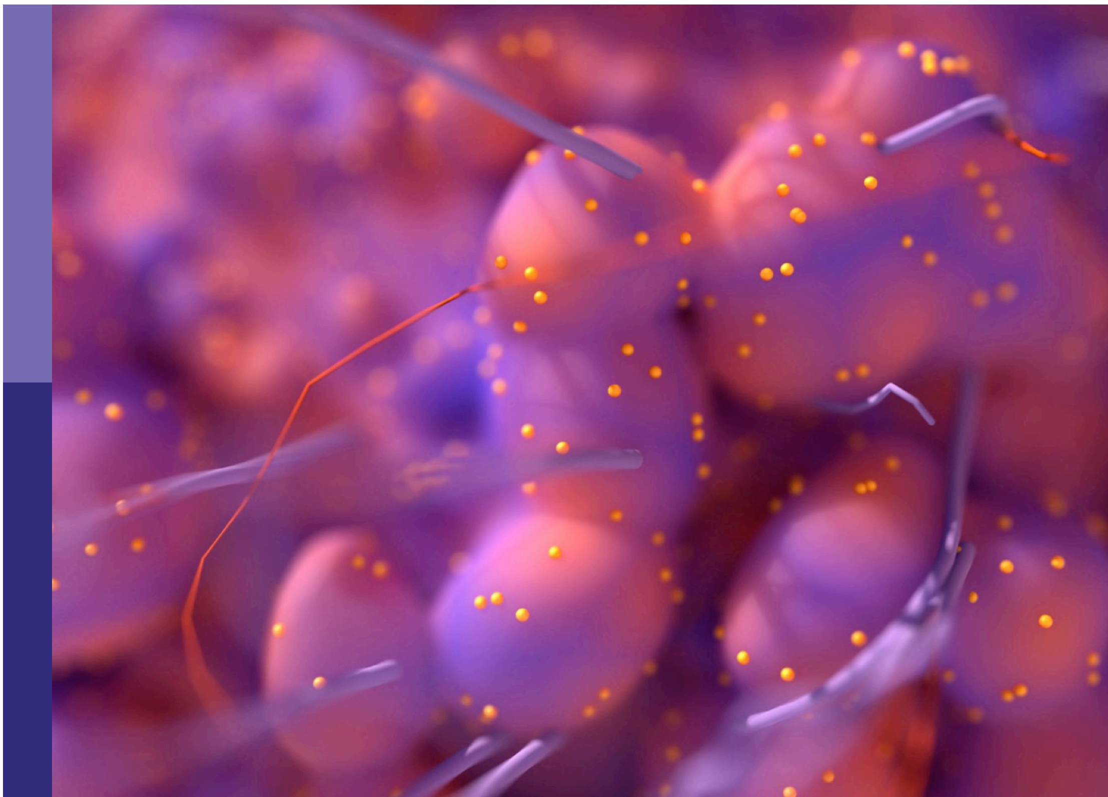
frontiers | Research Topics

The use of deep learning in mapping and diagnosis of cancers

Edited by
Fu Wang, Abhishek Mahajan and Haibin Shi

Published in
Frontiers in Oncology



ISSN 1664-8714
ISBN 978-2-83251-169-5
DOI 10.3389/978-2-83251-169-5

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

The use of deep learning in mapping and diagnosis of cancers

Topic editors

Fu Wang — Xi'an Jiaotong University, China

Abhishek Mahajan — The Clatterbridge Cancer Centre, United Kingdom

Haibin Shi — Soochow University, China

Citation

Wang, F., Mahajan, A., Shi, H., eds. (2023). *The use of deep learning in mapping and diagnosis of cancers*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-169-5

Table of contents

TYPE Editorial
PUBLISHED 13 December 2022
DOI 10.3389/fonc.2022.1077341

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Giuseppe Esposito,
MedStar Georgetown University Hospital, United States

*CORRESPONDENCE
Abhishek Mahajan
drabhishek.mahajan@yahoo.in

SPECIALTY SECTION
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

RECEIVED 22 October 2022
ACCEPTED 22 November 2022
PUBLISHED 13 December 2022

CITATION
Mahajan A and Chakrabarty N (2022) Editorial: The use of deep learning in mapping and diagnosis of cancers. *Front. Oncol.* 12:1077341. doi: 10.3389/fonc.2022.1077341

Editorial: The use of deep learning in mapping and diagnosis of cancers

Abhishek Mahajan[1]* and Nivedita Chakrabarty[2]

[1]Department of Radiology, The Clatterbridge Cancer Liverpool, Liverpool, United Kingdom, [2]Department of Radiodiagnosis, Tata Memorial Hospital, Homi Bhabha National Institute (HBNI), Mumbai, India

KEYWORDS

deep learning - artificial neural network, artificial intelligence-AI, oncology, cancer imaging and image-directed interventions, systematic review and meta-analysis, original research article, machine learning and AI, oncoimaging

Editorial on the Research Topic
The use of deep learning in mapping and diagnosis of cancers

Deep Learning (DL) is a subset and an augmented version of Machine Learning (ML), which in turn is a subgroup of Artificial Intelligence (AI), that uses layers of neural networks, similar to human brain, for performing complex tasks quickly and accurately. AI can recognize patterns in a large volume of data and extract characteristics imperceptible to the human eye (1). Convolutional Neural Network (CNN) is the most commonly used network of DL, which contains multiple layers, with weighted connections between neurons that are trained iteratively to improve performance. DL can be supervised or unsupervised, but most of the practical uses of DL in cancer has been with supervised learning where labelled images are used for data training (2). Despite the growing number of uses of DL in cancer mapping and diagnosis, there are uncharted territories in DL which remain to be explored to utilize it to its full capacity. Also, in spite of the revolution in cancer research that DL has ushered in, there are a lot of challenges to overcome, before DL can be widely used and accepted in every corner of the world.

