (0.033)	0.565 (0.018)	0.578 (0.048)	0.723 (0.042)	0.653 (0.04)	0.594 (0.046)	0.47 (0.046)	0.476 (0.043)
SBL	0.548 (0.03)	0.607 (0.036)	0.62 (0.04)	0.548 (0.03)	0.607 (0.036)	0.62 (0.04)	0.838 (0.045)	0.793 (0.031)	0.761 (0.056)	0.257 (0.076)	0.42 (0.07)	0.48 (0.058)
SDSU	0.603 (0.027)	0.648 (0.052)	0.668 (0.024)	0.6 (0.028)	0.662 (0.058)	0.68 (0.023)	0.613 (0.039)	0.597 (0.045)	0.63 (0.038)	0.587 (0.051)	0.728 (0.091)	0.729 (0.036)
STANFORD	0.52 (0.03)	0.522 (0.035)	0.527 (0.015)	0.525 (0.03)	0.528 (0.035)	0.533 (0.014)	0.346 (0.069)	0.286 (0.054)	0.299 (0.035)	0.703 (0.063)	0.77 (0.039)	0.766 (0.028)
TRINITY	0.512 (0.032)	0.535 (0.029)	0.444 (0.043)	0.51 (0.032)	0.53 (0.03)	0.44 (0.044)	0.541 (0.053)	0.609 (0.048)	0.494 (0.05)	0.479 (0.044)	0.451 (0.065)	0.387 (0.06)
UCLA	0.628 (0.025)	0.646 (0.024)	0.592 (0.024)	0.631 (0.026)	0.648 (0.023)	0.601 (0.022)	0.656 (0.048)	0.676 (0.027)	0.691 (0.03)	0.605 (0.033)	0.621 (0.036)	0.511 (0.044)
UM	0.644 (0.019)	0.664 (0.017)	0.629 (0.028)	0.647 (0.02)	0.67 (0.017)	0.632 (0.028)	0.584 (0.026)	0.58 (0.029)	0.57 (0.038)	0.711 (0.035)	0.759 (0.029)	0.695 (0.022)
USM	0.618 (0.019)	0.607 (0.042)	0.555 (0.042)	0.642 (0.019)	0.646 (0.04)	0.593 (0.035)	0.725 (0.031)	0.779 (0.043)	0.721 (0.031)	0.56 (0.027)	0.513 (0.051)	0.465 (0.062)
YALE	0.602 (0.023)	0.622 (0.031)	0.618 (0.026)	0.602 (0.023)	0.622 (0.031)	0.618 (0.026)	0.768 (0.035)	0.59 (0.044)	0.655 (0.062)	0.436 (0.048)	0.654 (0.044)	0.581 (0.041)
Mean	0.571	0.587	0.574	0.573	0.592	0.578	0.604	0.615	0.615	0.542	0.569	0.540
Weighted Mean	0.592	0.604	0.587	0.594	0.609	0.592	0.614	0.619	0.622	0.574	0.599	0.561

ACC: ACCURACY; SEN: SENSITIVITY; SPE: SPECIFICITY; AUC: Area under curve score; Pos: Unchanged pipeline; Neg: Anti-correlation pipeline; Abs: Absolute value pipeline.

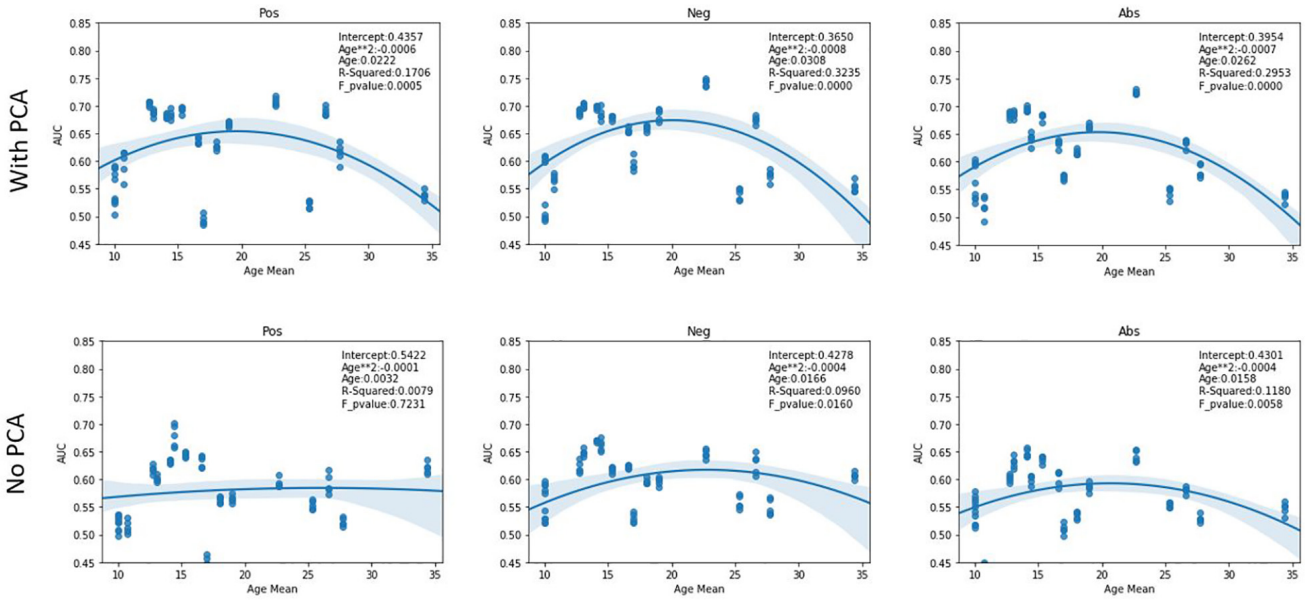


FIGURE 2 | AUC scores vs Age. The figures show the quadratic regression line of AUC score of a LOSO fold based on mean age associated with the left-out site. Top row plots the results from using both PCA and graph features while bottom row shows result from using only graph features. The sites that have extreme age ranges (either lowest or highest) perform worse than those closer to the population median. An inverted U-shape is observed in all pipelines other than the Pos pipeline in the case of using only graph features. Abbreviations: Pos, Positive pipeline; Neg, negative pipeline; Abs, Absolute value pipeline.

Although not the main goal of this study, our model was able to, based on the weighted accuracy, perform on-par with the deep learning model described in Heinsfeld et al. (2018) when using the graph and PCA features. The sites CALTECH, KKI, MAX_MUN, OHSU, SBL, STANFORD and TRINITY showcased lower accuracy in our model.

Section 4 of the **Supplementary Material** reports the LOSO results of 3 other models: Logistic regression, random forest classification and gaussian support vector classifiers. These tables show that our neural network model was able to outperform these 3 classical machine learning algorithms in most sites.

DISCUSSION

Our results consistently showed that, on average, using graph theoretical features from the negative pipeline increases the MLP model accuracy in distinguishing between HC and ASD. This suggests that anti-correlations in the brain network contain important information that helps our model in distinguishing between ASD and HC. One possible reason for the higher performance of the anti-correlation method could be that contrary to some previous studies (Murphy et al., 2009; Murphy and Fox, 2017), GSR introduced anti-correlations may in fact stem from real neurological basis and are not spurious (Khosla et al., 2018). Unfortunately, as we do not know of any gold standards to test this hypothesis in real data, it should only be interpreted as a speculation made based on the results of this study.

Previous studies have concluded that the use of absolute value graph metrics in the presence of GSR may be compromised (Bartholomew, 2010). This was attributed to the fact that in the presence of GSR, the topologies of the anti-correlation and correlation matrix will be mixed when using an absolute value pipeline as the anti-correlations have comparable magnitude to the correlations. However, our model trained with the absolute value graph features was able to perform on-par with the positive correlation pipeline.

The low number of significantly different within-site metrics between the three pipelines may be attributed to the way they were calculated. Each LOSO fold was run 5 times in order to offset any effects of random weight initialization and random training-validation set splits. Increasing the number of trials should add to the strength of this test by capturing the metric distributions more accurately. Another possible explanation is that there is enough information available in all pipelines for most of these sites.

In the case of added PCA, plotting the accuracies of the negative pipeline against number of participants did not provide further insight into why some sites performed worst. Plotting the same against the age of the participants revealed that there may be a link between age and the ability of our model to make accurate predictions. Age has been previously shown to affect network properties of subject within the ABIDE dataset (Henry et al., 2018). KKI and STANFORD are the two sites with the lowest average ASD and HC age. Of the 4 sites with the highest average ASD age, 3 of them, CALTECH, MAX_MUN and SBL showed lower accuracies. Interestingly, CMU, did not follow this behavior. However, the standard deviation of the age of CMU's

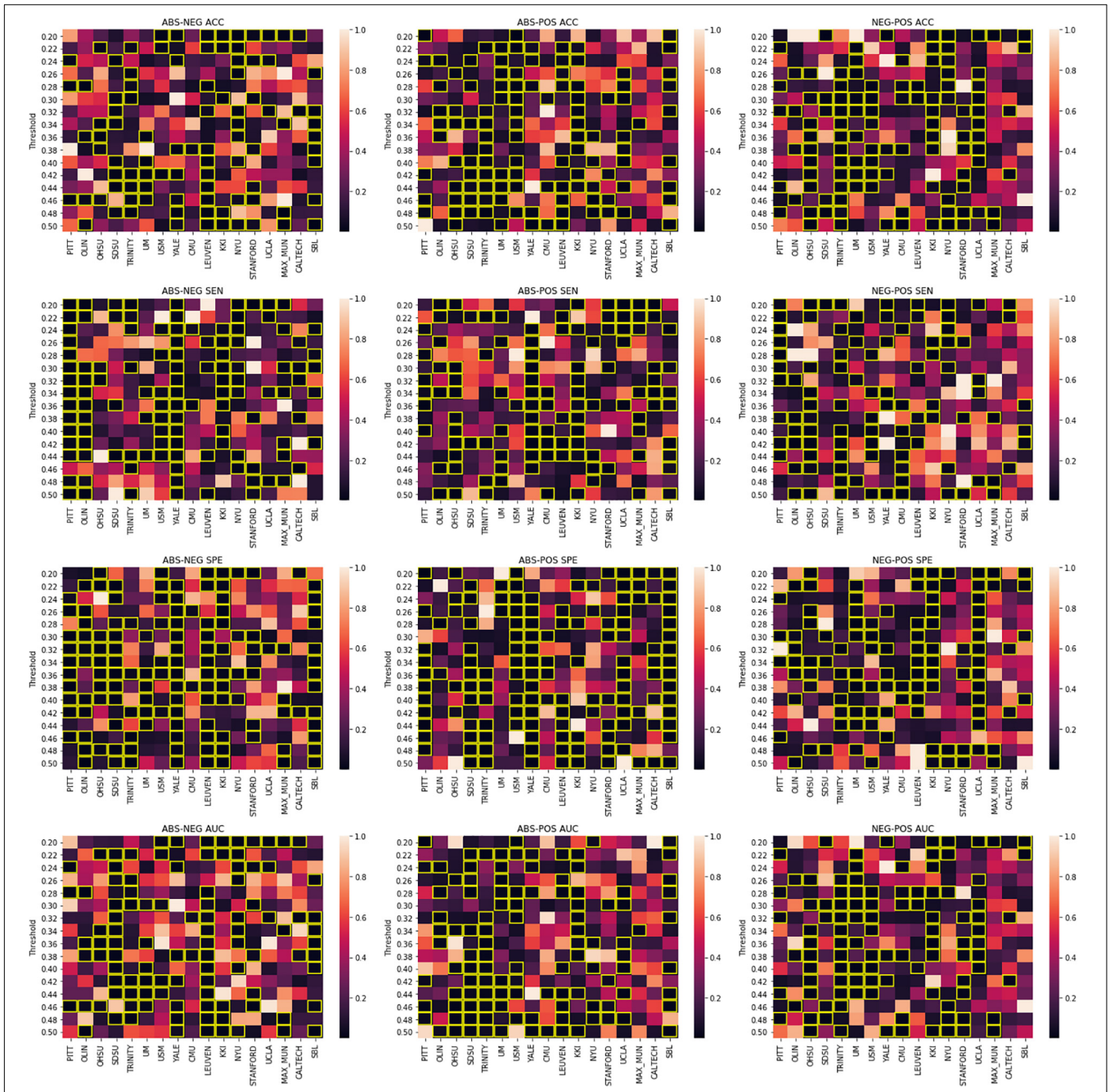


FIGURE 3 | Welch's *t*-test results. Figure columns represent the pipelines being compared and the rows represent the metric that is being compared. In each sub-figure, the title shows the pipelines being compared following by the metric. The X-axis represents different imaging sites and the Y-axis shows different thresholds. The heatmap values are *p*-values distributed based on the colormap to the right of each figure. Significant *p*-values (0.05) are highlighted in yellow.

participants was lower than those of the other 3. The larger age SD may have hindered the algorithm's ability to perform well for the sites with high average age.

Interestingly, the model trained only on the graph features was able to perform relatively well on SBL despite that site having the highest age mean and performed worst on OHSU, the site with the least number of subjects. While this could have happened purely by chance, it is an interesting avenue for further insights about the nature of our model. Furthermore, these models were more robust to the effects of age as shown in **Figure 2** suggesting age-related information is better captured in the graph theoretical features than the PCA transformation of the correlation matrix. However, this conclusion assumes that a quadratic regression fits the performance results of the graph only pipelines. The positive pipeline does not show a significant f-test result, thus perhaps a higher order regression would fit its

TABLE 3 | Paired *t*-test results (*p*-values) between the average metrics of the specified pipelines, corrected by the holm-sidak method.

	ACC	SEN	SPE	AUC
POS-NEG	0.451	0.966	0.367	0.382
POS-ABS	0.846	0.867	0.931	0.645
NEG-ABS	0.086	0.867	0.490	0.028

POS: positive correlation pipeline; NEG: anti-correlation pipeline; ABS: absolute correlation pipeline; ACC, SEN, SPE, AUC like above tables. Green cells show statistical significance.

results better. During our experiments, a third-degree regression did in fact show significant f-test values across all members of the graph only pipelines.

Another intriguing observation about our results is that some thresholds and pipelines were better suited for classifying specific sites. This suggest that an ensemble model, such as training different models on each of these combinations and assigning the most selected label for each subject between these models, may achieve better performance on this dataset.

In this study we utilized a leave-one-site out cross validation approach in order to capture the multi-site nature of the ABIDE I dataset. We believe this approach leads to a better estimate of how this model will perform in the real world as it allows for better generalization across previously unseen imaging sites and protocols.

Using graph theoretical measures to analyze and classify neurological and neurodevelopmental diseases is not a novel idea. However, the practice of thresholding the connectivity matrix in order to produce a binary graph introduces an artificial bias toward some of the correlations. While this bias can be avoided by using the absolute value of the correlation matrix, this may disregard the important information that the positive correlation and anti-correlation networks hold. We have shown that, after running a GSR pipeline, focusing on the anti-correlation network may lead to better classification performance in the case of ASD vs HC. We also showed that by adding non-graph features to the anti-correlation graph features, classification metrics are improved suggesting that a different process may be needed to accurately capture the graph properties of both the positive and negative correlation networks.

One limitation in the present study may be the use of multiple threshold values in constructing the functional graph. Here we reported the average the results of all thresholds for each site. Another approach may be to report results from the threshold resulting in the highest AUC for each site. This would lead to a significant increase in model performance (achieving 70% weighted accuracy when using graph and PCA features). However, the negative correlation pipeline still outperformed the other pipeline.

Another limitation of this study was the inclusion of all age ranges in our dataset. As different sites have different demographics, this may affect the power of our model. Furthermore, identifying ASD is most important in earlier stages of life. The ABIDE I dataset included only a small number of children under the age of 10 (146 subjects) spread across multiple sites and none below the age of 5. Thus, the results presented in this study may not generalize to studies on younger children.

The ABIDE I data exhibits numerous inherent variability due to its multi-site nature. Here we presented a deep learning model that was able to navigate the intricacies of this data and generalize over multiple sites by using graph theoretical features. Our model was able to, on average, perform on-par with previously reported deep learning models using the same number of subjects.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

We used the data collected as part of the ABIDE database and complied with everything that they have asked to be included in any manuscript using that data. The original ethics statement form the (Di Martino et al., 2014) manuscript is as follows: All contributions were based on studies approved by local IRBs, and data were fully anonymized (removing all 18 HIPAA protected health information identifiers, and face information from structural images). All data distributed were visually inspected prior to release.

AUTHOR CONTRIBUTIONS

AK designed the experiments including writing the experiment code as well as writing the manuscript. RS provided advice and guidance on the experimental design and contributed to many manuscript revisions and final proofreading. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was partially funded by the Department of Biomedical Engineering Research and Academic Excellence awards at the University of Calgary.