Role of DL in oncology

There has been an unprecedented surge in DL based research in oncology due to the availability of big data, powerful hardware and robust algorithms. Screening and diagnosis of cancer, prediction of treatment response, and survival outcome and recurrence prediction, are the various roles of ML and DL in cancer management. AI algorithms integrated with clinical decision support (CDS) tools can automatically mine electronic health record (EHR) and identify cohort that would benefit maximum from cancer screening programmes (3). For successful implementation of AI in cancer diagnosis, it is imperative for the radiologists and pathologists to collaborate with the key stakeholders, industrial partners and scientists (4). With ever increasing cancer burden worldwide, and availability of molecular targeted therapies, DL has served as an elixir, by its ability to screen, detect and diagnose tumours rapidly, and predict biomarkers non-invasively on imaging (5). Studies have shown that DL can be used to stage and grade tumours quickly and provide non-invasive histopathological diagnosis in cases where obtaining an invasive sample is risky. Patients, clinicians, radiologists and the pathologists, all have the potential to be benefitted by this DL technology as the utility of DL is no longer limited to tumour diagnosis, but to the cancer care as a whole. Prediction of overall survival, progression free survival, and disease free survival, assessment of response to treatment and outcome prediction are few of the many ways DL can benefit patients afflicted with cancer, the mere thought of which was previously unfathomable (5). Treatment planning and patient management can be hastened through the wider applications of DL based image interpretation, for example, non-responders to treatment detected on DL based baseline image interpretation, can be spared of further invasive treatment, and a change in management strategy may be considered for them.

Major applied uses of DL technology

Image classification and regression

DL can be used for classifying a lesion into benign or malignant, for treatment response evaluation and survival prediction. If DL models can be trained using a large dataset from a source domain, then it can be used in a target domain with a small sample size (2).

Object detection

DL can be used in tumour localization.

Semantic segmentation

DL can mark specific areas of concern on an image and assist the radiologists in decision making (2).

Image registration

Images acquired at different times can be accurately linked using DL, thus, enabling the radiologists to compare the images (2).

Federated learning

Robust deployable model can be built notwithstanding geographic boundaries, if multiple organizations/institutions/hospitals jointly train a model on a large data after de-identification of patient information (6).

Systematic review and meta-analysis data

A systematic review and meta-analysis from 1st January 2012 to 6th June 2019, comparing the diagnostic accuracy of health-care professionals with deep learning algorithms using imaging, found 10 studies on breast cancer, 9 studies on skin cancer, 7 studies on lung cancer, 5 studies on gastroenterological or hepatological cancers, 4 studies on thyroid cancer, 2 studies on oral cancer, and 1 study on nasopharyngeal cancer (7). Another systematic review on AI techniques in cancer diagnosis and prediction from articles published from 2009 to April 2021, revealed 10 articles pertaining to brain tumours, 13 articles related to breast cancer, 8 articles each related to cervical, liver, lung, and skin cancers, 6 articles related to colorectal cancer, 5 articles each related to renal and thyroid cancers, 2 articles each related to oral and prostate cancers, 7 articles related to stomach cancer, and 1 article each related to neuroendocrine tumours and lymph node metastasis (8). Few studies involving AI in cancer diagnosis and management include:

a. Histology prediction and screening of breast cancer on mammography (9, 10).
b. Brain tumour segmentation (11–14).
c. Lung nodule segmentation on computed tomography (CT) (15–17).
d. Liver tumour segmentation on CT (17, 18).
e. Prostate gland tumour detection on magnetic resonance imaging (MRI) (19, 20).
f. Brain tumour survival prediction (21–23).
g. F. Glioblastoma recurrence prediction (24).

Challenges and limitations of DL

a. Requirement of a large data: DL models need a large data (in thousands) to be trained and availability of such a huge data may not be possible in every institution.
b. Precise data annotation: Tumour region needs to be annotated or labelled accurately without contamination from surrounding non-tumour regions. This may not always be possible as many a times, tumours are

infiltrative in nature and not discrete, and may be located within a region containing some other pathology, for example, infiltrative lung tumour located within a collapsed lung, in which case precise margin delineation may not be possible.

c. There is need for equal representation of data on training and test sets failing which data gets skewed and bias is introduced (2).
d. Heterogeneity of data: Difference in training set of images and deployable image sets may affect the performance of a model, for example if the CT scanner used while acquiring images for training is different from the one on which the model is validated, then performance may be reduced.
e. Patient privacy concerns: Despite the available methods for deidentification of patient information, the problems of patient privacy still loom large (2).
f. Problem of hidden layers: DL uses multiple layers of neural network to analyse data, which remain hidden, and the exact reasoning of outcome is not decipherable, which makes it difficult to be relied upon and convincingly used.
g. Infrastructure: Use of DL requires a robust infrastructure which may not be available everywhere.
h. Lack of trained personnel and expertise and lack of awareness about collaboration for implementation of AI projects (25).

Imaging biobanks

Repositories of human tissue sample stored in an organized manner for research purpose is known as "biobank", and collection of medical image data for long term storage and retrieval for research is known as "imaging biobank" (26, 27) Digital Imaging and Communications in Medicine (DICOM) is the universal format for Picture Archiving and Communication System (PACS) storage and data sharing across all institutions (26). The data needs to be de-identified and informed consent of the patient obtained prior to data archiving (28). Few examples of such open-source platforms include The Cancer Genome Atlas (TCGA) program, The Cancer Imaging Archive (TCIA), and European Genome–phenome Archive (EGA) (29, 30). In India, collaboration between the Department of Biotechnology (Government of India) under the guidance of the National Institution for Transforming India (NITI) Aayog, and Tata Memorial Centre has led to the creation of The Tata Memorial Center Imaging Biobank (31). World's biggest multi-modality imaging study was commenced by the UK Biobank in 2014 to have a repository of neuro, cardiac, and abdominal MRI imaging, dual energy x-ray absorptiometry (DEXA) and carotid ultrasonography (32). Similarly, CAN-I-AID (Cancer Imaging Artificial Intelligence Database) biobank project has been initiated by Dr. Abhishek Mahajan at the Clatterbridge Cancer Centre, Liverpool, United Kingdom (UK). Such imaging biobanks for public use should be encouraged as it fulfils the requirement of large image data to promote DL based research across the globe.

Articles in research topic

In this Research Topic, we present 20 topics, 19 of which are original articles and one is a systematic review. *There is one article on cervical cancer screening*: Sun et al. used Stacking-Integrated Machine Learning Algorithm based on demographic, behavioural, and clinical factors to accurately identify women at high risk of developing cervical cancer and suggested the use of this model to personalise cervical cancer screening programme. *Three articles on lung cancer*: Shen et al. showed that DL based CT images have the potential to accurately predict malignancy and invasiveness of pulmonary subsolid nodules on CT Images and thus aid in management decisions. Sun et al. conducted a study to establish the role of Convolutional Neural Network-Based Diagnostic Model to differentiate between benign and malignant lesions manifesting as a solid, indeterminate solitary pulmonary nodule (SPN) or mass (SPM) on computed tomography (CT). Xia et al. compared and fused DL and Radiomics features of ground-glass nodules to predict the invasiveness risk of stage-I lung adenocarcinomas in CT scan and concluded that fusion of DL and radiomics features can refine the classification performance for differentiating non-invasive adenocarcinoma (non-IA) from IA and the prediction of invasiveness risk of GGNs is similar to or better than radiologists using AI scheme. *One article on thyroid cancer:* Wu et al. combined ACR TI-RADS with DL by training three commonly used deep learning algorithms to differentiate between benign and malignant in TR4 and TR5 thyroid nodules with available pathology and concluded that irrespective of the type of TI-RADS used for the classification competition, DL algorithms outperformed radiologists. *One article on bladder cancer:* Zhang et al. proposed a DL model based on CT images to predict muscle-invasive status of bladder carcinoma pre-operatively and concluded that DL model exhibited relatively good prediction ability with capability to enhance individual treatment of bladder carcinoma. *One article on periampullary region:* Tang et al. used DL to identify periampullary regions on MRI images and achieved optimal accuracies in the segmentation of the peri-ampullary regions on both T1 and T2 MRI images concordant with manual human assessment. *One article on rectal cancer:* Zhang et al. segmented rectal cancer *via* 3D V-Net on T2WI and DWI and then compared the radiomics performance in predicting KRAS/NRAS/BRAF status between DL-based auto segmentation and manual-based segmentation. They concluded that 3D V-Net architecture could conduct reliable rectal cancer segmentation on T2WI and DWI images. *One article on jaw lesions:* Chai et al. showed that AI-based cone-beam CT can distinguish between Ameloblastoma and Odontogenic Keratocyst with better accuracy than the surgeons. *Two articles on spine:* Ouyang et al. evaluated the efficiency of DL-based automated detection of primary spine tumours on MRI using the turing test. Hallinan et al. developed a DL model for classifying metastatic epidural spinal cord compression on MRI and which had comparable agreement to a subspecialist radiologist and clinical specialists. *One article on kidney tumour:* Sun et al. conducted a study on kidney tumour segmentation based on FR2PAttU-Net model. *One article on brain tumour:* Kandalgaonkar et al. conducted a study predicting IDH subtype of Grade 4 Astrocytoma and Glioblastoma from tumour radiomic patterns extracted from Multiparametric MRI using a machine learning approach and inferred that it may be used in either escalating or de-escalating adjuvant therapy for gliomas or for using targeted agents in future. *One article on survival rate prediction in cancer patients:* Sinzinger et al. developed Spherical Convolutional Neural Networks for survival rate prediction in cancer patients and concluded that it is beneficial in cases where expert annotations are not available or difficult to obtain. *One systematic review and meta-analysis:* Guha et al. performed a systematic review and meta-analysis differentiating primary central nervous system lymphoma (PCNSL) from glioblastoma (GBM) using deep learning and radiomics based ML approach. *There are five non-imaging related articles:* Zhu et al. developed transparent machine learning pipeline to efficiently predict Microsatellite instability (MSI), thus, helping pathologists to guide management decisions. Wang et al. conducted a study to reveal the heterogeneity in the tumor microenvironment of pancreatic cancer and analyze the differences in prognosis and immunotherapy responses of distinct immune subtypes. Menon et al. explored the histological similarities across cancers from a deep learning perspective. Huang et al. studied the effects of biofilm nano-composite drugs OMVs-MSN-5-FU on cervical lymph node metastases from oral squamous cell carcinoma (OSCC) on the animal model. Zormpas-Petridis et al. prepared a DL pipeline for mapping tumour heterogeneity on low-resolution whole-slide digital histopathology images. Figure 1 shows the list of authors based on type of articles submitted towards Research Topic.

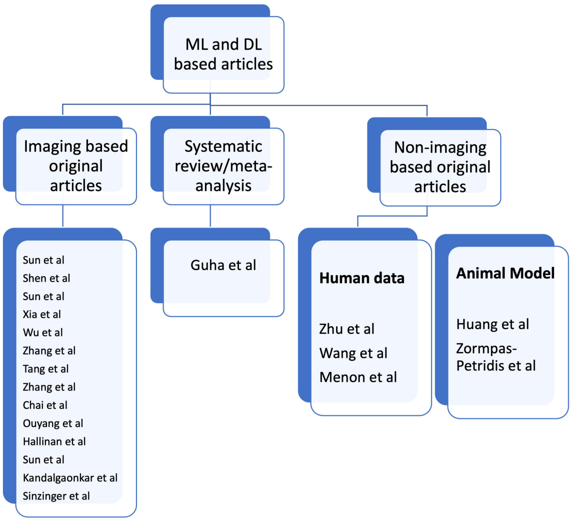


FIGURE 1
List of authors based on type of articles submitted towards research topic.