ACKNOWLEDGMENTS

Part of the code for the preprocessing pipeline was reused from the open source code provided by the Heinsfeld et al. manuscript. This work could not be completed without the immense contribution of open sourcing the ABIDE datasets. This manuscript has been released as a Pre-Print at bioRxiv (Kazeminejad and Sotero, 2019a).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00676/full#supplementary-material

REFERENCES

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Barbaresi, W. J., Katusic, S. K., and Voigt, R. G. (2006). Autism: a review of the state of the science for pediatric primary health care clinicians. *Arch. Pediatr. Adolesc. Med.* 160, 1167–1175.

Bartholomew, D. J. (2010). *Principal Components Analysis in International Encyclopedia of Education*. Berlin: Springer.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Bos, D. J., van Raalten, T. R., Oranje, B., Smits, A. R., Kobussen, N. A., van Belle, J., et al. (2014). Developmental differences in higher-order resting-state networks in Autism Spectrum disorder. *Neuroimage Clin.* 4, 820–827. doi: 10.1016/j.nicl.2014.05.007

Bullmore, E., and Sporns, O. (2009). Erratum: complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 312–312. doi: 10.1038/nrn2618

Cameron, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., et al. (2013a). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* 7:41. doi: 10.3389/conf.fninf.2013.09.00041

Cameron, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S. S., Yan, C., et al. (2013b). Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* 7:42. doi: 10.3389/conf.fninf.2013.09.00042

Cardillo, G. (2006). *Holm-Sidak T-Test: A Routine For Multiple T-Test Comparisons*. Available online at: http://www.mathworks.com/matlabcentral/fileexchange/12786

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* 33, 1914–1928. doi: 10.1002/hbm.21333

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667.

Dichter, G. S. (2012). Functional magnetic resonance imaging of autism spectrum disorders. *Dialog. Clin. Neurosci.* 14, 319–351.

Fox, M. D., and Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Front. Syst. Neurosci.* 4:19. doi: 10.3389/conf.fninf.2013.09.00019

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. (2006). *Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults*. Berlin: Springer.

Hallquist, M. N., and Hillary, F. G. (2019). Graph theory approaches to functional network organization in brain disorders: a critique for a brave new small-world. *Netw. Neurosci.* 3, 1–26. doi: 10.1162/netn_a_00054

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Henry, T. R., Dichter, G. S., and Gates, K. (2018). Age and gender effects on intrinsic connectivity in autism using functional integration and segregation. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* 3, 414–422. doi: 10.1016/j.bpsc.2017.10.006

Kazeminejad, A., Golbabaei, S., and Soltanian-Zadeh, H. (2017). "Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI," in *Proceedings of the 2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz.

Kazeminejad, A., and Sotero, R. (2019a). The importance of anti-correlations in graph theory based classification of autism spectrum disorder. *bioRxiv* [Preprint]. Available online at: https://www.biorxiv.org/content/10.1101/557512v1 (accessed May 10, 2019).

Kazeminejad, A., and Sotero, R. C. (2019b). Topological properties of resting-state fMRI functional networks improve machine learning-based autism classification. *Front. Neurosci.* 12:1018. doi: 10.3389/conf.fninf.2013.09.01018

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2016). Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. *Brain Imaging Behav.* 10, 799–817. doi: 10.1007/s11682-015-9448-7

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2018). "3D convolutional neural networks for classification of functional connectomes," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. 137–145.

Lee, M. H., Smyser, C. D., and Shimony, J. S. (2013). Resting-state fMRI: a review of methods and clinical applications. *AJNR Am. J. Neuroradiol.* 34, 1866–1872. doi: 10.3174/ajnr.a3263

Li, J., Kong, R., Liégeois, R., Orban, C., Tan, Y., Sun, N., et al. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *Neuroimage* 196, 126–141. doi: 10.1016/j.neuroimage.2019.04.016

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A four-dimensional probabilistic atlas of the human brain. *J. Am. Med. Inform. Assoc.* 8, 401–430.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage* 44, 893–905. doi: 10.1016/j.neuroimage.2008.09.036

Murphy, K., and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* 154, 169–173. doi: 10.1016/j.neuroimage.2016.11.052

Nielsen, J. A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. *Front. Hum. Neurosci.* 7:599. doi: 10.3389/conf.fninf.2013.09.0599

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209.

Plitt, M., Barnes, K. A., and Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *YNICL* 7, 359–366. doi: 10.1016/j.nicl.2014.12.013

Redcay, E., Moran, J. M., Mavros, P. L., Tager-Flusberg, H., Gabrieli, J. D. E., and Whitfield-Gabrieli, S. (2013). Intrinsic functional network organization in high-functioning adolescents with autism spectrum disorder. *Front. Hum. Neurosci.* 7:573. doi: 10.3389/conf.fninf.2013.09.0573

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Rudie, J. D., Brown, J. A., Beck-Pancer, D., Hernandez, L. M., Dennis, E. L., Thompson, P. M., et al. (2013). Altered functional and structural brain network organization in autism. *Neuroimage Clin.* 2, 79–94. doi: 10.1016/j.nicl.2012.11.006

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Berg, H. J., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, S208–S219.

Sotero, R. C., and Trujillo-Barreto, N. J. (2007). Modelling the role of excitatory and inhibitory neuronal activity in the generation of the BOLD signal. *Neuroimage* 35, 149–165. doi: 10.1016/j.neuroimage.2006.10.027

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM Using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

van den Heuvel, M. P., and Hulshoff Pol, H. E. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

Welch, B. L. (1947). The generalization of 'student's' problem when several different population varlances are involved. *Biometrika* 34, 28–35. doi: 10.1093/biomet/34.1-2.28

Zablotsky, B., Black, L. I., Maenner, M. J., Schieve, L. A., and Blumberg, S. J. (2015). Estimated prevalence of autism and other developmental disabilities following questionnaire changes in the 2014 national health interview survey. *Natl. Health Stat. Rep.* 87, 1–20.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Kazeminejad and Sotero. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 27 August 2020
doi: 10.3389/fnhum.2020.00320

Gender Perception From Gait: A Comparison Between Biological, Biomimetic and Non-biomimetic Learning Paradigms

Viswadeep Sarangi[1], Adar Pelah[1], William Edward Hahn[2] and Elan Barenholtz[2]*

[1] Department of Electronic Engineering, University of York, York, United Kingdom, [2] Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL, United States

This paper explores in parallel the underlying mechanisms in human perception of biological motion and the best approaches for automatic classification of gait. The experiments tested three different learning paradigms, namely, biological, biomimetic, and non-biomimetic models for gender identification from human gait. Psychophysical experiments with twenty-one observers were conducted along with computational experiments without applying any gender specific modifications to the models or the stimuli. Results demonstrate the utilization of a generic memory based learning system in humans for gait perception, thus reducing ambiguity between two opposing learning systems proposed for biological motion perception. Results also support the biomimetic nature of memory based artificial neural networks (ANN) in their ability to emulate biological neural networks, as opposed to non-biomimetic models. In addition, the comparison between biological and computational learning approaches establishes a memory based biomimetic model as the best candidate for a generic artificial gait classifier (83% accuracy, $p < 0.001$), compared to human observers (66%, $p < 0.005$) or non-biomimetic models (83%, $p < 0.001$) while adhering to human-like sensitivity to gender identification, promising potential for application of the model in any given non-gender based gait perception objective with superhuman performance.

Keywords: motion perception, biological motion, gait, machine learning, human perception, machine perception

OPEN ACCESS

Edited by:
Reza Lashgari,
Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Jean-Francois Daneault,
Rutgers, The State University of New Jersey, United States
Yousef Salimpour,
Johns Hopkins University, United States

****Correspondence:***
Adar Pelah
adar.pelah@york.ac.uk

Specialty section:
This article was submitted to Motor Neuroscience, a section of the journal Frontiers in Human Neuroscience

Received: *21 August 2019*
Accepted: *20 July 2020*
Published: *27 August 2020*

Citation:
Sarangi V, Pelah A, Hahn WE and Barenholtz E (2020) Gender Perception From Gait: A Comparison Between Biological, Biomimetic and Non-biomimetic Learning Paradigms. Front. Hum. Neurosci. 14:320. doi: 10.3389/fnhum.2020.00320

INTRODUCTION

A person's gait carries information about the individual along multiple dimensions. In addition to indicating biologically intrinsic properties, like gender and identity, the gait of a person changes dynamically based on their emotional state (Pollick et al., 2002) and state of health (Cesari et al., 2005). Humans are adept at identifying whether a given sparse motion pattern is biological or not (Johansson, 1973, 1976) as well as detecting properties such as gender or mood. However, the origin of these abilities remains unclear. One the one hand, the ability to distinguish biological from non-biological motion appears at a very young age (Fox and McDaniel, 1982), suggesting there may be some expert-system capacities present at birth. Indeed, some theorists have suggested that biological motion perception served as an evolutionary and developmental precursors to the theory of the mind (Frith, 1999). However, recognition of biological motion could also be attributed to generic learning systems that are trained with experience (Fox and McDaniel, 1982;

Bertenthal and Pinto, 1994; Thompson and Parasuraman, 2012), where adults have been shown to be able identify biological motion which was synthetically created using machines (while the infants could not), indicating a learning system that tunes itself based on experience. One way to address this is to compare the behavior of human observers to computational learning models of different types that can be trained "from scratch," i.e., without specialized mechanisms pre-tuned to the properties of biological motion. This provides the opportunity to assess whether generic learning models trained on biological motion stimuli serve as a reasonable model of human behavior or whether additional mechanisms, such as pre-tuned expert systems, should be posited. In addition, we classify the computational models into two groups: (1) biomimetic models, that functionally replicate the neural learning systems in humans, especially biological memory, and (2) non-biomimetic models, which utilize statistical techniques to identify discerning features in data for classification. For clarification, we use the term "biomimetic" as, the study of the structure and function of living things as models for the creation of materials or products by reverse engineering (Farber, 2010). A strong resemblance between humans and biomimetic models, but not with non-biomimetic models, would provide further evidence that the mechanisms underlying human biological motion perception are well captured by a generic learning model. To do so, we compared performance in a simple binary biological motion classification task (gender recognition) and compared human performance to a range of computational learning models.

Biological Models

Johansson (1973, 1976) first demonstrated that human observers were sensitive to biological motion through point light displays representing joints of a human walkers. Despite their sparsity, human observers readily interpreted the stimuli as human gait. Subsequent research with point-light walkers demonstrated the ability of humans in the identification of familiar people (Loula et al., 2005). In case of unfamiliarity, observers could extract certain general categories such as approximate age and gender (Kozlowski and Cutting, 1977; Barclay et al., 1978; George and Murdoch, 1994; Lee and Grimson, 2002; Pollick et al., 2005) with significantly higher than chance accuracy. In case of gender identification, humans achieved the best performance when presented with the stimuli in the coronal plane, due to the prevalence of dynamic cues (George and Murdoch, 1994). However, the biological nature of human perception hinders its replicability and transferability. The learning is highly variable, volatile and susceptible to fatigue, illness and mortality, leading to the need for automation of gait classification. Automation of gait classification has been extensively studied with a high focus on performance outcome through classification accuracy. However, mimicking human perception closely would ensure versatility of the artificial classifier, enabling the same classifier to be used in other non-gender related gait classification tasks.

Machine Learning Models

Machine learning (ML) models can be trained to identify the relevant attributes in gait such as gender with high speed and fidelity. The models can broadly be divided into two categories: (1) Memory based models comprised of artificial neural networks (ANN) such as the Long Short Term Memory (LSTM) cells (Graves and Schmidhuber, 2005), which operate on time series data, and (2) Static models, such as the Random Decision Forests (RDFs) (Kaur and Bawa, 2017) and Support Vector Machines (SVMs) (Huang et al., 2014), which operate on static data. The LSTM model shall be referred to as the "biomimetic" models crediting the functional implementation of the biological neural network and memory using artificial neurons, while the SVM and RDF shall be referred to as the "non-biomimetic" models. Prior studied have promising results in terms of the ability of biomimetic ML in being able to mimic human observation of gait (Pelah et al., 2019; Sarangi et al., 2019; Stone et al., 2019). However, an in-depth exploration of the above mentioned models and direct comparison to human observers on the same stimuli has not been conducted.

Biomimetic Models

Artificial neural networks aim to mimic the flow of information in the biological brain by creating a network of neurons, based on the perceptron model (Rosenblatt, 1958). Recurrent neural networks (RNN), an implementation of the ANN, simulates the memory capabilities of the human brain, by creating an additional feedback loop for processing latent network state along with new data (Mikolov and Zweig, 2012). RNNs operate on a sequence of vectors as input data. The sequence resembles the time series information and the vector represents the features of the input at each timestamp. However, RNNs suffer from vanishing and exploding gradient rendering them ineffective in processing long sequences (Graves and Schmidhuber, 2005). LSTM cells overcome this problem by introducing additional gates in the network to regulate the flow of information, enabling them to remember relevant temporal patterns over long periods of time (Graves and Schmidhuber, 2005).

Non-biomimetic Models

The non-biomimetic ML techniques considered in this paper learn to classify information based on (1) linear separability, following non-linear projections, and (2) reduction of information entropy based on feature thresholds. SVM models learn to fit a linear hyperplane to maximize the separation between the classes in the training dataset. Decision trees, learn to classify information based on the learned numerical thresholds of features. RDF is a collection of randomly initialized decision trees with majority vote of the cohort considered as the predicted class. SVMs and RDFs accept static representations of data as input and thus cannot process temporal sequences of information, unlike humans and LSTMs.

DATA COLLECTION

Forty one consenting healthy adults (26 male, 15 female) between the ages of 18 and 50 years old were recorded walking on the treadmill. Participants volunteered and received credit toward a participation grade for their class. Appropriate

consent forms were signed and anonymity maintained. Gait data was recorded as spatiotemporal three-dimensional joint trajectories for 20 tracked joints of the body. The tracked points on the walker's skeleton included the head, neck, shoulders, elbows, wrists, fingertips, mid spine, back, hips, knees, hips, ankles, and toes. The collection of the joint positions formed a static frame. Data was captured at 24 frames per second, each frame represented by 60 numbers (3D coordinates of 20 joints) and a corresponding timestamp of capture of the frame. Data was recorded for 6 sessions per participant. Each session consisted of a minute of walking on the treadmill at a self-selected speed followed by a minute's rest. The joints were extracted utilizing a consumer-level time-of-flight based RGB-D sensor, the Microsoft Kinect v2. The sensor provides an anthropomorphic representation of the human skeleton through 3D joint coordinates. The sensor was placed approximately 1.5 m in front of the treadmill with the front board removed to avoid issues with occlusion. The ML based skeletal motion capture method mentioned in Shi et al. (2018) is used for capturing the PLD representation of the biological motion of the walkers. When compared with the state-of-art optical motion tracking methods [such as Vicon (Clark et al., 2012; Pfister et al., 2014)], the anatomical landmarks from the Kinect-generated point clouds can be measured with high test-retest reliability, and the differences in the interclass coefficient correlation between Kinect and Vicon are <0.16 (Clark et al., 2012, 2013; Pfister et al., 2014; van Diest et al., 2014). Both systems have been shown to effectively capture >90% variance in full-body segment movements during exergaming (van Diest et al., 2014). The validity of biological motion captured using the Kinect v2 sensor is established in Shi et al. (2018) with human observers through reflexive attentional orientation and extraction of emotional information from the upright and inverted PLD.

EXPERIMENT 1: BIOLOGICAL MODELS

Studies have shown humans to require no longer than two complete gait cycles to correctly identify gender from human gait (Huang et al., 2014). In terms of duration, this translates to less than 2.7 s of walking animation. Although humans can decipher biological motion from point light animation of walking human figure within 200 msec, at least 1.6 s of stimulus is required for significantly above chance performance. This experiment aims to establish the change in gender identification performance in humans as a function of increasing duration of stimulus exposure.

Method

Biological Model

Fifteen female and six male healthy observers with age ranging from 20 to 43 years old, participated in the experiment. All had some experience of biological motion displays, although none had been required to make judgments about gender.

Stimuli

A PC-compatible computer monitor with a high performance raster graphics system displayed stimuli on an Iiyama ProLite B2283HS color monitor (1920 × 1080 resolution, 60 Hz refresh rate). Human figures were defined by 20 circular white dots of 5 pixel radius overlaid on a black background, located on the head, neck, shoulders, elbows, wrists, fingertips, back, spine, hips, knees, ankles, and toes. None of the dots were occluded by other subjective parts of the figure. Animated sequences were created by placing the dots at the three-dimensional trajectory of each of the 20 tracked joints, and temporally sampling the coordinates to produce 24 static frames per second, as shown in **Figure 1**. The stimulus size was 6 degrees wide and 8 degrees long for the whole frame, including zero (black) padding. A degree of visual angle is defined as the subtended angle at the nodal point of the eye. The actual walking clip was 2.5 degrees wide and 4 degrees long. The presented stimuli was height normalized to fit the given aspect ratio, to prevent the observers from identifying gender based on height of the animated walker.