Conclusions

DL has ushered in revolution in the field of oncology research, from cancer screening and diagnosis, to response assessment and survival prediction, thus positively influencing patient management. With the increasing cancer burden and limited number of specialized healthcare providers, there is a growing inclination to use DL at various levels of cancer diagnosis to cater to the needs of patients and the healthcare providers alike. Despite the umpteen benefits, there are a few challenges that DL needs to conquer, before it can be ubiquitously used. Through this Research Topic, we wish to acquaint the readers with the latest ongoing DL based research in cancer diagnosis, which can pave the way for further innovations and research in this field, as full potential of DL is still underutilized.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. *Radiology* (2018) 288(2):318–28. doi: 10.1148/radiol.2018171820

2. Cherian Kurian N, Sethi A, Reddy Konduru A, Mahajan A, Rane SUA. 2021 update on cancer image analytics with deep learning. *WIREs Data Min Knowl Discov* (2021) 11:e1410. doi: 10.1002/widm.1410

3. Bizzo BC, Almeida RR, Michalski MH, Alkasab TK. Artificial intelligence and clinical decision support for radiologists and referring providers. *J Am Coll Radiol* (2019) 16(9 Pt B):1351–6. doi: 10.1016/j.jacr.2019.06.010

4. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian association of radiologists (CAR) artificial intelligence working group. canadian association of radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* (2018) 69(2):120–35. doi: 10.1016/j.carj.2018.02.002

5. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* (2021) 13(1):152. doi: 10.1186/s13073-021-00968-x

6. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med* (2020) 3(1):119. doi: 10.1038/s41746-020-00323-1

7. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit Health* (2019) 1(6):e271–97. doi: 10.1016/S2589-7500(19)30123-2

8. Kumar Y, Gupta S, Singla R, Hu Y-C. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch Comput Methods Eng* (2022) 29(4):2043–70. doi: 10.1007/s11831-021-09648-w

9. Sapate S, Talbar S, Mahajan A, Sable N, Desai S, Thakur M. Breast cancer diagnosis using abnormalities on ipsilateral views of digital mammograms. *Biocybernetics Biomed Eng* (2020) 40(1):290–305. doi: 10.1016/j.bbe.2019.04.008

10. Sapate SG, Mahajan A, Talbar SN, Sable N, Desai S, Thakur M. Radiomics based detection and characterization of suspicious lesions on full field digital mammograms. *Comput Methods Programs Biomed* (2018) 163:1–20. doi: 10.1016/j.cmpb.2018.05.017

11. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint* (2021). arXiv:2107.02314.

12. Baid U, Talbar S, Rane S, Gupta S, Thakur MH, Moiyadi A, et al. A novel approach for fully automatic intra-tumor segmentation with 3D U-net architecture for gliomas. *Front Comput Neurosci* (2020) 14:10. doi: 10.3389/fncom.2020.00010

13. Mehta R, Filos A, Baid U, Sako C, McKinley R, Rebsamen M, et al. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation–analysis of ranking metrics and benchmarking results. *arXiv e-prints* (2021).

14. Pati S, Baid U, Zenk M, Edwards B, Sheller M, Reina GA, et al. The federated tumor segmentation (fets) challenge. *arXiv preprint* (2021).

15. Singadkar G, Mahajan A, Thakur M, Talbar S. Deep deconvolutional residual network based automatic lung nodule segmentation. *J Digit Imaging* (2020) 33(3):678–684. doi: 10.1007/s10278-019-00301-4

16. Singadkar G, Mahajan A, Thakur M, Talbar S. Automatic lung segmentation for the inclusion of juxtapleural nodules and pulmonary vessels using curvature based border correction. *J King Saud University-Computer Inf Sci* (2021) 33 (8):975–87. doi: 10.1016/j.jksuci.2018.07.005

17. Kumar YR, Muthukrishnan NM, Mahajan A, Priyanka P, Padmavathi G, Nethra M, et al. Statistical parameter-based automatic liver tumor segmentation from abdominal CT scans: A potential radiomic signature. *Proc Comput Science* (2016) 93:446–52. doi: 10.1016/j.procs.2016.07.232

18. Rela M, Krishnaveni BV, Kumar P, Lakshminarayana G. Computerized segmentation of liver tumor using integrated fuzzy level set method. *AIP Conf Proc* (2021) 2358(1):60001. doi: 10.1063/5.0057980

19. Hambarde P, Talbar SN, Sable N, Mahajan A, Chavan SS, Thakur M. Radiomics for peripheral zone and intra-prostatic urethra segmentation in MR imaging. *Biomed Signal Process Control* (2019) 51:19–29. doi: 10.1016/j.bspc.2019.01.024

20. Bothra M, Mahajan A. Mining artificial intelligence in oncology: Tata memorial hospital journey. *Cancer Res Stat Treat* (2020) 3:622–4. doi: 10.4103/CRST.CRST_59_20

21. Davatzikos C, Barnholtz-Sloan JS, Bakas S, Colen R, Mahajan A, Quintero CB, et al. AI-Based prognostic imaging biomarkers for precision neuro-oncology: theReSPOND consortium. *Neuro-oncology* (2020) 22(6):886–8. doi: 10.1093/neuonc/noaa045

22. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. MahIdentifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint* (2018).

23. Baid U, Rane SU, Talbar S, Gupta S, Thakur MH, Moiyadi A, et al. Overall survival prediction in glioblastoma with radiomic features using machine learning. *Front Comput Neurosci* (2020) 14:61. doi: 10.3389/fncom.2020.00061

24. Akbari H, Mohan S, Garcia JA, Kazerooni AF, Sako C, Bakas S, et al. Prediction of glioblastoma cellular infiltration and recurrence using machine learning and multi-parametric mri analysis: Results from the multi-institutional respond consortium. *Neuro-Oncology* (2021) 23(Supplement_6):vi132–3. doi: 10.1093/neuonc/noab196.522

25. Mahajan A, Vaidya T, Gupta A, Rane S, Gupta S. Artificial intelligence in healthcare in developing nations: The beginning of a transformative journey. *Cancer Research Statistics Treat* (2019) 2(2):182. doi: 10.4103/CRST.CRST_50_19

26. Mantarro A, Scalise P, Neri E. Imaging biobanks, big data, and population-based imaging biomarkers. In: *Imaging biomarkers: Development and clinical integration* (2017). Switzerland:Springer International Publishing. p. 153–7.

27. Woodbridge M, Fagiolo G, O'Regan DP, et al. MRIdb: Medical image management for biobank research. *J Digit Imaging* (2013) 26(5):886–90. doi: 10.1007/s10278-013-9604-9

28. Available at: https://car.ca/news/new-car-white-paper-on-ai-provides-guidance-on-de-identification-of-medical-imaging-data/ (Accessed on 17/10/2022).

29. Geis JR, Brady A, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in radiology: Summary of the joint European and north American multisociety statement. *Insights Into Imaging* (2019) 293(2):436–440. doi: 10.1148/radiol.2019191586

30. Available at: https://www.cancerimagingarchive.net/ (Accessed on 17/10/2022).

31. Bothra M, Mahajan A. Mining artificial intelligence in oncology: Tata memorial hospital journey. *Cancer Res Stat Treat* (2020) 3:622–4. doi: 10.4103/CRST.CRST_59_20

32. Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, et al. The UK biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nat Commun* (2020) 11(1):2624. doi: 10.1038/s41467-020-15948-9

frontiers in Oncology

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fonc.2020.00418

Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan

Xianwu Xia [1†], Jing Gong [2,3†], Wen Hao [2,3], Ting Yang [1], Yeqing Lin [1*], Shengping Wang [2,3*] and Weijun Peng [2,3*]

[1] Department of Radiology, Municipal Hospital Affiliated to Medical School of Taizhou University, Taizhou, China, [2] Department of Radiology, Fudan University Shanghai Cancer Center, Shanghai, China, [3] Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Peng Huang,
Johns Hopkins University,
United States

Reviewed by:
Yutong Li,
University of Illinois at
Urbana-Champaign, United States
Zhongxiang Ding,
Zhejiang University, China

***Correspondence:**
Yeqing Lin
yegu303@163.com
Shengping Wang
shengpingwang2007@126.com
Weijun Peng
cjr.pengweijun@vip.163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to
Cancer Imaging and Image-directed
Interventions,
a section of the journal
Frontiers in Oncology

Received: 29 September 2019
Accepted: 10 March 2020
Published: 31 March 2020

Citation:
Xia X, Gong J, Hao W, Yang T, Lin Y, Wang S and Peng W (2020) Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan. Front. Oncol. 10:418. doi: 10.3389/fonc.2020.00418

For stage-I lung adenocarcinoma, the 5-years disease-free survival (DFS) rates of non-invasive adenocarcinoma (non-IA) is different with invasive adenocarcinoma (IA). This study aims to develop CT image based artificial intelligence (AI) schemes to classify between non-IA and IA nodules, and incorporate deep learning (DL) and radiomics features to improve the classification performance. We collect 373 surgical pathological confirmed ground-glass nodules (GGNs) from 323 patients in two centers. It involves 205 non-IA (including 107 adenocarcinoma *in situ* and 98 minimally invasive adenocarcinoma), and 168 IA. We first propose a recurrent residual convolutional neural network based on U-Net to segment the GGNs. Then, we build two schemes to classify between non-IA and IA namely, DL scheme and radiomics scheme, respectively. Third, to improve the classification performance, we fuse the prediction scores of two schemes by applying an information fusion method. Finally, we conduct an observer study to compare our scheme performance with two radiologists by testing on an independent dataset. Comparing with DL scheme and radiomics scheme (the area under a receiver operating characteristic curve (AUC): 0.83 ± 0.05, 0.87 ± 0.04), our new fusion scheme (AUC: 0.90 ± 0.03) significant improves the risk classification performance ($p < 0.05$). In a comparison with two radiologists, our new model yields higher accuracy of 80.3%. The kappa value for inter-radiologist agreement is 0.6. It demonstrates that applying AI method is an effective way to improve the invasiveness risk prediction performance of GGNs. In future, fusion of DL and radiomics features may have a potential to handle the classification task with limited dataset in medical imaging.

Keywords: lung adenocarcinoma, deep learning, radiomics, invasiveness risk, ground-glass nodule, CT scan

INTRODUCTION

As the most common histologic subtype of lung cancer, lung adenocarcinomas accounts for almost half of lung cancers. The persistent presence of ground-glass nodules (GGN) in computed tomography (CT) image usually serves as an indicator of the presence of lung adenocarcinoma or its precursors (1). According to the guideline of the 2011 International Association for the

Study of Lung Cancer/American Thoracic Society/European Respiratory Society International (IASLC/ATS/ERS) classification, lung adenocarcinoma includes atypical adenomatous hyperplasia (AAH), adenocarcinoma *in situ* (AIS), and minimally invasive adenocarcinoma (MIA) and invasive adenocarcinoma (IA) (2). Previous reported studies has depicted that the different subtypes of lung adenocarcinoma have different 3-years and 5-years disease-free survival (DFS) rates (3). For stage-I lung adenocarcinoma, the 5-years DFS of AIS and MIA is 100%, but IA is only 38–86% (4, 5). Meanwhile, the standard surgical treatment for lung adenocarcinoma is still lobectomy, but non-IA patients may be candidates for limited surgical resection (6). Thus, it is important to discriminate between IA and non-IA (including AIS and MIA) by using non-invasive CT image.

In order to classify between non-IA and IA GGNs, investigators and researchers have proposed two kinds of computer-aided diagnosis (CADx) schemes including CT radiomics feature analysis method and deep learning (DL) architecture based scheme (7). The radiomics feature analysis approach mainly includes tumor segmentation, radiomics feature extraction and selection (8), and machine-learning classifier training/testing process, respectively (9–11). The related studies usually compute a large number of handcrafted imaging features to decode the different tumor phenotypes (6, 12–14). Unlike radiomics feature analysis scheme, DL based scheme use the convolutional neural network (CNN) to build an end-to-end classification model by learning a hierarchy of internal representations (15–17). Although DL scheme can improve the classification performance and reduce the workload of hand-craft feature engineering (i.e., tumor boundary delimitation), it needs to be trained with larger dataset than radiomics feature based scheme (18, 19). However, under common medical diagnosis conditions, collecting, and building a large uniform image dataset is very difficult because of the inconformity of CT screening standard and lacking surgical pathological confirmed GGNs. Thus, how to improve the CADx performance with a limited dataset is a challenge task.

To address this issue, we have fused the DL and radiomics features to build a new AI scheme to classify between non-IA and IA GGNs. We first collected 373 surgical pathological confirmed GGNs from 323 patients in two centers. To segment the GGNs in CT images, we trained a recurrent residual convolutional neural network (RRCNN) based on U-Net model. Then, we respectively built a DL model and radiomics feature analysis mode to classify between IA and non-IA GGNs. Finally, we applied an information fusion method to fuse the prediction scores generated by the two models. In order to evaluate the performance of our new scheme, we used an independent dataset to conduct an observer study by comparing our prediction score with two radiologists (an experienced senior radiologist S.P. Wang and a junior radiologist W. Hao).