When the static frames were played in quick succession, a vivid impression of a walking person emerged. There was no progressive component to the walking animation, thus the human figure appeared to walk on an unseen treadmill with the walking direction oriented toward the observer. The height range for males was 144–208 cm, and for females was 129–152 cm. None were notably over- or underweight (see **Table 1**). The x and z component were sampled to display the walker in the coronal plane to emphasize lateral sway and maximize the provision of dynamic cues to the observer (Barclay et al., 1978; Troje, 2002). The recorded gait sequences were converted into an animation sequence in the same fashion to be presented as visual stimuli. The observers were seated in a well-lit room in front of the monitor and had access

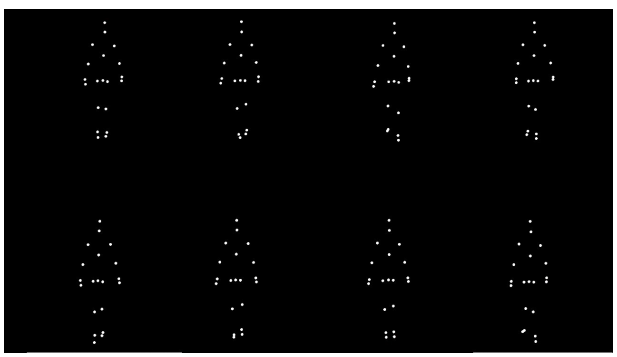


FIGURE 1 | Point light representation of a walking stimulus at eight different stages of a gait cycle.

TABLE 1 | Description of the walking subjects taking part in the stimulus set.

	Height (cm)	Weight (kg)	Age (years)
Male	176.23 ± 32.43	80.49 ± 2.86	26.06 ± 6.42
Female	128.56 ± 23.51	73.3 ± 4.59	21.29 ± 1.23

to a standard computer mouse for interaction. The randomly chosen walker stimuli were presented for exposure durations of 0.4, 1.5, 2.5, and 3.8 s, followed by an on-screen prompt in the form of two buttons requesting the observer prediction of binary gender through a mouse click on either of the labeled buttons. Following the response from the observer, the next stimulus was presented. A total of 200 walking clips were shown per observer per exposure duration and the responses recorded for each.

Results

Human observers correctly identified 63% of all the trials across all exposure durations, $t(20) = 7.8$, $p < 0.001$, two tailed. All t-tests reported in this paper are two-tailed, unless otherwise indicated. Chance performance for the binary gender classification is 50% as expected. Correct identification at 0.4 s, which consisted of a quarter of a step cycle, was above chance at 60%, $t(20) = 3.7$, $p < 0.01$, conforming with (Barclay et al., 1978; Troje, 2002), however, was in disagreement with (Barclay et al., 1978). This could be attributed to the presentation of the stimulus in the coronal plane as opposed to the sagittal plane (Sarangi et al., 2019), leading to higher emphasis on the dynamic cues. Performance at 1.5 s is 66%, $t(20) = 3.8$, $p < 0.005$, which is higher than the performance at 2.5 s of 61%, $t(20) = 4.8$, $p < 0.001$. Barclay et al. (1978) explains this anomalous phenomenon due to an additional partial step at 2.5 s by highlighting the preferred perception of velocity over positional cues, where sensitivity to gender identification decreases mid-swing in the gait cycle. Humans were able to identify gender with highest accuracy at 3.8 s with 69%, $t(20) = 3.4$, $p < 0.01$. Details of the results obtained have been listed in **Table 2**. Overall, the performance of the human observers taking part in the experiment conforms to the results of perceptual experiments in literature (Barclay et al., 1978; Troje, 2002), providing a reliable baseline for comparison with the biomimetic perception on the same stimulus set.

In summary, human observers were able to identify gender from gait with significantly above chance performance from moving dots presentations of joints, while conforming with existing human perception literature. There is a significant increase in gender identification performance between 0.4 and 3.8 s of stimulus exposure duration. The increased gender sensitivity at 1.5 s is attributed to the prevalence of dynamic, velocity based cues at the phase of the step cycle corresponding to that time (George and Murdoch, 1994), thus demonstrating the preference of humans toward dynamic velocity based cues compared to structural position based cues for gender identification.

TABLE 2 | Gender identification accuracy in % as a function of exposure duration of the stimulus.

Model/Stimulus Duration	0.4 s	1.5 s	2.5 s	3.8 s
Biological (Human)	60 ($p < 0.01$)	66 ($p < 0.005$)	61 ($p < 0.001$)	65 ($p < 0.05$)

EXPERIMENT 2: BIOMIMETIC MODELS

Long short term memory network's capability to process a temporal sequence of data aims to mimic the temporal pattern recognition capabilities of humans. The learning gates inherent in the network parallel the short and long term memory of the human brain, enabling the network to remember the relevant temporal pattern while ignoring patterns that don't contribute toward the classification objective. This experiment aims to present an LSTM network with the temporal evolution of joint trajectories during human gait and train it for gender identification to evaluate for resemblance with human observers.

Method

Biomimetic Model

A standard LSTM model consisting of 128 hidden states in the cell (as shown in **Figure 2**), is initialized. The cell state weights were initialized as a random normal distribution. The final cell state was ReLU activated (Maas et al., 2013) and connected to an affine output layer, which represented the one-hot labeled gender identity of the walker during training. During testing, the output layer represented the prediction values. The error of prediction was evaluated using a cross-entropy function (de Brébisson and Vincent, 2015) for updating of the weights using an Adam optimizer (Kingma and Ba, 2014) based on the error differentials and a learning rate of 0.001. The most probable output was taken as the class label during prediction. 10 LSTM models, mimicking 10 random human perceptions, were generated for inferring the gender from gait input.

Data Input

The three-dimensional trajectories of each of the 20 tracked joints were concatenated to form a vector representation of a static frame with a cardinality of 60, representing the location of the head, neck, shoulders, elbows, wrists, fingertips, mid-back, hips, knees, ankles, and toes. Gait input to the model consisted of a sequence of vector representations of subsequent static frames, sampled at 24 frames per second. Joint trajectories were size normalized (Troje, 2002) and standardized with a zero mean and unit standard deviation. Model training sessions included, initialization of the model weights, prediction of the output probabilities based on the gait input, propagation of the prediction error and

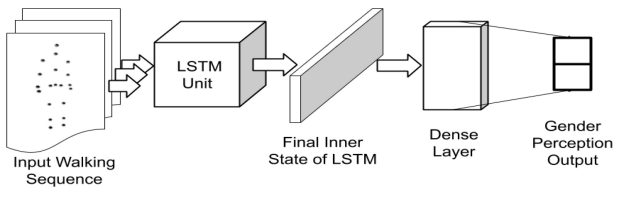


FIGURE 2 | Implementation of the LSTM network architecture for processing gait sequences.

updating the network weights. Model training was executed in batches of 50 and repeated for 100 epochs. Input sequence durations mirrored the exposure durations in the corresponding human perception experiment and varied incrementally for 10 durations from 0.4 to 3.8 s in steps of 0.4 s (10 static frames). 10-fold cross validation was carried out to ensure model generalizability and a total of 250 gender predictions were obtained per input sequence duration. The models trained per session per duration are stored locally for future analyses.

Results

Long Short Term Memory models correctly identified 76% of all the gait inputs presented across all the input durations, $t(9) = 9.2$, $p < 0.001$. Chance performance remains same at 50%. Correct identification at a quarter of a step cycle at 0.4 s was 71%, $t(9) = 5$, $p < 0.001$, higher than the same with human observers, $F(9,20) = 3.6$, $p < 0.1$. The difference in performance indicates a higher inference capacity from a limited amount of available data. The inference performance increases slightly with increase in the amount of information available from 0.4 to 3.8 s, $F(9,9) = 2$, $p < 0.1$. At 3.8 s, the model correctly identified gender with 81% accuracy, $t(9) = 9.6$, $p < 0.001$, considerably higher than human observers, $F(9,20) = 9$, $p < 0.01$. Generalizing across all the input (or exposure) durations, the biomimetic model identified gender with a significantly higher accuracy than the human observers, $F(9,20) = 39.9$, $p < 0.001$. Details of results obtained for the LSTM model have been presented in **Table 3** with the corresponding trend plotted in **Figure 3**. As shown in the figure, mean performance peaks temporarily at 1.6 s (halfway completion of one gait step) with 79% accuracy, $t(9) = 10.1$, $p < 0.001$ suggesting a dependence on dynamic and velocity cues similar to humans at 1.5 s.

In summary, the biomimetic LSTM model performed significantly better than chance in gender classification from 3D moving point representations of human gait. There was a significant increase in gender identification accuracy from 0.4 to 3.8 s of gait information exposure, corresponding to humans. The increased gender sensitivity at 1.6 s could be attributed to an inherent sensitivity to dynamic velocity based cues in LSTM networks for gender identification, similar to humans. One could argue that the presentation of the skeleton stimulus as facing toward the camera could potentially limit real-world applications. However, although specific deployments would need to be assessed, the availability of 3D data could be leveraged to apply a simple preprocessing rotational step to the skeleton to correct for any misalignment in global skeletal configuration.

TABLE 3 | Gender identification accuracy as a function of exposure duration of the stimulus with $p < 0.001$ for all the durations.

Model/Stimulus Duration	0.4 s	1.5 s	2.5 s	3.8 s
Biological (Human)	60 ($p < 0.01$)	66 ($p < 0.005$)	61 ($p < 0.001$)	65 ($p < 0.05$)
Biomimetic (LSTM)	71 ($p < 0.001$)	73 ($p < 0.001$)	77 ($p < 0.001$)	81 ($p < 0.001$)

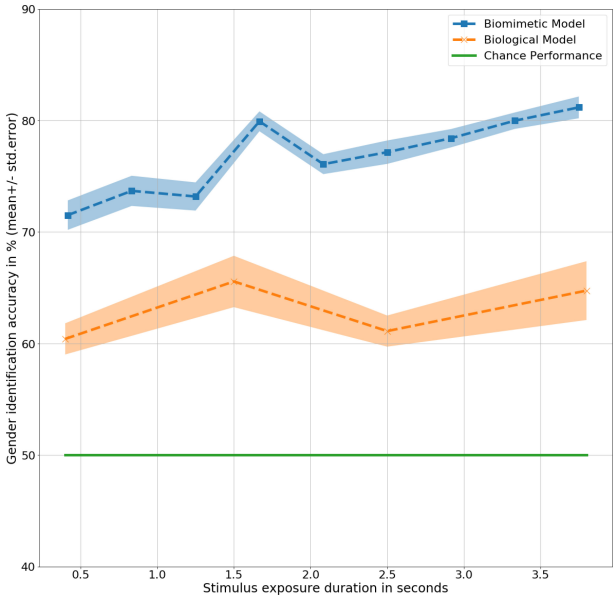


FIGURE 3 | Gender identification performance in mean ± standard error % by the models as a function of exposure duration in seconds.

EXPERIMENT 3: NON-BIOMIMETIC MODELS

The non-biomimetic models, unlike humans and LSTMs are capable of analyzing static data only. Their reliance on the principles of linear separability and information entropy to create rules for classification, resembles expert systems. The models require a static representation of the spatiotemporal gait data for gait classification. Thus data was represented as, (1) static descriptions of the temporal signals, and (2) extracted metrics used in a clinical setting to describe gait for diagnosis and rehabilitation monitoring. In this experiment, we evaluate the SVMs and RDFs on the two static representations of gait data for resemblance with human observation.

Method

Non-biomimetic Models

Support vector machines are designed to linearly separate a set vectors to achieve maximum classification accuracy, whereas DTs learn to classify by choosing optimal split conditions of attributes to minimize ambiguity in classification. SVMs with linear (SVM-Linear), radial basis function (SVM-RBF) with gamma as 0.99 and sigmoid (SVM-Sigmoid) kernels were evaluated. The non-biomimetic models also included RDFs of 10 randomly generated ID3 DTs (Kaur and Bawa, 2017) with a minimum requirement of two samples for splitting and a maximum of three features for splitting consideration.

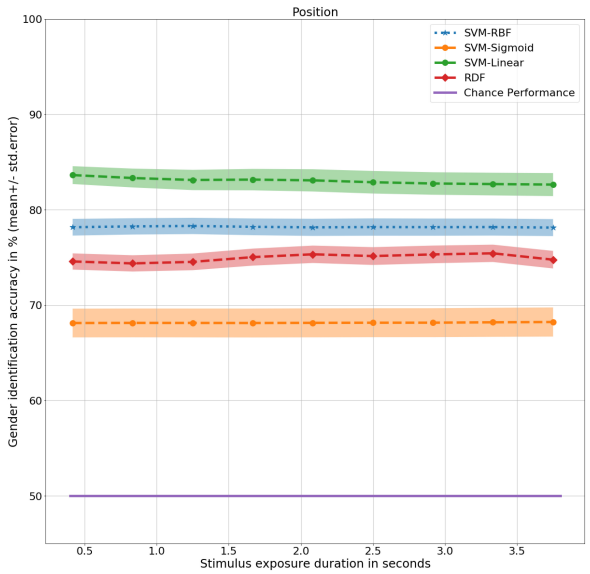


FIGURE 4 | Gender identification performance of non-biomimetic models as a function of duration of gait data used to generate the static representation. The shaded region around the central mean line represents the standard error in performance.

Data Input

Gait data, originally represented as a temporal sequence of vectors, was described with four first-order statistics, namely, minimum, maximum, mean, and standard deviation of each dimension in the multi-dimensional time series signal. The temporal sequence duration of the signal was varied from 0.4 to 3.8 s in steps of 0.4 s (10 frames). The resulting static dataset was normalized and standardized to have a mean of zero and unit standard deviation.

The study was conducted in conjunction with the Cambridge University Hospitals, thus the gait metrics utilized for clinical gait analysis in the gait analysis laboratory were mirrored as static representations of gait. 12 spatiotemporal metrics, including, stride length, cadence, single-double support, stance-swing phase ratio, speed of walking and knee flexion for each leg during stance and swing were provided as input feature sets to the static learning models. All the features were standardized to have a mean value of zero and a standard deviation of one. The features were further normalized to lie within the $[-1, 1]$ range for uniformity and to discourage the models from learning the gender from the structural information and to rely solely on the gait dynamics.

Results

At 0.4 s, the RDFs and SVMs of all the kernels were able to identify gender with significantly better than chance performance with RDF at 75%, $t(9) = 9$, $p < 0.001$, SVM-Linear at 84%, $t(9) = 12$, $p < 0.001$, SVM-RBF at 78%, $t(9) = 10$, $p < 0.001$, and SVM-Sigmoid at 68%, $t(9) = 4$, $p < 0.01$, as shown in **Figure 4**. There was no significant difference in performance in the non-biomimetic models between 0.4 and 3.8 s of duration, unlike humans and LSTM models. Gender sensitivity remained similar across all the durations of gait input, demonstrating a dynamic cue agnostic learning mechanism. Performance output has been detailed in **Table 4**. All the non-biomimetic models performed close to chance performance. The best performing out of the cohort was the SVM with radial basis function with a gender identification accuracy of 59%, $t(9) = 1.8$, $0.1 < p < 0.2$, followed by the RDF classifier with an accuracy of 59%, $t(9) = 1.5$, $0.1 < p < 0.2$. The statistical significance of the results is yet to be established with the collection of more data. However, the motivation for testing the non-biomimetic models further is reduced based on the results.

In summary, the biomimetic models that were trained on the four static representations of the temporal signals performed significantly better than chance and the corresponding models trained on clinical gait metrics performed at or below chance performance. Notably, there was no significant change in performance with increasing duration of exposure of gait input. In addition, there was no change in sensitivity to gender identification with increasing availability of information at different phases of the step cycle. Both the above characteristics have been observed in humans and LSTMs, suggesting a deviation of the non-biomimetic models from a common learning mechanism shared between humans and LSTMs. Humans also possess the trait of being sensitive to dynamic velocity based cues for gender identification. In order to explore the existence of the trait in the artificial classifiers, the next experiment trains models on velocity cues exclusively, to evaluate the change in performance.

EXPERIMENT 4: BIOMIMETIC MODELS WITH VELOCITY CUES

Humans are known to rely on dynamic velocity based cues when determining gender from gait. This experiment focuses on training the biomimetic and non-biomimetic models on velocity cues exclusively. An increase in performance would determine a common trait shared with the humans.

TABLE 4 | Performance of non-biomimetic models in % correctly identified gender.