MATERIALS AND METHODS

Image Dataset

In this study, we respectively collected 373 surgical pathological confirmed GGNs from two centers. For the cases with multifocal ground-glass nodules (multi-GGNs), we treated each GGN as an independent primary lesion (20). The inclusion criteria were: (1) diagnosed with stage-I lung adenocarcinoma cancer; (2) histopathologically confirmed AIS, MIA and IA pulmonary nodules; (3) available CT examination within 1 month before surgery; and (4) the tumor manifesting as GGN on CT with a maximum diameter of (3 mm, 30 mm). The exclusion criteria were: (1) preoperative systemic therapy; (2) lacking CT images before surgery; (3) histopathologically described GGN not identifiable on CT; and (4) artifacts appeared in CT images. We only collected the latest CT examination images of each patient before surgery. The time interval between chest CT examination and operation was 1–30 days (mean, 8.3 days). The institutional review board of two centers approves this retrospective study, and written informed consents were waived from all patients. The details of GGNs in the two centers were depicted as follows.

In the first dataset, we collected 246 GGNs from 229 patients (involving 82 males and 147 females) in Taizhou Municipal Hospital (Zhejiang, China). Among these nodules, 55 GGNs were AIS, 64 GGNs were MIA, and 127 GGNs were IA. All the CT scans were reconstructed by using the standard convolution kernel, and each slice was reconstructed with a matrix 512 × 512 pixels (GE scanner). CT parameters were as follows: 120 kVp tube voltage, and 100–250 mA tube current. The pixel spacing of CT scan ranged from 0.684 to 0.703 mm, and the slice thickness was 1.25 or 5 mm.

The other 127 GGNs were collected from 94 patients (involving 35 males and 59 females) in Fudan University Shanghai Cancer Center (Shanghai, China). In this dataset, 52 AIS GGNs, 34 MIA GGNs, and 41 IA GGNs were involved. The CT examinations were performed with a fixed tube voltage of 120 kVp and a tube current of 200 mA. The pixel spacing of CT image ranged from 0.684 to 0.748 mm, and the slice thickness was 1 or 1.5 mm. Each slice was reconstructed with an image matrix of 512 × 512 pixels.

In order to train and test our proposed schemes, we divided the GGNs into two parts. We used 246 GGNs in the first dataset to build a training and validation dataset to train our scheme. Meanwhile, to evaluate our new scheme performance, we selected the 127 GGNs in the second part to build an independent testing dataset. The details of our dataset were listed in **Table 1**.

Methods

In this study, we first built a DL based model and a radiomics feature based model, respectively. Then, to improve the scheme performance, we used an information-fusion method to fuse the prediction scores of the two schemes. The framework of our proposed scheme was illustrated in **Figure 1**.

Before building the scheme, we first used a series of preprocessing technique to process the initial CT images. To avoid the biases caused by the variant spacing of CT scans in our dataset, we applied a cubic spline interpolation algorithm to resample CT images to a new spacing of 1 mm × 1 mm × 1 mm. Then, we used an intensity window range of [−1,200, 600] to scale the resampled axial CT images to an intensity range of 0–255. After normalized all the CT images, we cropped the GGN into a 3D cubes with a patch of 64 × 64× 64 mm. During this process, we used the position of GGN center point in Cartesian

TABLE 1 | Demographic characteristics of 323 patients with 373 GGNs in two datasets.

Characteristic		**Training and validation dataset (*N* = 246)**			**Testing dataset (*N* = 127)**		
		Non-IA	**IA**	***P***	**Non-IA**	**IA**	***P***
		119	127		86	41	
Sex	Male	40	42	0.15	19	16	0.15
	Female	73	74		43	16	
Age (mean ± SD, year)		56.5 ± 11.8	59.7 ± 10.3	0.03	51.8 ± 12.1	58.1 ± 8.6	0.03
Location	RUL	48 (19.5%)	52 (21.1%)	0.64	28 (22.0%)	18 (14.2%)	0.13
	RML	6 (2.4%)	9 (3.7%)		6 (4.7%)	3 (2.4%)	
	RLL	17 (6.9%)	19 (7.7%)		15 (11.8%)	7 (5.5%)	
	LUL	34 (13.8%)	32 (13.0%)		25 (19.7%)	7 (5.5%)	
	LLL	14 (5.7%)	15 (6.1%)		12 (9.4%)	6 (4.7%)	
Diameter (mm)	(3, 10)	72 (29.3%)	42 (17.1%)	0.004	67 (52.8%)	8 (6.3%)	<0.0001
	(10, 20)	39 (15.9%)	68 (27.6%)		19 (15.0%)	22 (17.3%)	
	(20, 30)	8 (3.3%)	17 (6.9%)		0 (0%)	11 (8.7%)	
Type	pGGN	88 (35.8%)	65 (26.4%)	0.0002	78 (61.4%)	18 (14.2%)	<0.0001
	sGGN	31 (12.6%)	62 (25.2%)		8 (6.3%)	23 (18.1%)	

IA, invasive adenocarcinoma; pGGO, pure ground glass nodule; sGGN, part-solid ground glass nodule.

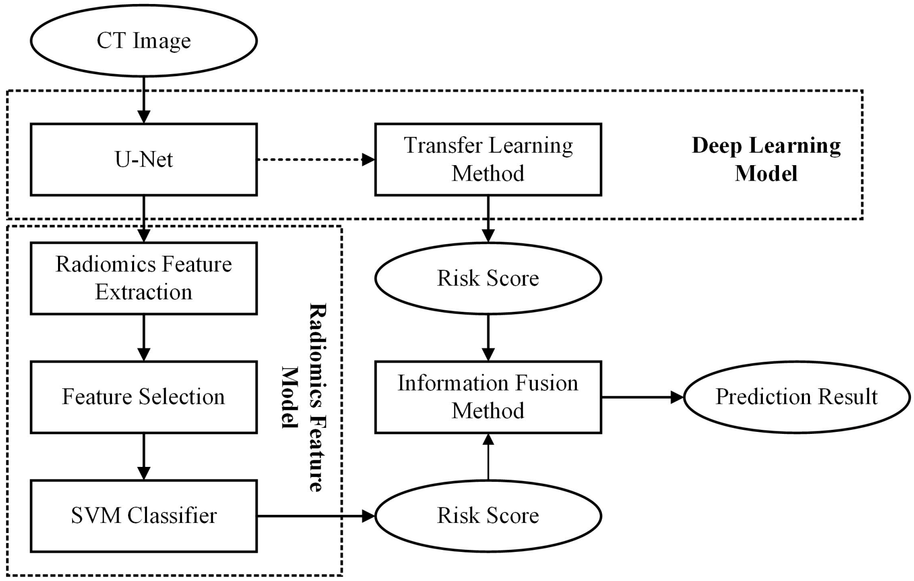


FIGURE 1 | Flowchart of the proposed scheme.

coordinates drawn by radiologist to locate each GGN in CT image. Last, in order to reduce the computational cost of our model, we normalized the intensity of cropped GGN cubes to an intensity range of 0–1.