Model/Duration	0.4 s	1.5 s	2.5 s	3.8 s
SVM-Linear	83.8 ($p < 0.001$)	83.5 ($p < 0.001$)	82.8 ($p < 0.001$)	82.5 ($p < 0.001$)
SVM-RBF	78 ($p < 0.001$)	78 ($p < 0.001$)	78.5 ($p < 0.001$)	77.9 ($p < 0.001$)
SVM-Sigmoid	68 ($p < 0.01$)	68.1 ($p < 0.01$)	68.2 ($p < 0.01$)	68.2 ($p < 0.01$)
RDF	74.9 ($p < 0.001$)	74.6 ($p < 0.001$)	74.2 ($p < 0.001$)	73.5 ($p < 0.001$)

Method

Biomimetic Model

The LSTM model architecture remains the same as mentioned in Experiment 2. The weights initialization, training regime and data input style is maintained. The only difference is brought about because of the difference in the data input being provided.

Biomimetic Data Input

In Experiment 2, gait was represented as a temporal evolution of the positional joint trajectories. For this experiment, temporal derivatives of the gait of the walkers were used for generating corresponding velocities of the joints. The positional data was smoothed with a 5-frame moving average filter before calculating the derivatives for adjacent frames. The training was performed as mentioned in Experiment 2 and 10-fold cross validation ensured generalizability of the results.

Non-biomimetic Model

The non-biomimetic models namely the SVM-Linear, SVM-RBF, SVM-Sigmoid and the RDF remain the same as the previous experiment, however, the training data provided is a static representation of the temporal derivative of the data provided in Experiment 3.

Non-biomimetic Data Input

The positional joint trajectories of 20 tracked joints are smooth using a 5-frame moving average filter followed by a temporal derivative of the smoothed signal to obtain the three-dimensional velocity of the joint trajectories. The result is represented as four static attributes, namely, minimum, maximum, mean and standard deviation of the temporal signal. The temporal duration of the signal is varied from 0.4 to 3.8 s in steps of 0.4 s (10 frames). The static representation is normalized between $[-1, 1]$ and standardized to have zero mean and unit standard deviation. The resulting dataset is used for training and testing of the non-biomimetic models with a 10-fold cross validation of the walkers.

Result

Biomimetic Model

The biomimetic LSTM model trained with three-dimensional velocity (LSTM – Velocity) achieved an overall accuracy of 81%, $t(9) = 9.4$, $p < 0.001$, significantly better than the human observers, $F(9,20) = 82$, $p < 0.001$ for all durations as well as the LSTM model trained with three-dimensional positions (LSTM – Position) of joint trajectories, $F(9,9) = 5.6$, $p < 0.05$. The model achieves its highest accuracy at 2.8 s of exposure with an accuracy of 83%, $t(9) = 9.4$, $p < 0.001$. As shown in **Figure 5**, the lack of velocity based cues is noticed at 2.2 s in LSTM – Velocity, corresponding to the similar lack dynamic cues at the same time in LSTM – Position, demonstrating the dependence of the LSTM network on velocity while determining gender.

Non-biomimetic Models

As shown in **Figure 6**, the performance of the non-biomimetic models decreased significantly upon training with velocity data, compared with the position data (details provided in **Table 5**). SVM-Sigmoid demonstrates a significant increase in performance with increasing duration of gait data provided, $F(9,9) = 4.5$, $p < 0.05$ with the best accuracy of 67%. However, the change in performance is not statistically significant in other SVM and RDF models. The behavior goes against the expected biological behavior and the results demonstrated by the LSTMs, denoting a loss of performance in a form of data which is biologically more conducive to gender identification.

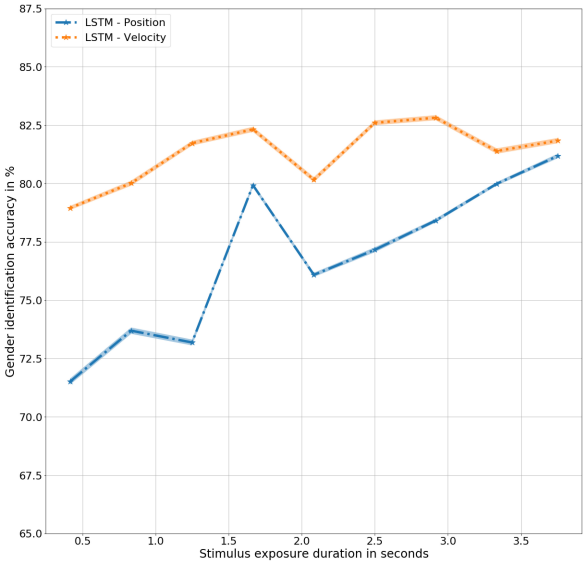


FIGURE 5 | Gender identification performance in mean ± standard error % by LSTM models trained with Position and Velocity as a function of exposure duration in seconds.

As shown in **Figure 6**, the performance of the non-biomimetic models decreased significantly upon training with velocity data, compared with the position data (details provided in **Table 5**). SVM-Sigmoid demonstrates a significant increase in performance with increasing duration of gait data provided, $F(9,9) = 4.5$, $p < 0.05$ with the best accuracy of 67%. However, the change in performance is not statistically significant in other SVM and RDF models. The behavior goes against the expected biological behavior and the results demonstrated by the LSTMs, denoting a loss of performance in a form of data which is biologically more conducive to gender identification.

In summary, training biomimetic and non-biomimetic models with joint velocities produced contradicting results when compared to the corresponding models trained with joint positional trajectories. The result of the non-biomimetic models goes against the established human dependence on dynamic velocity based cues for gender identification. The results obtained through the biomimetic LSTM networks not only conforms to the expected biological behavior, but also demonstrates the shared gender sensitivity trend observed at different phases of the walk cycle between the biological and biomimetic models. However, one could argue that the provision of three-dimensional gait information to the biomimetic models but the two-dimensional screen based input to the humans could cause a gap in direct comparison between the models, making it difficult to

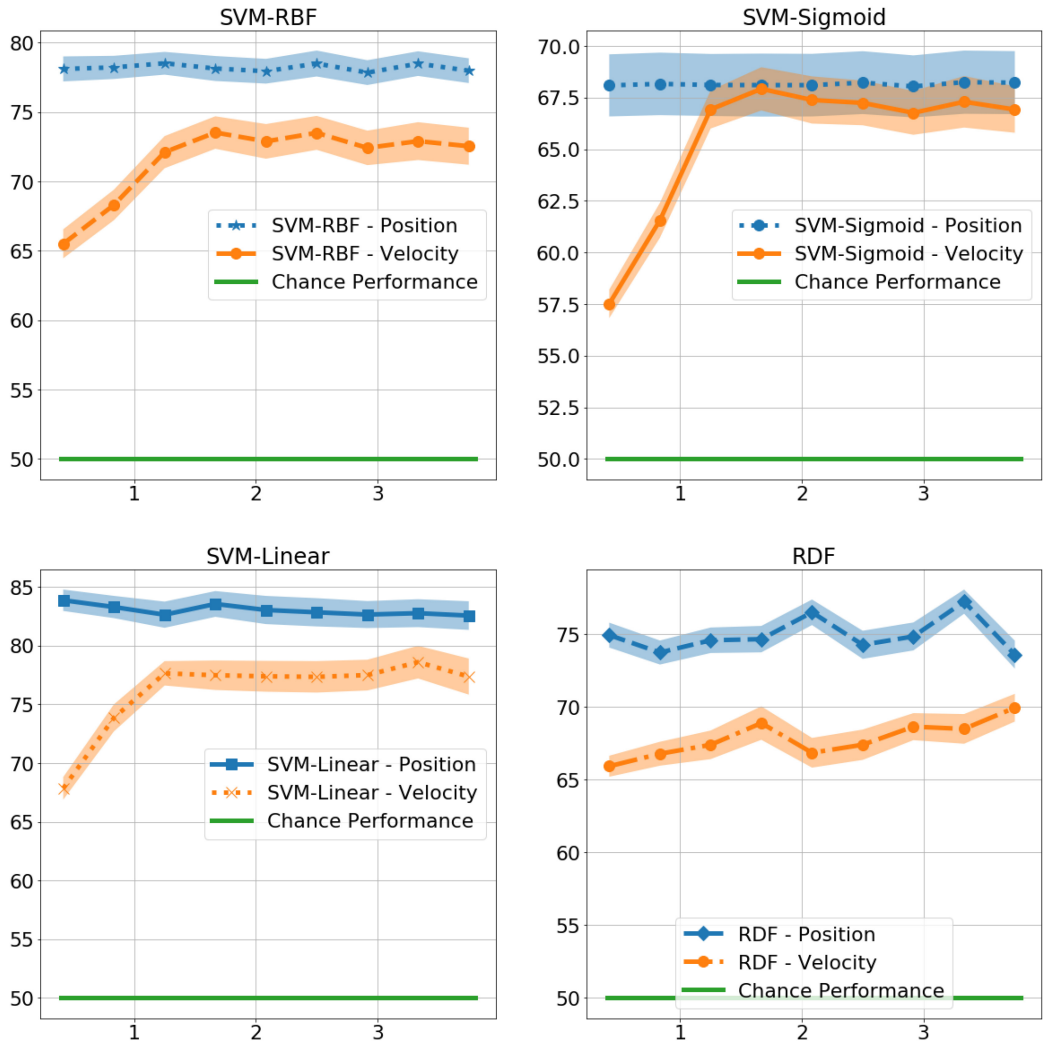


FIGURE 6 | Gender identification performance of non-biomimetic models trained with Position and Velocity information, as a function of duration of gait considered for extracting the static representation.

draw parallels between the learning mechanisms. In the next experiment, the biomimetic LSTM model is trained and tested on two-dimensional gait information by omitting the depth information, which the humans had to infer from the screen.

EXPERIMENT 5: BIOMIMETIC MODELS WITH TWO-DIMENSIONAL INPUT

Long Short Term Memory networks have demonstrated a close resemblance with the human observers, making it conducive to draw parallels between the learning mechanisms. However, humans observed the moving point animations on a two-dimensional screen for gender identification while the LSTMs were provided with three-dimensional motion information. This experiment trains the LSTMs with two-dimensional gait information to form a direct comparison with the human observers, without assuming any depth inference capabilities of humans.

Method

Biomimetic Model

The LSTM model architecture is similar to the previous experiment. However, the number of inputs are reduced by a third, owing to the loss of the z-components of the joint trajectories. The weights initialization, training regime and data input style is maintained. The only difference is brought about because of the difference in the data input being provided.

Data Input

In Experiment 2, gait was represented as a temporal evolution of the positional joint trajectories. The three-dimensional trajectories of each of the 20 tracked joints were concatenated to form a vector representation of a static frame with a cardinality of 60. This experiment maintains the same data style but omitting

TABLE 5 | Gender identification performance of non-biomimetic models trained with Position and Velocity gait information and the difference in performance between the corresponding models denoted through *F*-test.

Model/Gait Data	Position	Velocity	*F*-Test, *F*(9,9)
SVM – Linear	83%	76%	37 ($p < 0.001$)
SVM – RBF	78%	72%	52 ($p < 0.001$)
SVM – Sigmoid	68%	65%	5 ($p < 0.05$)
RDF	75%	68%	149 ($p < 0.001$)

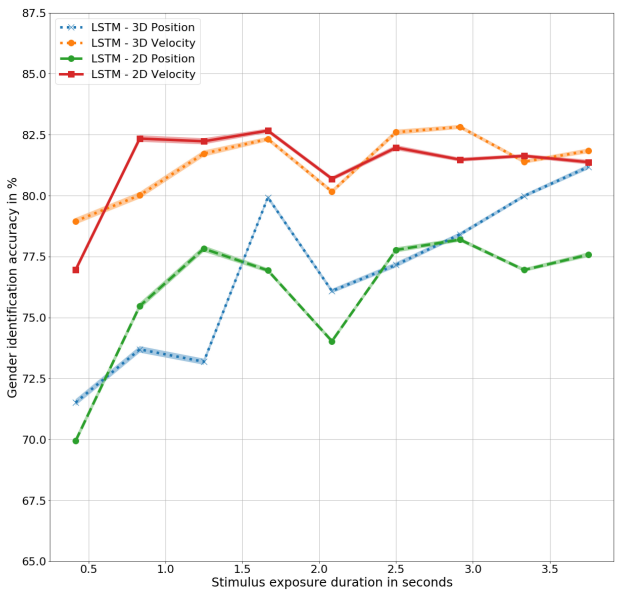


FIGURE 7 | Gender identification accuracy using biomimetic LSTM models trained in three- and two- dimensional position and velocity representations of the joint trajectories.

the z-component, modifying the vector representation to have a cardinality of 40. The gait information is varied from 0.4 to 3.8 s in steps of 0.4 s (10 frames). The resulting data is normalized between [−1, 1] and standardized to have zero mean and unit standard deviation.

Result

There is no statistically significant difference in the outcomes of the LSTMs trained with three- and two-dimensional joint trajectories, in corresponding position and velocity information, as shown in **Figure 7**. However, the difference between the models trained with corresponding 3D and 2D values is significant with $F(9,9) = 6$, $p < 0.05$ for LSTM – 2D Position and LSTM – 2D Velocity, with accuracies of 76% and 80%, respectively. Notably, the performance of the models trained with 3D and 2D velocities are significantly higher than the models trained with the corresponding position representations. The two-dimensional models also demonstrate the unique gender sensitivity trait possessed by human observers and the biomimetic models trained with three-dimensional representations of gait, further supporting the close resemblance of the LSTM models with humans.

In summary, the loss of depth information didn't cause any significant change in performance accuracy all the traits that were shared with humans and corresponding biomimetic models trained with three-dimensional gait information were maintained, further supporting the close resemblance of the artificial model with human perception.

DISCUSSION

Biomimetic models share the following characteristics with human observers, which the non-biomimetic models either do not share or have the opposite trait of: (1) Increase in gender identification performance with increasing temporal availability of gait information, (2) Preference toward dynamic velocity based cues as opposed to structural position based cues for gender identification, leading to higher performance in the former data type, and (3) Unique trend of gender sensitivity during different phases of the walk cycle. The results support a closer congruence in biological motion perception between humans and biomimetic models, compared to non-biomimetic models. Additionally, the close resemblance confirms the ability of the biomimetic learning models in emulating human learning dynamics. This study presents a correlation between the human visual perception of motion as well as the biomimetic memory-based neural network. This, however, does not necessarily indicate causation, but provides a bidirectional mode of understanding human perception using models as simple as the LSTM. The objective of finding correlations between human performance and models for more general use, in a non-arbitrary manner, while minimizing model parameters and assumptions is accomplished. It is interesting that a simple LSTM model is able to capture certain dynamics of human perception of gender from gait from an input-output perspective. This is consistent with previous findings for the inversion effect, a failure to accurately classify gender from gait in upside down stimuli (Sarangi et al., 2020). Observing such correlations between human and biomimetic models encourages further investigations that may help to elucidate both types of "black box" models.

From an application standpoint, the results encourage the potential of using biomimetic models for gait classification. Although the paper uses gender identification as the gait classification objective, the resemblance with human observers may widen the scope of application to non-gender related classification tasks as well. The results could also be further improved through a more conservative approach to cross-validation such as the leave-one-out cross validation. Interestingly, the availability of 3D skeletal data tests the effect of rotation of the skeleton and its correlation with gender identifiability for human observers. The same should not affect the performance of the biomimetic model as a simple preprocessing rotation step could correct for any misalignment in global skeletal configuration. Finally, although treadmill walking may not be congruent with over ground walking, the experiment establishes the ability of the machine based models to extract relevant features from spatiotemporal skeletal data.

The non-biomimetic models on the other hand, require explicitly hand crafted features which may not be applicable to a generic classification task. As shown in Experiment 3, the clinical gait metrics representation of gait didn't possess information about gender, while the static four attribute representation possessed enough information about gender for a better than chance performance. The transferability of the same features toward a new objective (such as person identification) is brought into question and requires further experiments.

CONCLUSION

Humans are highly adept at classification of gait for a multitude of objectives, from gender identification to clinical diagnosis, while relying on a common learning mechanism of spatiotemporal perception of gait. This paper applies a number of ML models, each with a different mode of learning, to the classification of gender from gait, comparing their performance to that of human observers under controlled conditions. The analysis aims to identify, firstly, specific correlations between humans (biological), biomimetic and non-biomimetic learning models, and secondly, to find an artificial model that best resembles human performance to potentially generalize its application to other gait classification tasks. Results are analyzed in terms of performance profiles and preferences of motion cues over structural cues. While not necessarily informative on underlying neural substrates in humans the findings demonstrate the parallel usefulness of the biomimetic approach. Although non-biomimetic and biomimetic models exhibit comparable levels of performance, the biomimetic models are more generalizable in not requiring hand-engineering of features for a given application. Biomimetic are also useful for modeling human perception at least from an input-output perspective, while, conversely, perceptual findings can improve the pragmatic effectiveness of models, contrasting with the synthetic approach that is typically employed in ML research.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available as there are privacy laws regarding data protection.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Internal Review Board, Florida Atlantic University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

VS and AP conceived of the presented idea. VS performed the computational experiments. WH performed the psychophysical experiments with supervision from EB. AP and EB helped fortify the argument in the manuscript. VS wrote the manuscript with help from EB and AP. All authors discussed the results and contributed to the final manuscript.

FUNDING

This research was partially supported by a grant (G0041501) from the Medical Research Council (UK) titled "Clinical Movement Markers of Osteoarthritis for Treatment with Regenerative Medicine".

REFERENCES

Barclay, C. D., Cutting, J. E., and Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Percept. Psychophys.* 23, 145–152. doi: 10.3758/bf03208295

Bertenthal, B. I., and Pinto, J. (1994). Global processing of biological motions. *Psychol. Sci.* 5, 221–225. doi: 10.1111/j.1467-9280.1994.tb00504.x

Cesari, M., Kritchevsky, S. B., Brenda, W. H., Nicklas, B. J., Simonsick, E. M., Newman, A. B., et al. (2005). Prognostic value of usual gait speed in well-functioning older Peopleâ"results from the health, aging and body composition study. *J. Am. Geriatr. Soc.* 53, 1675–1680. doi: 10.1111/j.1532-5415.2005.53501.x

Clark, R. A., Bower, K. J., Mentiplay, B. F., Paterson, K., and Pua, Y.-H. (2013). Concurrent validity of the Microsoft Kinect for assessment of spatiotemporal gait variables. *J. Biomech.* 46, 2722–2725. doi: 10.1016/j.jbiomech.2013.08.011

Clark, R. A., Pua, Y.-H., Fortin, K., Ritchie, C., Webster, K. E., Denehy, L., et al. (2012). Validity of the Microsoft Kinect for assessment of postural control. *Gait Posture* 36, 372–377. doi: 10.1016/j.gaitpost.2012.03.033

de Brébisson, A., and Vincent, P. (2015). An exploration of softmax alternatives belonging to the spherical loss family. *arXiv[preprint]* Available online at: https://arxiv.org/abs/1511.05042 (accessed August 17, 2019).

Farber, D. (2010). Book review: steen Hyldgaard Christensen, Bernard Delahousse, and Martin Meganck (Eds.) engineering in context Aarhus, Denmark: Academica, 2009. ISBN 978-87-7675-700-7. *Bull. Sci. Technol. Soc.* 30, 144–145. doi: 10.1177/0270467610363141

Fox, R., and McDaniel, C. (1982). The perception of biological motion by human infants. *Science* 218, 486–487. doi: 10.1126/science.7123249

Frith, C. D. (1999). Interacting minds–a biological basis. *Science* 286, 1692–1695. doi: 10.1126/science.286.5445.1692

George, M., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 258, 273–279. doi: 10.1098/rspb.1994.0173

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

Huang, X., Shi, L., and Suykens, J. A. K. (2014). Asymmetric least squares support vector machine classifiers. *Comput. Stat. Data Anal.* 70, 395–405. doi: 10.1016/j.csda.2013.09.015

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/bf03212378

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.* 38, 379–393. doi: 10.1007/bf00309043

Kaur, S., and Bawa, R. K. (2017). "Implementation of an expert system for the identification of drug addiction using decision tree ID3 algorithm," in *Proceedings of the 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall)*, (Dehradun: IEEE), doi: 10.1109/icaccaf.2017.8344677

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv[preprint]* Available online at: https://arxiv.org/abs/1412.6980 (accessed August 17, 2019).

Kozlowski, L. T., and Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Percept. Psychophys.* 21, 575–580. doi: 10.3758/bf03198740

Lee, L., and Grimson, W. E. L. (2002). "Gait analysis for recognition and classification," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, (Washington, DC: IEEE), doi: 10.1109/afgr.2002.1004148

Loula, F., Prasad, S., Harber, K., and Shiffrar, M. (2005). Recognizing people from their movement. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 210–220. doi: 10.1037/0096-1523.31.1.210

Maas, A. L., Awni, H. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML* 30:3.

Mikolov, T., and Zweig, G. (2012). "Context dependent recurrent neural network language model," in *Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT)*, (Miami, FL: IEEE), doi: 10.1109/slt.2012.6424228

Pelah, A., Sarangi, V., Hahn, W. E., and Barenholtz, E. (2019). *Do Machines "See" Like Us? A Comparative Study on Classification of Gender from Gait Between Human, Bio-Inspired and Non Bio-Inspired Learning Systems*. London: BMVA Symposium on Visual Image Interpretation in Humans and Machines: Machines that see like us? (accessed April 10, 2019).

Pfister, A., West, A. M., Bronner, S., and Noah, J. A. (2014). Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis. *J. Med. Eng. Technol.* 38, 274–280. doi: 10.3109/03091902.2014.909540

Pollick, F. E., Kay, J. W., Heim, K., and Stringer, R. (2005). Gender recognition from point-light walkers. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1247–1265. doi: 10.1037/0096-1523.31.6.1247

Pollick, F. E., Lestou, V., Ryu, J., and Cho, S.-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Res.* 42, 2345–2355. doi: 10.1016/s0042-6989(02)00196-7

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Sarangi, V., Pelah, A., Hahn, W. E., Barenholtz, E., Stone, T., and Kazakov, D. (2019). *Clinical Evaluation of Machine Learning Approaches for the Classification of 3D Gait Using Static & Dynamic Models in Comparison to Human Perception*. London: BMVA Symposium - Deep Learning in 3 Dimensions (accessed February 20, 2019).

Sarangi, V., Pelah, A., Hahn, W. E., and Barenholtz, E. (2020). Biological and biomimetic perception: a comparative study of gender classification from human gait. *J. Percept. Imaging* 1,

Shi, Y., Ma, X., Ma, Z., Wang, J., Yao, N., Gu, Q., et al. (2018). Using a Kinect sensor to acquire biological motion: toolbox and evaluation. *Behav. Res. Methods* 50, 518–529. doi: 10.3758/s13428-017-0883-9

Stone, T., Sarangi, V., and Pelah, A. (2019). *Application of Skeletal Pose Estimation from Video to Clinical Gait Analysis in Lower Limb Prosthetics in Comparison to Gold Standard Vicon Motion Capture*. London: IET Conference on Human Motion Analysis for Healthcare Applications (accessed June 26, 219).

Thompson, J., and Parasuraman, R. (2012). Attention, biological motion, and action recognition. *NeuroImage* 59, 4–13. doi: 10.1016/j.neuroimage.2011.05.044

Troje, N. F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *J. Vis.* 2, 371–387.

van Diest, M., Stegenga, J., Wörtche, H. J., Postema, K., Verkerke, G. J., and Lamoth, C. J. C. (2014). Suitability of Kinect for measuring whole body movement patterns during exergaming. *J. Biomech.* 47, 2925–2932. doi: 10.1016/j.jbiomech.2014.07.017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sarangi, Pelah, Hahn and Barenholtz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/fnins.2020.00931

Machine Learning Methods for Fluorescence Lifetime Imaging (FLIM) Based Label-Free Detection of Microglia

*Md Abdul Kader Sagar[1,2], Kevin P. Cheng[1], Jonathan N. Ouellette[1,2], Justin C. Williams[1], Jyoti J. Watters[3] and Kevin W. Eliceiri[1,2,4]**

[1] Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI, United States, [2] Laboratory for Optical and Computational Instrumentation, University of Wisconsin-Madison, Madison, WI, United States, [3] Department of Comparative Biosciences, University of Wisconsin-Madison, Madison, WI, United States, [4] Morgridge Institute for Research, Madison, WI, United States

Automated computational analysis techniques utilizing machine learning have been demonstrated to be able to extract more data from different imaging modalities compared to traditional analysis techniques. One new approach is to use machine learning techniques to existing multiphoton imaging modalities to better interpret intrinsically fluorescent cellular signals to characterize different cell types. Fluorescence Lifetime Imaging Microscopy (FLIM) is a high-resolution quantitative imaging tool that can detect metabolic cellular signatures based on the lifetime variations of intrinsically fluorescent metabolic co-factors such as nicotinamide adenine dinucleotide [NAD(P)H]. NAD(P)H lifetime-based discrimination techniques have previously been used to develop metabolic cell signatures for diverse cell types including immune cells such as macrophages. However, FLIM could be even more effective in characterizing cell types if machine learning was used to classify cells by utilizing FLIM parameters for classification. Here, we demonstrate the potential for FLIM-based, label-free NAD(P)H imaging to distinguish different cell types using Artificial Neural Network (ANN)-based machine learning. For our biological use case, we used the challenge of differentiating microglia from other glia cell types in the brain. Microglia are the resident macrophages of the brain and spinal cord and play a critical role in maintaining the neural environment and responding to injury. Microglia are challenging to identify as most fluorescent labeling approaches cross-react with other immune cell types, are often insensitive to activation state, and require the use of multiple specialized antibody labels. Furthermore, the use of these extrinsic antibody labels prevents application in *in vivo* animal models and possible future clinical adaptations such as neurodegenerative pathologies. With the ANN-based NAD(P)H FLIM analysis approach, we found that microglia in cell culture mixed with other glial cells can be identified with more than 0.9 True Positive Rate (TPR). We also extended our approach to identify microglia in fixed brain tissue with a TPR of 0.79. In both cases the False Discovery Rate was around 30%. This method can be further extended to potentially study and better understand microglia's role in neurodegenerative disease with improved detection accuracy.

Keywords: FLIM, CNS, machine learning, brain metabolism, neural networks, microglia, NAD(P)H

OPEN ACCESS

Edited by:
Reza Lashgari, Brain Engineering Research Center, Institute for Research in Fundamental Sciences, Iran

Reviewed by:
Per Niklas Hedde, University of California, Irvine, United States
Hamed Azarnoush, Amirkabir University of Technology, Iran

***Correspondence:**
Kevin W. Eliceiri
eliceiri@wisc.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *15 June 2019*
Accepted: *11 August 2020*
Published: *03 September 2020*

Citation:
Sagar MAK, Cheng KP, Ouellette JN, Williams JC, Watters JJ and Eliceiri KW (2020) Machine Learning Methods for Fluorescence Lifetime Imaging (FLIM) Based Label-Free Detection of Microglia. Front. Neurosci. 14:931. doi: 10.3389/fnins.2020.00931

INTRODUCTION

Unlike external fluorescent labeling approaches, label-free microscopic identification methods can provide equally useful information while leaving the cellular microenvironment unperturbed. Identification of unique metabolic fingerprints based on quantitative data obtained from endogenous cellular properties has been recently explored to develop biomarkers of different cell types and/or disease states. These techniques take advantage of different optical imaging modalities and the intrinsic properties revealed by them followed by quantification techniques to identify different biomarkers. Examples include: diffused optical tomography for breast cancer (Flexman et al., 2013), collagen signature (Kirkpatrick et al., 2006), Stimulated Raman Scattering (SRS) based label-free chemical contrast (Freudiger et al., 2008), and fluorescence lifetime based macrophage signature (Szulczewski et al., 2016).

Fluorescence Lifetime Imaging Microscopy (FLIM) is a well-suited modality for identifying candidate biomarkers as it can be used to assess intrinsic cellular metabolism. Fluorescence lifetime depends on physiological parameters such as pH and ion/oxygen concentrations; it is also independent of intensity, concentration, sample absorption, and sample thickness (Berezin and Achilefu, 2010; Suhling et al., 2015). FLIM can monitor metabolism by taking advantage of the intrinsic fluorescence of the ubiquitous metabolic coenzyme NAD(P)H (Nicotinamide Adenine Dinucleotide). NAD(P)H is a key electron donor/acceptor involved in many metabolic processes, especially redox reactions (Lakowicz et al., 1992; Skala et al., 2007; Mongeon et al., 2016). FLIM can quantify the ratio between free and bound NAD(P)H, and calculate the mean fluorescence lifetime based on the relative quantity of free: bound components and the individual component's lifetime (Lakowicz et al., 1992; Bird et al., 2005; Skala et al., 2005; Provenzano et al., 2008). The mean lifetime, long lifetime component, or free: bound ratios of NAD(P)H are indicative of whether a cell's metabolism is in a more glycolytic or oxidative state. For example, more free NAD(P)H measured via FLIM can be used to show a shift toward glycolysis in cancer per the Warburg theory. As a result, FLIM is gaining widespread acceptance as a way to probe the cellular microenvironment (Wang et al., 1992; Suhling et al., 2005, 2015; Provenzano et al., 2008; Berezin and Achilefu, 2010). FLIM is also increasingly used to probe brain metabolism and neuronal function *in vivo*. For instance, quantitative FLIM data has been used by researchers to (1) find contrast between glioblastoma and normal brain tissue (Leppert et al., 2006; Sun et al., 2010; Kantelhardt et al., 2016), (2) map alterations in cerebral metabolism based on NAD(P)H binding (Chia et al., 2008; Yaseen et al., 2017), (3) non-invasively, optically image Alzheimer's Disease (Das et al., 2018), (4) visualize redox activities in the brain (Mongeon et al., 2016), and (5) quantify neuronal dysfunction in neuroinflammation using FLIM instrumentation (Rinnenthal et al., 2013).

One type of cells of growing imaging interest has been those of the central nervous system (CNS) such as microglia. Microglia are a critical cell type in the nervous system whose activities are implicated in virtually all neuropathologies including traumatic injury, neurodegenerative disease, ischemia, and infection (Watters et al., 2005; Garden and Möller, 2006; Tambuyzer et al., 2009; Charles et al., 2011). These CNS tissue-resident macrophages influence brain development, maintain the neural environment, respond to injury and infection, and orchestrate repair processes among other important functions. Their production of neurotoxic inflammatory molecules often exacerbates neuronal damage. Due to this biological significance and the need for improved tools for further characterization, we chose microglia as our biological use case for developing a non-invasive label free imaging workflow that combines fluorescence lifetime imaging with machine learning.

The goal of this study was to develop a FLIM-based fingerprint for microglia, as has been done for macrophages (Alfonso-García et al., 2016) and bacteria (Bhattacharjee et al., 2017). Our lab previously demonstrated that FLIM-based, label-free imaging could be used to distinguish unique glycolytic type FLIM signatures between tumor-associated macrophages and mouse mammary tumors cells (Szulczewski et al., 2016). However, this has not been done in the context of the CNS and microglia. Previous studies have used FLIM to show neural stem cell differentiation (Stringari et al., 2012) or characterize the metabolic response of astrocytes (Stuntz et al., 2017), but not to identify microglia identity. In order to optimally exploit differences in metabolism-induced lifetime changes for microglia identification, a fast, quantitative approach that can identify their unique metabolic signature would be ideal. Accordingly, we have coupled FLIM with a machine learning-based solution to detect microglia based on their FLIM parameters.

One of the most common ways FLIM acquisition is performed is Time Correlated Single Photon Counting (TCSPC) connected to multiphoton laser scanning microscopes. TCSPC offers picosecond level time resolution, which is arranged in histograms based on timing information of the detection photon. In a typical time domain FLIM analysis workflow, the lifetime decay curve is subject to one- or two-component exponential curve fitting to estimate the lifetime parameters such as free: bound NAD(P)H lifetime, the amplitude of their decay curve, and goodness of fit (chi-squared error). If a machine learning (ML) algorithm has raw unfitted decay data, it could be possible to estimate lifetime accurately and subsequently characterize cell types. This approach would be advantageous in two ways: (i) it will reduce time for calculating lifetime from a decay curve (the only remaining bottleneck would then be collection time; the lifetime can be estimated instantly without the time-consuming exponential curve fitting); and (ii) possibly identify an optical/metabolic fingerprint which would otherwise not be apparent with regular analysis. This approach of calculating lifetime directly from the decay curve is relatively new, to our knowledge, there is one previous published report that has estimated lifetimes using artificial neural networks (Wu et al., 2016). Another approach of possible utility would be to use the estimated lifetime parameters calculated using standard curve fitting approaches and train where the label is created by antibody staining. Recently there is work by Smith et al. (2019) have utilized deep neural network with FLIM data to estimate lifetime in fit free fashion. Another group demonstrated machine

learning approach using random forest classifier on FLIM data for classification of tumor FLIM image (Unger et al., 2020).

In this study, we have used both approaches to train an Artificial Neural Network (ANN) to identify the location of microglia, i.e., (i) a fitting based method (FBM) where the fitted data are exported from standard curve fitting routines, and (ii) an experimental decay based method (DBM) training with the exponential decay curve consisting of 256 time-bins directly. To our knowledge, this is the first application of a FLIM-based machine learning application for intrinsic fluorescence signature development.