Second, we built a 3D RRCNN based on U-Net model to segment the GNNs in CT images. The architecture of our segmentation DL model were showed in **Figure 2**. The inputs of 3D RRCNN model were our cropped GGN patches, and the outputs were the segmented 3D masks. For each layer of the 3D RRCNN, we used a RRCNN block with a 3 × 3 × 3 convolutional layer, a batch normalization layer and a standard rectified linear unit (ReLU). In each convolutional layer, we also embedded a residual unit and a recurrent unit into the block (21). To build the segmentation model, we used the 257 GGNs in the lung image database consortium and image database resource initiative (LIDC-IDRI) to train our proposed RRCNN model (22). Four radiologists delineated the boundaries of nodules in LIDC-IDRI database. We used the boundary voted by three or more radiologists as the "ground-truth" of each nodule. To generate the training GGNs for RRCNN model, we

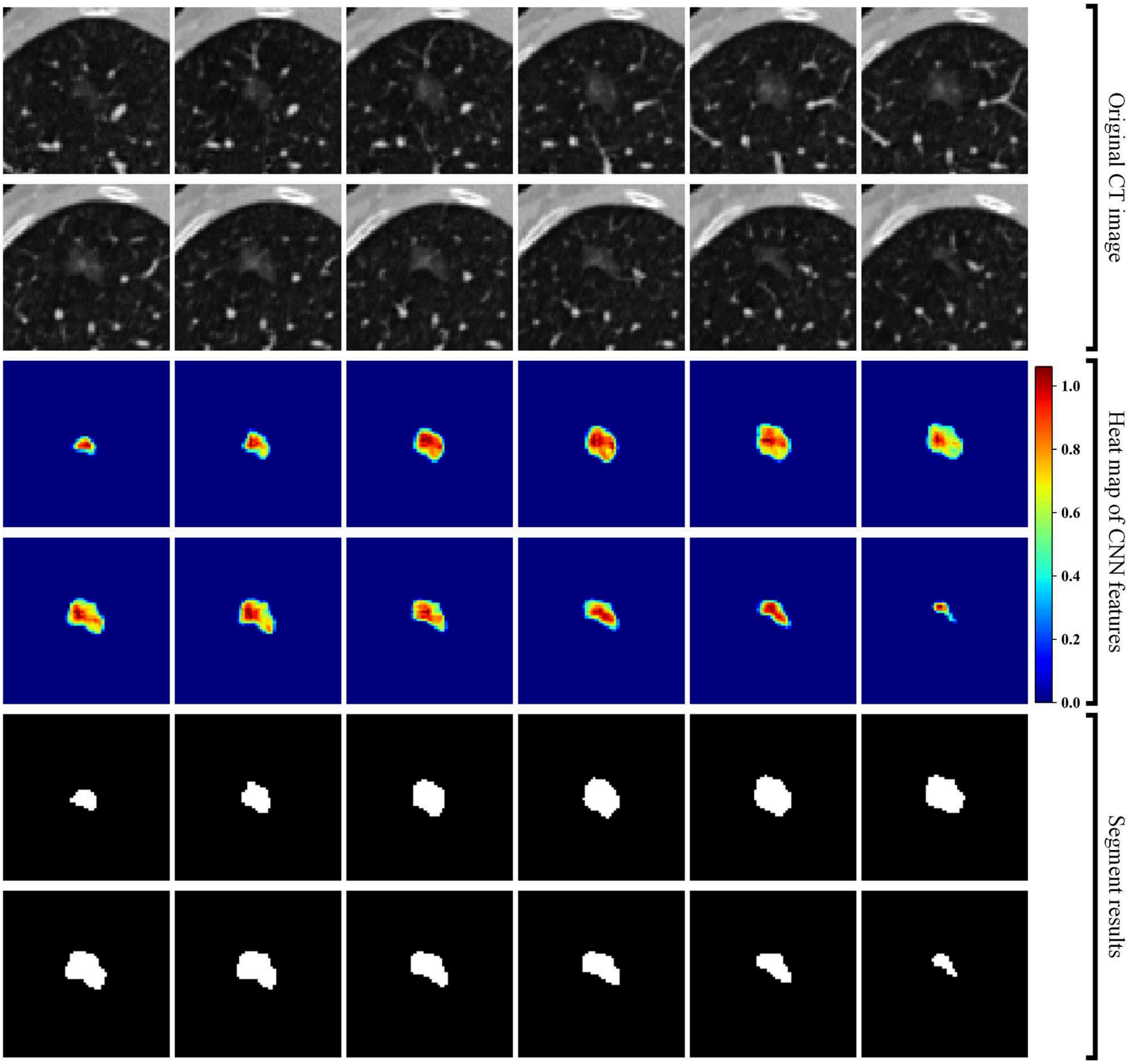


FIGURE 2 | Segmentation results of a GGN. From top to bottom: original CT images, heat map of CNN features, and segment masks of the GGN.

applied some data augmentation techniques (i.e., rotation of image by 90° increments, left-right flipping, up-down flipping) to augment the dataset. Moreover, we applied the Dice similarity coefficient (DSC) of nodule to define the loss function of our segmentation model (23). **Figure 2** shows an example of GGN segmentation results.

Third, we used a transfer learning method to build a DL based invasiveness risk prediction model. In this model, we fixed the parameters in CNN-pooling processes of the segmentation model. To build a classification model, we added two fully connected (FC) layers into the DL model, and used deep features generated by the CNN-pooling layers of segmentation model to feed into the FC layers. Then, we used the GGNs in our training and validation dataset to fine-tune our classification CNN model. In this process, we selected the cross entropy to calculate the loss, and used an Adam optimizer with a weight decay of 1e-4 to update the parameters. **Figure 3** shows the architectures of our proposed DL model.

Fourth, we built a radiomics feature analysis model to classify between non-IA and IA GGNs. For each CT scan in our dataset, we used the RRCNN model to segment 3D GGNs. Then, we computed 1,218 radiomics features to quantify each GGN. These imaging features involved: 430 LoG features, 688 wavelet features, 18 histogram features, 14 shape features, and 68 texture features. The LoG features and wavelet features were computed by using

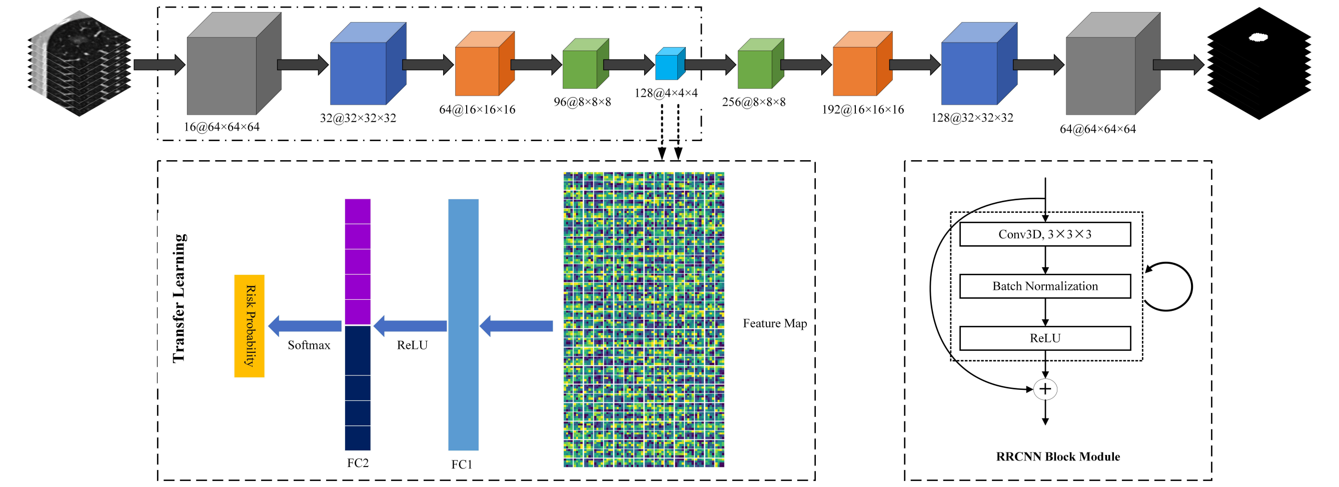


FIGURE 3 | The architectures of Recurrent Residual Convolutional Neural Network (RRCNN) based on U-Net model and the transfer learning method based risk prediction model.

the Laplacian of Gaussian (LoG) filter and wavelet filter to filter the initial image, respectively. The LoG image was obtained by convolving the original image with the second derivative of a Gaussian kernel. Five sigma values including 1, 2, 3, 4, and 5 were used to calculate the LoG features. In Among the 68 texture features, 22 were gray level co-occurrence matrix texture features (GLCM), 14 were gray level dependence matrix texture features (GLDM), 16 were gray level run length matrix texture features (GLRLM), and 16 were gray level size zone matrix texture features (GLSZM). After extracting the radiomics features, we scaled each feature to [0, 1] by using a feature normalization technique. To reduce the dimensionality of initial features, we applied the univariate feature selection method with ANOVA *F*-value to select the best features and remove the redundant features (24). After feature selection processing, we used these selected imaging features to train a support vector machine (SVM) classifier and build a radiomics feature based model.

Finally, we used an information-fusion method to fuse the prediction scores of two classification models. In brief, the information-fusion strategies includes the maximum, minimum, and weighting average fusion. For maximum and minimum strategy, we compared two prediction scores of each GGN, and selected the maximum or minimum value as the fusion prediction score. For weighting average strategy, we systematically increased the weighting factor of prediction score generated by DL based scheme from 0.1 to 0.9 (or 0.9–0.1 for the prediction score generated by radiomics feature based scheme) to compute the fusion prediction score. A similar method was applied in our previously reported literature (25).

Performance Evaluation

After obtaining the prediction scores, we generated the receiver operating characteristic (ROC) curves and computed the area under a ROC curve to evaluate the performance of our proposed models. In order to compare the new scheme performance with radiologists, we conducted an observer study by testing on an independent testing dataset. Two radiologists (a junior radiologist: Wen Hao with 5-years experience; a senior radiologist: Shengping Wang with 14-years experience in CT interpretation) were independently to diagnose all the GGNs in testing dataset by blinding to the histopathologic results and clinical data. Since two radiologists only provided a binary result for each case, we calculated some additional metrics to assess and compare the prediction performance. The evaluation indexes were accuracy (ACC), F1 score, weighted average F1 score, and Matthews correlation coefficient ($\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$), respectively. The equation of F1 score was defined as follows.

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

where TP, FP, TN, FN denoted true positive, false positive, true negative, and false negative, respectively. Precision denoted the precision value ($\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$), and Recall denoted the recall value ($\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$).

In this study, we implemented the above model building and performance evaluation processes on the Python 3.6 by using a computer with Intel Core i7-8700 CPU 3.2 GHz × 2, 16 GB RAM and a NVIDIA GeForce GTX 1,070 graphics processing unit. To build the DL and radiomics feature based scheme, we applied some publicly available Python packages, i.e., SimpleITK, pyradiomics (26), Pytorch, scikit-learn, scikit-feature, scipy. We used the default configuration of performance evaluation functions. Thus, the scheme performance can be easily compared and evaluated in future studies.

All the codes of our proposed models were open source available at https://github.com/GongJingUSST/DL_Radiomics_Fusion.

RESULTS

Table 1 listed the detailed demographic characteristics of the patients in two datasets. A total of 323 patients [117 (36.2%)