MATERIALS AND METHODS

Animals

All animals were maintained in an AALAC-accredited animal facility with a 12-h light/dark cycle regime and had access to food and water *ad libitum*. All experiments were performed in accordance with the University of Wisconsin-Madison Institutional Animal Care and Use Committee.

For FLIM imaging, 100 μm thick coronal slices were prepared from the fixed brains of young adult C57BL/6J and CX3CR1-GFP mice (Jackson Labs), aged 6–8 weeks. Animals were euthanized by isoflurane overdose and transcardially perfused with ∼30 ml of ice-cold PBS, followed by a second perfusion with an ice-cold solution of 4% PFA in PBS. Brains were then dissected, post-fixed for 24 h in a solution of 4% PFA in PBS, and then moved to HBSS (all performed at 4°C and protected from light).

Preparation of Primary Neonatal Mixed Glial Cultures

Mixed primary glial cultures were prepared from 3 to 7 day old, CX3CR1-GFP mouse pups as previously described (Crain et al., 2013). Briefly, brains were dissected immediately after decapitation, and the brain stem, olfactory bulbs, meninges, and visible blood vessels were removed. The remaining tissue was finely minced, thoroughly triturated with a serological pipette in 0.25% trypsin-EDTA containing 0.1 mg/ml deoxyribonuclease I, and then incubated at 37°C for 20 min. The reaction was immediately stopped by the addition of an equal volume of heat-inactivated horse serum. The dissociated cells were resuspended in DMEM supplemented with 10% FBS and 100 units/ml penicillin/streptomycin. Brains were processed individually for each pup, and the resulting cell suspension was divided equally and plated in 35 mm dishes (4–6 plates/brain). The plated cells were cultured for 7–14 days in a 37°C incubator supplemented with 5% CO_2; the culture medium was replaced every 3–4 days.

Immunohistochemistry

One hundred micrometers thick coronal sections were prepared from the midbrain region of each brain using a Leica Vibratome. Two slices from each animal were used for immunohistochemical staining. Briefly, slices were washed at room temperature with 0.3% TritonX-100 in PBS, before incubating in blocking buffer (1% BSA, 0.3% TritonX-100/PBS) for 2 h at room temperature. Slices were then incubated with anti-Iba1 antibodies (1:1000; Wako Catalog No. 019-19741) in blocking buffer in the dark at 4°C overnight. Slices were washed again at room temperature with 0.3% TritonX-100 in PBS followed by incubation in the dark for 2 h with AlexaFlour594 anti-rabbit IgG antibodies (1:200) in blocking buffer, at room temperature. Slices were washed with 0.3% TritonX-100 in PBS and mounted on 1 mm slides using Cytoseal60 mounting medium. Mounted sections were stored at room temperature, protected from light until they could be imaged.

Multiphoton Lifetime Imaging

The multiphoton based (Denk et al., 1990) lifetime and intensity imaging was performed on a custom multiphoton laser scanning system built around an inverted Nikon Eclipse TE2000U at the Laboratory for Optical and Computational Instrumentation (Yan et al., 2006). A 20× air immersion objective (Nikon Plan Apo VC, 0.75 NA) (Melville, NY, United States) was used for all imaging. For NAD(P)H imaging, data was collected using an excitation wavelength of 740 nm, and the emission was filtered at 457 ± 50 nm (Semrock, Rochester, NY) for the spectral peak for NAD(P)H/NADPH. For GFP intensity imaging, the excitation was set at 890 nm, and an emission 520 ± 35 filter was used (Semrock, Rochester, NY). For AlexaFluor594 imaging, excitation was set at 810 nm, and a 615/20 (Semrock, Rochester, NY) bandpass emission filter was used for emission. We used time domain FLIM imaging where the FLIM decays curves were built with TCSPC (Time Correlated Single Photon Counting) electronics. FLIM images of 256 × 256 pixels were collected with 120 s collection using SPC-150 Photon Counting Electronics (Becker & Hickl GmbH, Berlin, Germany) and Hamamatsu H7422P-40 GaAsP photomultiplier tube (Hamamatsu Photonics, Bridgewater, NJ). Urea crystals were used to determine the Instrumentation Response Function (IRF) with a 370/10 bandpass emission filter (Semrock, Rochester. NY). For each sample, around 20 neighboring FOVs were randomly selected, and the average value of lifetime and free NAD(P)H ratio was calculated based on masking described in the "Data Analysis" section. The instrument response function of the optical system was calibrated during each imaging session. Autofluorescence intensity and fluorescence lifetime data were analyzed in SPCImage (Becker & Hickl GmbH, Berlin, Germany) where a Levenberg–Marquardt routine for nonlinear fitting was used to fit the fluorescence decay curve collected for each decay after binning. Data were assessed by the minimized chi-square value generated during the fit so that the analyses were unbiased. To eliminate background fluorescence, a threshold for analysis was applied based on photon counts.

Data Analysis

The cell cultures used for experiments were CX3CR1-GFP positive, and the GFP intensity image was used to create the mask for identifying microglia. For the brain tissue imaging, anti-Iba1 antibodies visualized with AlexaFluor594 was used to identify microglia (**Figure 1A**) and the mask is shown in **Figure 1B**. The lifetime fitted data from SPCImage (Becker & Hickl GmbH, Berlin, Germany) and all the parameters were exported for more

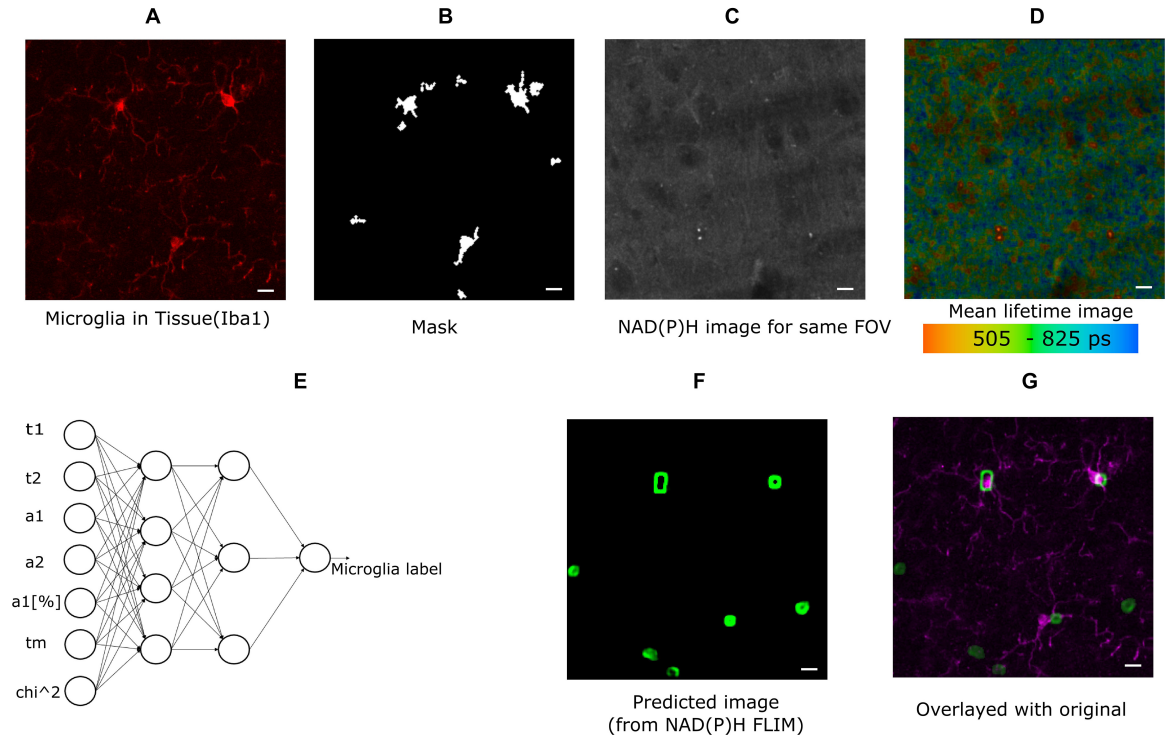


FIGURE 1 | Method demonstrating microglia prediction using ANN. **(A)** Microglia image using Iba1. **(B)** Mask created from Iba1 image **(C)** NAD(P)H intensity image from the same FOV. **(D)** Lifetime image of the same FOV created from SPCImage (Becker and Hickl GmbH, Berlin, Germany) **(E)** ANN showing the inputs used for training instances and output **(F)** predicted microglia image. **(G)** Composite image showing fusion of predicted microglia image and actual microglia image created from Iba1 (scale bar 10 μm, **A–D,E,F**).

custom operations. **Figure 1C** shows the NAD(P)H intensity image created from the lifetime image. **Figure 1D** shows the lifetime image (mean lifetime) created by curve fitting. Mean lifetime is one of the parameters exported for ANN training. The exported data were imported in MATLAB (MathWorks Inc., Natick, MA) for calculating the means of custom regions and statistical analyses. The masks were created from intensity images in MATLAB. The time-resolved Becker and Hickl SDT data were read using Bio-Formats (Linkert et al., 2010) MATLAB support package. For generating the threshold of the predicted image, we tuned a optimal threshold that maximizes overlap on the training images. Then the trained network is applied to the test images and the threshold is applied. A cell is considered a positive detection when it overlaps with the cell body or the processes, otherwise it is considered a false positive.

ANN Implementation, Fitting-Based Method (FBM)

In this approach, the fitted data for NAD(P)H lifetime was exported in *.asc format from SPCImage and read in MATLAB for post-processing and training. The exported data were mean free/bound/mean NAD(P)H lifetimes (τ_1, τ_2, τ_m), amplitude of their decay curve and free bound ratio (a_1, a_2, a_1[%]), goodness of fit (chi-squared error, $\chi 2$). The neural network for training and testing is applied on the image data after some preprocessing to create smaller block. All exported images were exported from 256 × 256 size lifetime images, and smaller overlapping blocks of 8 × 8 size were taken from each dataset for training instances; the average of each block was calculated for each parameter. The neural network is applied on the each pixel of the new calculated point from the 8 × 8 block. In this way, 8,494,137 labeled training datasets were created for cells, and 5,456,088 training datasets were created for tissue training. Subsequently, the data was used for training using MATLAB's neural fitting toolbox. Seventy percentage of the dataset was used for training, and Fifteen percentage was set aside for validation and testing. Bayesian regularization (Foresee and Hagan, 1997) backpropagation was used as a network training function, that updates the values according to Levenberg-Marquardt optimization. We used a feed-forward network with one hidden layer with 10 neuron in the hidden layer. As a cost function, we used mean squared error. The performance index is the mean square error, generated by the neural network toolbox comparing the result of test and training dataset (random splitting between training/testing/validation) with a number ranging from 0 to 1, 0 indicating maximum match 1 indicating no match. The performance with 10 hidden neuron and Bayesian Regularization was 0.0202 for mixed glial cell culture and 0.013 for tissue; increasing the number of hidden neurons further did not improve performance. For example, with 20 hidden neuron with cell the performance was 0.020 but with five hidden neuron the performance was 0.024. With additional hidden layers, the performance remained the same. We also compared training performance using Naive Bayes classifier,

support vector machine (SVM) and K-nearest neighbor (Knn). We reached performance of 0.92 with Naïve bayes and 0.802 with Knn. With SVM the performance varied from 0.07 to 0.15. Given the performance achieved by ANN, we decided to used it for further classification. The accuracy for microglia detection was determined by first exporting the FLIM parameters from NAD(P)H lifetime from a separate testing dataset. The input data for feeding the ANN (**Figure 1E**) was created the same way as the training process. Following this step, the label for each block was predicted using the trained model, and the label for that block was created. Repeating this step for all pixels generated a 2D image with microglia detection probabilities. Pixels with lower probability was discarded and smaller detected regions were also discarded. The calibration was done on training images to ensure overlap. Then, these thresholds were used to create a probability image. **Figure 1F** shows such a mask which has seven microglia in the FOV. The predicted image was compared with antibody-labeled (**Figure 1A**) images to locate microglia positions. The composite image (**Figure 1G**) shows the overlap between the predicted microglia region and antibody detected microglia region. We calculated the sensitivity (True Positive Rate), Positive Predictive Rate (PPV), False Negative Rate (FNR), and False Discovery Rate (FDR) (Sammut and Webb, 2017) from the detected microglia after prediction and post-processing. We did not calculate the rest of the parameters of the confusion matrix as we do not have the true negatives with our current approach of imaging. We do not have labeled data for non-microglia cells, as it was out of the scope of this paper.

ANN Implementation, Decay-Based Method (DBM)

This approach exports the decay directly, which has 256 times bins in the histogram. The data were directly read from time-resolved SDT file using the Bio-Formats file reader library (Linkert et al., 2010) in MATLAB. Instead of using the fitted parameters as input to the ANN, we used an ANN with 10 hidden neuron and 256 input nodes. The rest of the training and testing was the same as the previous method. The performance with these settings for cells was 0.0227.

RESULTS

ANN-FLIM Can Detect Microglia in Mixed Glial Cell Cultures

In this section, we demonstrate the ability of ANN-based techniques to identify microglia in mixed glial cell cultures. ANN is applied on exported data after curve fitting from SPCImage software. NAD(P)H lifetime fitted data are exported from SPCImage (see section "Materials and Methods"), and exported parameters are used to compute training instances for individual blocks. The training is performed on training sets and tested on a separate testing dataset. **Figure 2A** shows the GFP intensity image created from GFP positive microglia. The predicted microglia image from a sample field of view is shown in **Figure 2B**. **Figure 2C** shows the fused image of the original microglia image and predicted image from the same FOV. It is evident from the fused image that, most of the microglia are properly detected when compared with actual microglia image created from GFP. There are some microglia which are not predicted and some false positive in the lower left corner where microglia are falsely identified. **Figure 2D** shows the error obtained by prediction of individual instances created from FLIM parameters of the testing dataset. The total number of microglia in all (testing) FOVs were 348 and of them 313 were correctly identified, 35 microglia was missed by the prediction algorithm, but 138 additional microglia was falsely identified. **Figure 2E** shows the result for microglia detection for five different dishes; we got TPR 0.90 ± 0.03, PPV 0.67 ± 0.10, FNR 0.09 ± 0.03, FDR 0.33 ± 0.10.

ANN-FLIM Can Detect Microglia in Brain Tissue

Next, we extended our approach to identifying microglia in fixed mouse brain tissue slices. Imaging brain tissue is more challenging than imaging cell cultures because of greater heterogeneous spatial structures, and a wider degree of variation in microglia lifetimes. The algorithm implementation was the same as in the *in vitro* cell culture experiments, where ANN was applied to the exported lifetime data from curve fitting software, SPCImage. An anti-Iba1 antibody with an AlexaFluor594 secondary antibody was used to visualize and create a microglial intensity image (**Figure 3A**). **Figure 3B** shows the predicted microglia image created from NAD(P)H lifetimes of the same field of view as in **Figure 3A**, in which the microglia are stained with Iba1. **Figure 3C** shows the fused image of the original microglia image (from tissue)- and predicted image from the same FOV. It is evident from the fused image that all of the microglia from this FOV is properly detected when compared with the Iba1 intensity positive microglia. **Figure 3D** shows the error obtained by while predicting using individual instances created from FLIM parameters of the testing dataset. The total number of microglia in the testing FOVs are 170 and 137 were correctly identified, but 76 microglia were falsely identified. **Figure 3E** shows the result for five different tissues where we got TPR 0.79 ± 0.08, PPV 0.638 ± 0.09, FNR 0.2 ± 0.08, FDR 0.36 ± 0.09. The TPR is reduced for microglia in tissue and FNR is increased as the heterogeneity and complexity of the structure complicates the prediction.

ANN Directly on Exponential Decay Can Detect Microglia in Mixed Cell Cultures

Finally, we implemented an experimental approach, where instead of exporting the lifetime fitted data, we used the decay data having 256-time bins as a training instance. A recent study showed the effectiveness of using ANN to calculate lifetime directly from decay (Wu et al., 2016). Instead of calculating lifetime, we directly used microglia locations as labels. This approach is simpler because it bypasses the steps involving exponential curve fitting routines and exporting the fitted lifetime parameters. **Figure 4A** shows the intensity image of microglia created from the GFP channel in the mixed glial cultures. **Figure 4B** shows the recreated image from ANN-predicted microglia from the exponential decay. **Figure 4C** shows the

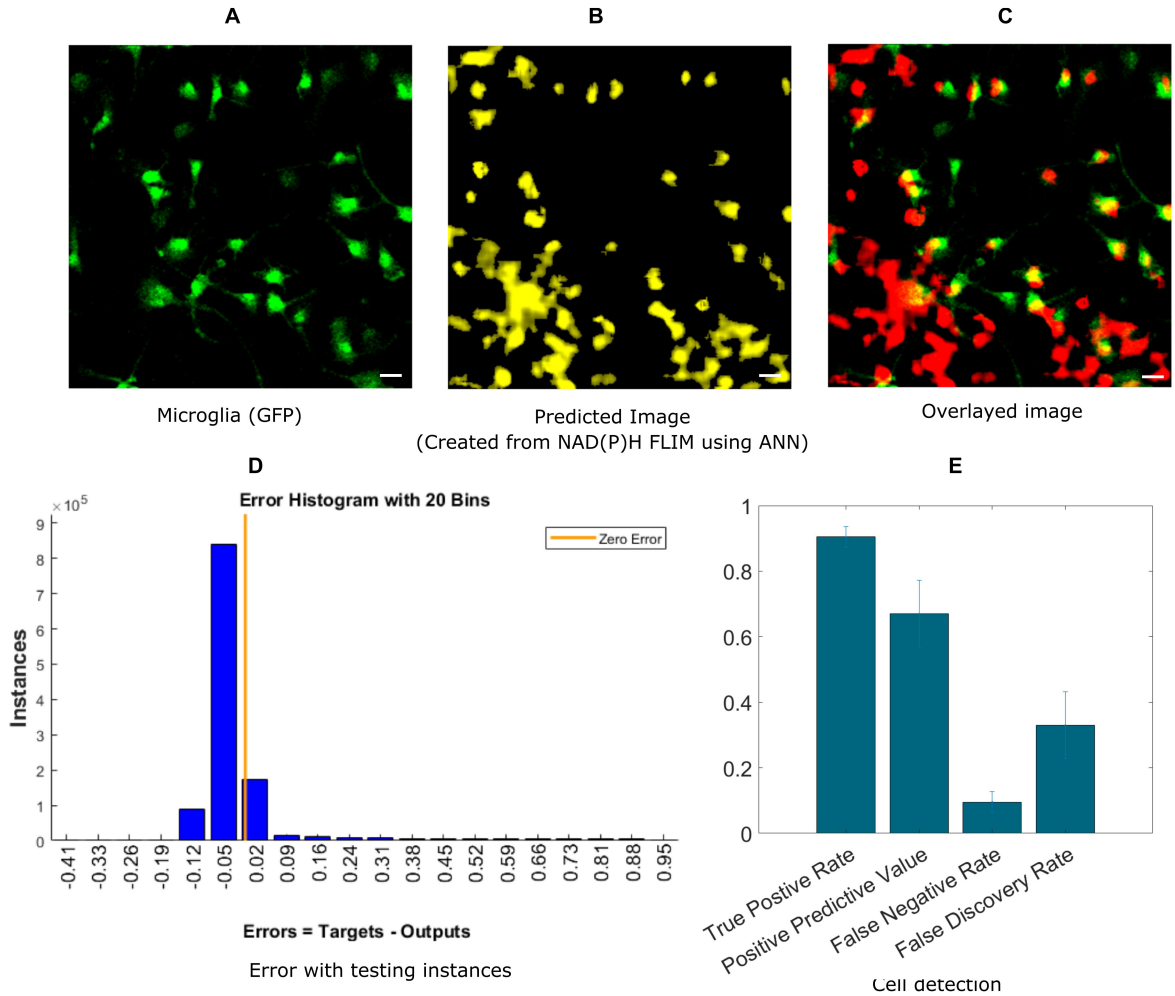


FIGURE 2 | Prediction of microglia from NAD(P)H lifetime data in mixed glial cell cultures where microglia cells are GFP-labeled. In this approach, ANN is applied on exported lifetime parameters of the endogenous fluorophore NAD(P)H. **(A)** Original fluorescence intensity image of microglia created from the GFP channel **(B)** Predicted microglia image from NAD(P)H lifetime data from the same field-of-view. **(C)** Composite image of predicted microglia image and original intensity image. Most of the microglia are accurately predicted as seen from the composite image. But there are a few microglia not detected and some false positive in the lower left corner **(D)** Error rate from testing instances, created from the testing microglia dataset **(E)** TPR, PPV, FNV and FDR from five different dishes (Scale bar 20 μm, **A–C,E**).

fused image of the original microglia image (from mixed glial culture)- and predicted image from the same FOV. **Figure 4D** shows the error obtained by while predicting using individual instances created from FLIM parameters of the testing dataset. The total number of microglia in the testing FOVs are 371 and 136 were correctly identified, but 24 microglia were falsely identified. **Figure 4E** shows the result for five different dishes where we got TPR 0.36 ± 0.09, PPV 0.82 ± 0.16, FNR 0.63 ± 0.08, FDR 0.17 ± 0.15.

DISCUSSION

In this paper, we have used ML and fluorescence lifetime to identify microglia in both mixed cell cultures and in brain tissue sections. To our knowledge, this is one of the first studies to apply ML algorithm to FLIM data to identify intrinsic cellular metabolic signatures. Moreover, applying ML methods directly on exponential decay data can be potentially augmented to calculate lifetime without curve fitting to help understand underlying trends in biological samples. The techniques applied in this paper can be extended in various ways in studies related to brain metabolism. Although there have been some FLIM studies to visualize brain metabolism, but ML-FLIM was not used to characterize cell types previously. This technique can be potentially extended to identify other CNS glial cells such as astrocytes, oligodendrocytes, and neurons. For *in vivo* brain studies, a ML-based, label-free technique can be used in the future where FLIM for different spectral channels can be used in conjunction with trained ML networks to identify and study specific cell types. Here, we used the lifetime information generated by exponential curve fitting from each

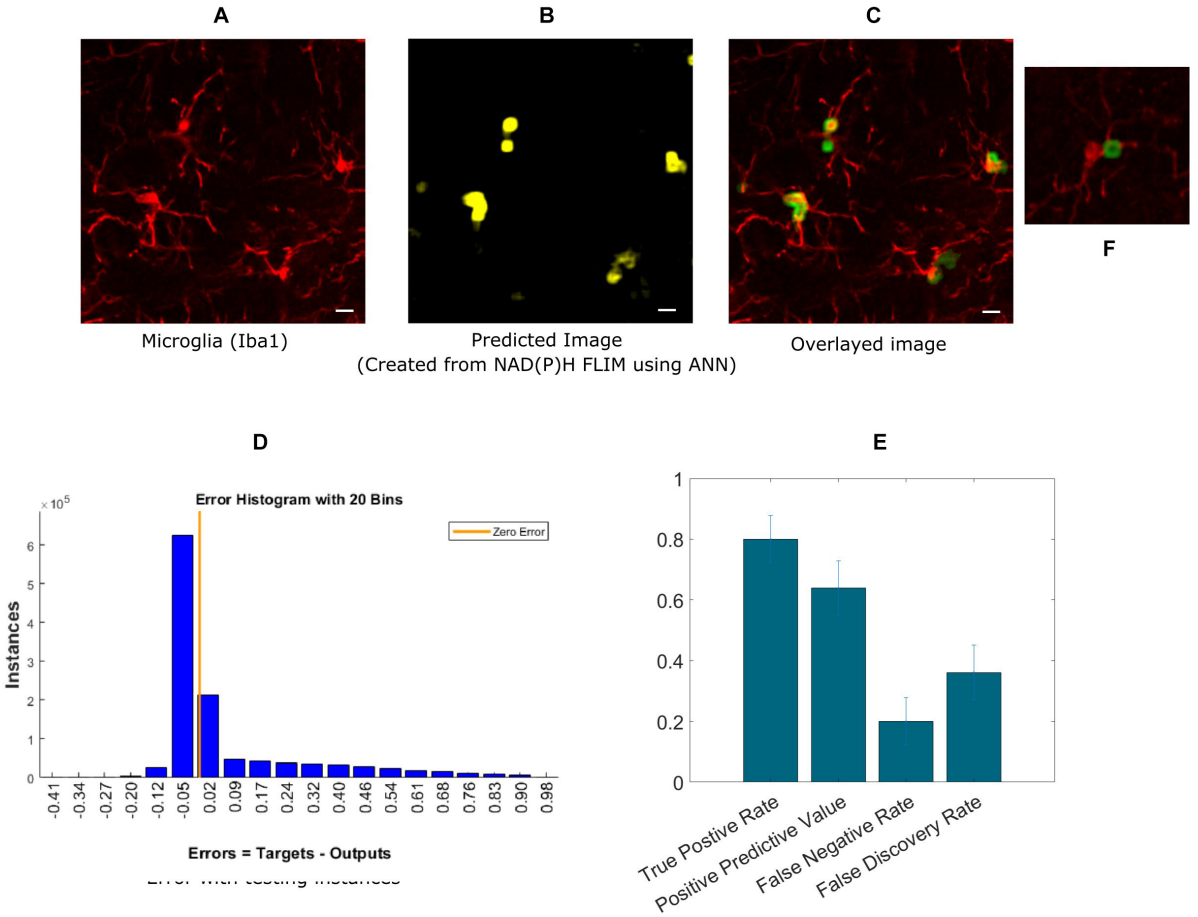


FIGURE 3 | Prediction of microglia from NAD(P)H lifetime data in mouse brain tissue. In this approach, ANN is applied on exported lifetime parameters of the endogenous fluorophore NAD(P)H. **(A)** Original intensity image of microglia location created from the anti-Iba1 antibody AlexaFluor594 channel **(B)** Predicted microglia image from NAD(P)H lifetime data from the same field-of-view. Composite image of predicted microglia image and original intensity image. **(C)** Composite image of predicted microglia image and original intensity image. **(D)** Error rate from training and **(E)** TPR, PPV, FNV, and FDR from five different tissues (Scale bar 10 μm, **A–C,E**). **(F)** Zoomed in image of a microglia where a process is identified as positive.

pixel as training instances, and antibody-stained microglia locations are used as positive microglia pixels. The goal was to be able to identify microglia using the NAD(P)H lifetime information alone. We have used two different approaches for training, (1) the calculated lifetime parameters created from curve fitting software SPCImage, FBM, and (2) ML directly on time-resolved data, DBM. The second approach bypasses the time-consuming curve fitting process and reduces the overall number of processing steps. We first implemented our ANN-based approach on microglia in mixed glial cell culture and then extended this to testing tissue samples. In the future, the tissue-based identification can be repurposed for *in vivo* identification of microglia given there was enough training data before detection testing.

Microglia are normally visualized with antibody-based methods, using antibodies, such as anti-TMEM119 (Bennett et al., 2016), anti-Iba1, and the combination of anti-CD11b and anti-CD45, among others. These standard labeling techniques have several limitations, including the extended sample preparation time and associated complexity, inefficient antibody penetration, and the potential for non-specific antibody binding. A FLIM based, label-free imaging technique is advantageous because it is simple, free from exogenous labeling, provides the ability to directly measure intrinsic cellular properties, and its potential for extending its use to observe cell activity *in vivo*. NAD(P)H-FLIM based endogenous biomarker visualization is an effective way to image intrinsic metabolism as NAD(P)H lifetimes and free: bound ratios change with alterations in metabolic state. The metabolic state of immune cells like microglia can be different from other non-immune cells and the surrounding tissue, and these differences are reflected by the alternations in their lifetime signature.

For fitting-based methods (FBM), where the fitted parameters were exported from SPCImage, we were able to achieve TPR of 0.90 ± 0.03 for five different microglia mixed culture group (**Figure 2**). However, we got some false positives with FNR being 0.09 ± 0.03 and FDR being 0.33 ± 0.10. For clinical application the false positive need to be reduced in future studies. We expect

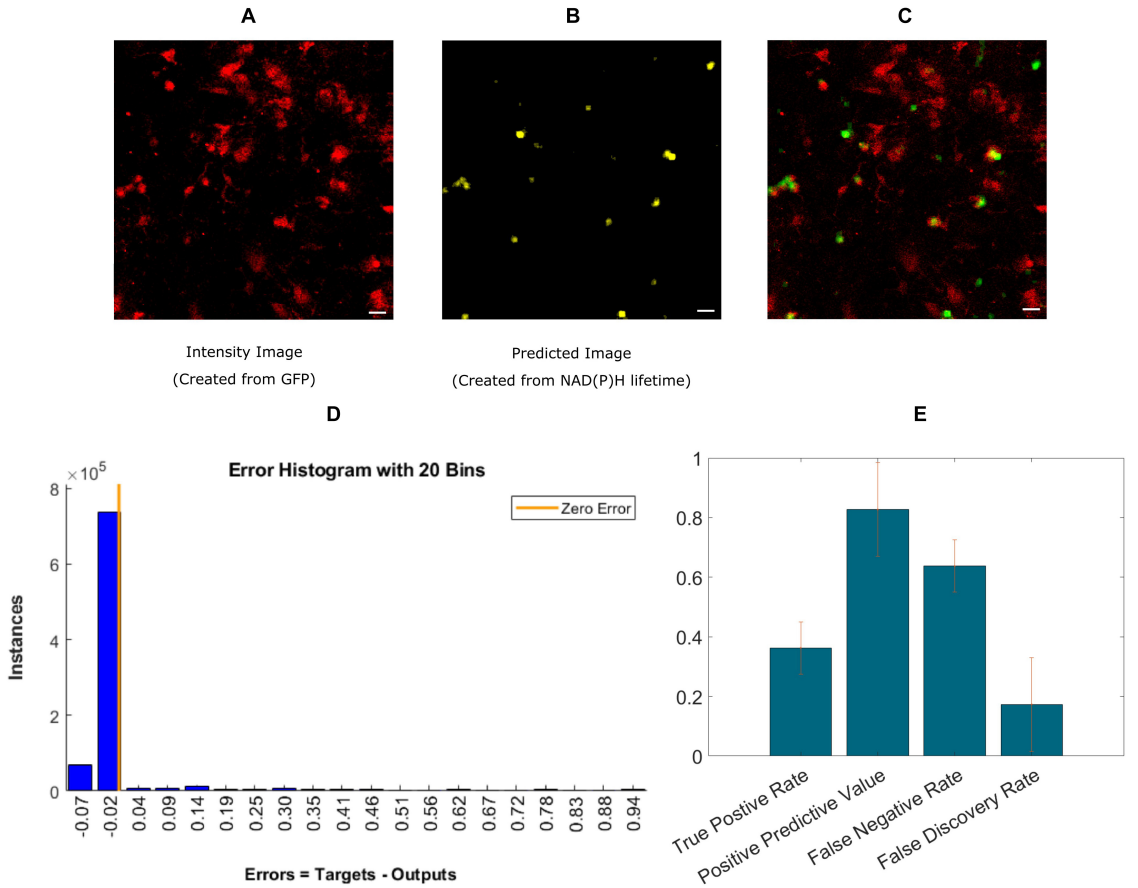


FIGURE 4 | Prediction of microglia from NAD(P)H lifetime data in mixed glial cell cultures where microglia cells are GFP-labeled. In this approach, ANN is applied directly on exponential decay curves of the endogenous fluorophore NAD(P)H. **(A)** Original intensity image of microglia location created from the GFP channel **(B)** Predicted microglia image from NAD(P)H lifetime data only. Both images are from the same field-of-view. **(C)** Composite image of predicted microglia image and original intensity image. **(D)** Error rate from training and **(E)** TPR, PPV, FNV, and FDR from five different dishes (Scale bar 20 μm, **A–C**).

this number will be improved by increasing the number of training samples with more diverse samples matching the realistic use case for *in vivo* imaging. This approach would require a significant amount of data acquisition and processing time as each FLIM image takes 1–2 min to acquire. This FBM approach was also applied to brain tissue sections, which achieved a lower accuracy than cell culture which we expected. The tissue is much more heterogeneous and different fluorescent signature add to the variation in lifetime alternations. Still, we managed to achieve TPR of 0.79 ± 0.08 and PPV of 0.64 ± 0.09 but increased false positive resulted in FNR of 0.2 ± 0.08 and FDR of 0.36 ± 0.09 (**Figure 3**). The reduction in TPR for tissue can be attributed to several factors. In the mixed glial cell, there were primarily glial cell of which majority of non-microglia cells were astrocyte. The lifetime variation is much smaller in this environment compared to an actual brain tissue where lots of different factors contribute to the NAD(P)H lifetime variation. We also had fewer number of microglia cell for training for the same number of FOVs, which can also contribute to the reduction in accuracy to some extent. Combination of these factors can contribute to the reduction in TPR and increase in FDR.

The second approach, DBM used the decay curve directly as the input training parameter. However, there is shortfall in the accuracy possibly introduced by curve shift between successive imaging, after pulsing of the detectors that requires taking into account Instrument Response Function (IRF). When the testing instances (from mixed glial culture) were reorganized to form image after classification, we obtained a TPR of 0.36 ± 0.09 but a PPV of 0.82 ± 0.15. But we also got an increased FNR of 0.63 ± 0.08 but reduced FDR 0.17 ± 0.15. This method has the lowest TPR but also achieved the best FDR. With more training samples this method could lead to significantly better performance. This method did not yield an acceptable result in tissue as the heterogeneity in the tissue might add to the variation. To make this method more accurate and clinically viable, we need more training datasets as well as take into account the FLIM instrumentation originated variation such as shift in decay and IRF. Moreover, surface-based markers such as CD11b for microglia could result in a better classification compared to Iba1 as they are able to properly identify all of cell body and processes which would reduce false negatives for training. We intend to explore this further in future work, as well as combining

advanced deep learning tools such as Convolutional Neural Networks (Krizhevsky et al., 2012) to better predict microglia location using lifetime information and morphological features.

One limitation of our approaches is inherent to the TCSPC approach itself, as TCSPC acquisition takes several minutes to finish a single frame depending on the sample fluorescence intensity. For *in vivo* live acquisitions, the TCSPC approach might be less useful if real-time visualization is required. Frequency domain FLIM (Gratton et al., 2003) acquisition followed by ML could be another approach to overcome acquisition time limitation. Another limitation is the variability of relative shift in the exponential decay, although this can be overcome with the fitting-based method. The FBM already deals with shifts during the fitting process by the fitting software. It could be an issue with the decay-based method where the shift is not taken into account. It might be one of the reasons why the decay-based method did not yield good results with tissue sections. We plan to address the issue with the decay-based approach in tissue sections by using more training data in the future, incorporating the IRF and adding more classification features that consider morphology and/or the intensity of different spectral channels. Other potential exponential curve fitting issues include fitting bias and local minima. One way of overcoming this issue would be avoiding curve fitting by either using raw data (which we demonstrated) and using phasor analysis methods with the TCSPC FLIM data. We ran some preliminary experiment to test the accuracy of the phasor-based approach (TPR 0.41 $\pm$ 0.03, PPV 0.49 $\pm$ 0.11, FDR 0.51 $\pm$ 0.12) and found that false positives were relatively high. But in the future, a hybrid approach where phasors are part of classification features alongside morphology could provide better discrimination for cell type. Another issue we would like to bring to attention is the tissue/cell fixation. All of our testing and training was performed using fixed cell/tissue. While there is some past concern with FLIM imaging with fixed samples for NAD(P)H imaging, our lab has demonstrated recently that, although fixation causes shift in NAD(P)H lifetime values, the metabolic signature and trend are not altered (Chacko and Eliceiri, 2019). For classification for live tissue, new training sample would be required to tune the parameters. Another limitation of our experiment was that we did not identify True Negative (TN) and as a result all the components of confusion matrix and accuracy can't be determined. For the mixed glial culture, the TNs would be the non-microglia glial cell. But it is more complicated in brain tissue to define TN as the brain tissue is heterogeneous and consist of components that are challenging to define in terms of TN. One last limitation would be the variability of NAD(P)H lifetimes from sample to sample. Based on the microenvironment, the NAD(P)H lifetime can vary even though the cells/tissues are treated similarly. One way to overcome this limitation would be to train with larger datasets with similar treatments.

CONCLUSION

We have demonstrated a novel machine learning based approach that can use FLIM data to identify microglia based on NAD(P)H lifetime parameters. We have successfully shown the effectiveness of the method in both cells and tissue slices and achieved close to 90% True Positive Rate and moderately low False Discovery Rate. Additionally, we have shown that the decay can be used to directly identify microglia using ANN without exponential curve fitting. This approach can be further enhanced to calculate lifetimes and other parameters from lifetime decay data directly using machine learning.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The animal study was reviewed and approved by the University of Wisconsin-Madison Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

MS and KE conceptualized the technique and application. MS, KC, JO, JCW, JJW, and KE designed the experiments. KC and JO performed the biology experiments. MS wrote the code for ANN and performed the data analysis. MS, JJW, and KE wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge funding from the Semiconductor Research Corporation (KE) and the Morgridge Institute for Research. This work was also supported by NIH R01 085226 (JJW). Non-animal work was sponsored by the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Tristan McClure-Begley through the Space and Naval Warfare Systems Center, Pacific Grant/Contract No. N66001-17-2-4010.

REFERENCES

Alfonso-García, A., Smith, T. D., Datta, R., Luu, T. U., Gratton, E., Potma, E. O., et al. (2016). Label-free identification of macrophage phenotype by fluorescence lifetime imaging microscopy. *JBO* 21:046005. doi: 10.1117/1.JBO.21.4.046005

Bennett, M. L., Bennett, F. C., Liddelow, S. A., Ajami, B., Zamanian, J. L., Fernhoff, N. B., et al. (2016). New tools for studying microglia in the mouse and human CNS. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1738–E1746. doi: 10.1073/pnas.1525528113

Berezin, M. Y., and Achilefu, S. (2010). Fluorescence Lifetime Measurements and Biological Imaging. *Chem. Rev.* 110, 2641–2684. doi: 10.1021/cr900343z

Bhattacharjee, A., Datta, R., Gratton, E., and Hochbaum, A. I. (2017). Metabolic fingerprinting of bacteria by fluorescence lifetime imaging microscopy. *Sci. Rep.* 7:3743. doi: 10.1038/s41598-017-04032-w

Bird, D. K., Yan, L., Vrotsos, K. M., Eliceiri, K. W., Vaughan, E. M., Keely, P. J., et al. (2005). Metabolic mapping of MCF10A human breast cells via multiphoton fluorescence lifetime imaging of the coenzyme NADH. *Cancer Res.* 65, 8766–8773. doi: 10.1158/0008-5472.CAN-04-3922

Chacko, J. V., and Eliceiri, K. W. (2019). NAD(P)H fluorescence lifetime measurements in fixed biological tissues. *Methods Appl. Fluoresc.* 7:044005. doi: 10.1088/2050-6120/ab47e5

Charles, N. A., Holland, E. C., Gilbertson, R., Glass, R., and Kettenmann, H. (2011). The brain tumor microenvironment. *Glia* 59, 1169–1180. doi: 10.1002/glia.21136

Chia, T. H., Williamson, A., Spencer, D. D., and Levene, M. J. (2008). Multiphoton fluorescence lifetime imaging of intrinsic fluorescence in human and rat brain tissue reveals spatially distinct NADH binding. *Opt. Express* 16, 4237–4249. doi: 10.1364/OE.16.004237

Crain, J. M., Nikodemova, M., and Watters, J. J. (2013). Microglia express distinct M1 and M2 phenotypic markers in the postnatal and adult central nervous system in male and female mice. *J. Neurosci. Res.* 91, 1143–1151. doi: 10.1002/jnr.23242

Das, B., Shi, L., Budansky, Y., Rodriguez-Contreras, A., and Alfano, R. (2018). Alzheimer mouse brain tissue measured by time resolved fluorescence spectroscopy using single- and multi-photon excitation of label free native molecules. *J. Biophotonics* 11:e201600318. doi: 10.1002/jbio.201600318

Denk, W., Strickler, J. H., and Webb, W. W. (1990). Two-photon laser scanning fluorescence microscopy. *Science* 248, 73–76. doi: 10.1126/science.2321027

Flexman, M. L., Kim, H. K., Gunther, J. E., Lim, E. A., Alvarez, M. C., Desperito, E., et al. (2013). Optical biomarkers for breast cancer derived from dynamic diffuse optical tomography. *J. Biomed. Opt.* 18:096012. doi: 10.1117/1.JBO.18.9.096012

Foresee, F. D., and Hagan, M. T. (1997). "Gauss-Newton approximation to Bayesian learning," in *Proceedings of International Conference on Neural Networks (ICNN'97)*, Houston, TX, 1930–1935.

Freudiger, C. W., Min, W., Saar, B. G., Lu, S., Holtom, G. R., He, C., et al. (2008). Label-free biomedical imaging with high sensitivity by stimulated raman scattering microscopy. *Science* 322, 1857–1861. doi: 10.1126/science.1165758

Garden, G. A., and Möller, T. (2006). Microglia biology in health and disease. *J. NeuroImmune Pharm.* 1, 127–137. doi: 10.1007/s11481-006-9015-5

Gratton, E., Breusegem, S., Sutin, J., Ruan, Q., and Barry, N. (2003). Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods. *J. Biomed. Opt.* 8, 381–390. doi: 10.1117/1.1586704

Kantelhardt, S. R., Kalasauskas, D., König, K., Kim, E., Weinigel, M., Uchugonova, A., et al. (2016). In vivo multiphoton tomography and fluorescence lifetime imaging of human brain tumor tissue. *J. Neurooncol.* 127, 473–482. doi: 10.1007/s11060-016-2062-8

Kirkpatrick, N. D., Hoying, J. B., Botting, S. K., Weiss, J. A., and Utzinger, U. (2006). In vitro model for endogenous optical signatures of collagen. *JBO* 11:054021. doi: 10.1117/1.2360516

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 1097–1105.

Lakowicz, J. R., Szmacinski, H., Nowaczyk, K., and Johnson, M. L. (1992). Fluorescence lifetime imaging of free and protein-bound NADH. *PNAS* 89, 1271–1275. doi: 10.1073/pnas.89.4.1271

Leppert, J., Krajewski, J., Kantelhardt, S. R., Schlaffer, S., Petkus, N., Reusche, E., et al. (2006). Multiphoton excitation of autofluorescence for microscopy of glioma tissue. *Neurosurgery* 58, 759–767. doi: 10.1227/01.NEU.0000204885.45644.22

Linkert, M., Rueden, C. T., Allan, C., Burel, J.-M., Moore, W., Patterson, A., et al. (2010). Metadata matters: access to image data in the real world. *J. Cell. Biol.* 189, 777–782. doi: 10.1083/jcb.201004104

Mongeon, R., Venkatachalam, V., and Yellen, G. (2016). Cytosolic NADH-NAD+ redox visualized in brain slices by two-photon fluorescence lifetime biosensor imaging. *Antioxid. Redox Signal.* 25, 553–563. doi: 10.1089/ars.2015.6593

Provenzano, P. P., Eliceiri, K. W., and Keely, P. J. (2008). Multiphoton microscopy and fluorescence lifetime imaging microscopy (FLIM) to monitor metastasis and the tumor microenvironment. *Clin. Exp. Metastasis* 26, 357–370. doi: 10.1007/s10585-008-9204-0

Rinnenthal, J. L., Börnchen, C., Radbruch, H., Andresen, V., Mossakowski, A., Siffrin, V., et al. (2013). Parallelized TCSPC for dynamic intravital fluorescence lifetime imaging: quantifying neuronal dysfunction in neuroinflammation. *PLoS One* 8:e60100. doi: 10.1371/journal.pone.0060100

Sammut, C., and Webb, G. I. (eds) (2017). *Encyclopedia of Machine Learning and Data Mining*, 2nd Edn. Cham: Springer.

Skala, M. C., Riching, K. M., Bird, D. K., Gendron-Fitzpatrick, A., Eickhoff, J., Eliceiri, K. W., et al. (2007). In vivo multiphoton fluorescence lifetime imaging of protein-bound and free nicotinamide adenine dinucleotide in normal and precancerous epithelia. *J. Biomed. Opt.* 12:024014. doi: 10.1117/1.2717503

Skala, M. C., Squirrell, J. M., Vrotsos, K. M., Eickhoff, J. C., Gendron-Fitzpatrick, A., Eliceiri, K. W., et al. (2005). Multiphoton microscopy of endogenous fluorescence differentiates normal, precancerous, and cancerous squamous epithelial tissues. *Cancer Res.* 65, 1180–1186. doi: 10.1158/0008-5472.CAN-04-3031

Smith, J. T., Yao, R., Sinsuebphon, N., Rudkouskaya, A., Un, N., Mazurkiewicz, J., et al. (2019). Fast fit-free analysis of fluorescence lifetime imaging via deep learning. *PNAS* 116, 24019–24030. doi: 10.1073/pnas.1912707116

Stringari, C., Nourse, J. L., Flanagan, L. A., and Gratton, E. (2012). Phasor fluorescence lifetime microscopy of free and protein-bound nadh reveals neural stem cell differentiation potential. *PLoS One* 7:e48014. doi: 10.1371/journal.pone.0048014

Stuntz, E., Gong, Y., Sood, D., Liaudanskaya, V., Pouli, D., Quinn, K. P., et al. (2017). Endogenous two-photon excited fluorescence imaging characterizes neuron and astrocyte metabolic responses to manganese toxicity. *Sci. Rep.* 7:1041. doi: 10.1038/s41598-017-01015-9

Suhling, K., French, P. M. W., and Phillips, D. (2005). Time-resolved fluorescence microscopy. *Photochem. Photobiol. Sci.* 4, 13–22. doi: 10.1039/B412924P

Suhling, K., Hirvonen, L. M., Levitt, J. A., Chung, P.-H., Tregidgo, C., Le Marois, A., et al. (2015). Fluorescence lifetime imaging (FLIM): basic concepts and some recent developments. *Med. Photonics* 27, 3–40. doi: 10.1016/j.medpho.2014.12.001

Sun, Y. H., Hatami, N., Yee, M., Phipps, J. E., Elson, D. S., Gorin, F., et al. (2010). Fluorescence lifetime imaging microscopy for brain tumor image-guided surgery. *JBO* 15:056022. doi: 10.1117/1.3486612

Szulczewski, J. M., Inman, D. R., Entenberg, D., Ponik, S. M., Aguirre-Ghiso, J., Castracane, J., et al. (2016). In vivo visualization of stromal macrophages via label-free FLIM-based metabolite imaging. *Sci. Rep.* 6:sre25086. doi: 10.1038/srep25086

Tambuyzer, B. R., Ponsaerts, P., and Nouwen, E. J. (2009). Microglia: gatekeepers of central nervous system immunology. *J. Leukocyte Biol.* 85, 352–370. doi: 10.1189/jlb.0608385

Unger, J., Hebisch, C., Phipps, J. E., Lagarto, J. L., Kim, H., Darrow, M. A., et al. (2020). Real-time diagnosis and visualization of tumor margins in excised breast specimens using fluorescence lifetime imaging and machine learning. *Biomed. Opt. Express* 11, 1216–1230. doi: 10.1364/BOE.381358

Wang, X. F., Periasamy, A., Herman, B., and Coleman, D. M. (1992). Fluorescence lifetime imaging microscopy (FLIM): instrumentation and applications. *Crit. Rev. Anal. Chem.* 23, 369–395. doi: 10.1080/10408349208051651

Watters, J. J., Schartner, J. M., and Badie, B. (2005). Microglia function in brain tumors. *J. Neurosci. Res.* 81, 447–455. doi: 10.1002/jnr.20485

Wu, G., Nowotny, T., Zhang, Y., Yu, H.-Q., and Li, D. D.-U. (2016). Artificial neural network approaches for fluorescence lifetime imaging techniques. *Opt. Lett.* 41, 2561–2564. doi: 10.1364/OL.41.002561

Yan, L., Rueden, C. T., White, J. G., and Eliceiri, K. W. (2006). Applications of combined spectral lifetime microscopy for biology. *Biotechniques* 41, 249–251. doi: 10.2144/000112251

Yaseen, M. A., Sutin, J., Wu, W., Fu, B., Uhlirova, H., Devor, A., et al. (2017). Fluorescence lifetime microscopy of NADH distinguishes alterations in cerebral metabolism in vivo. *Biomed. Opt. Express* 8, 2368–2385. doi: 10.1364/BOE.8.002368

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Sagar, Cheng, Ouellette, Williams, Watters and Eliceiri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

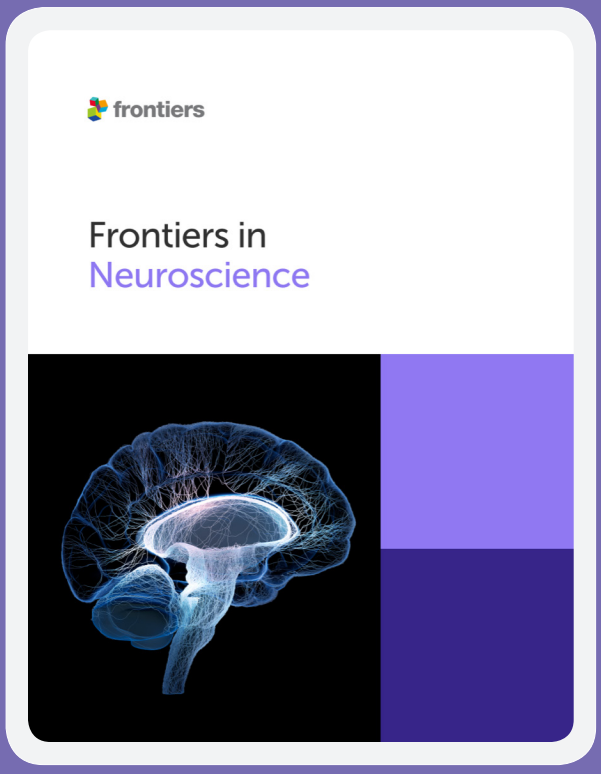
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact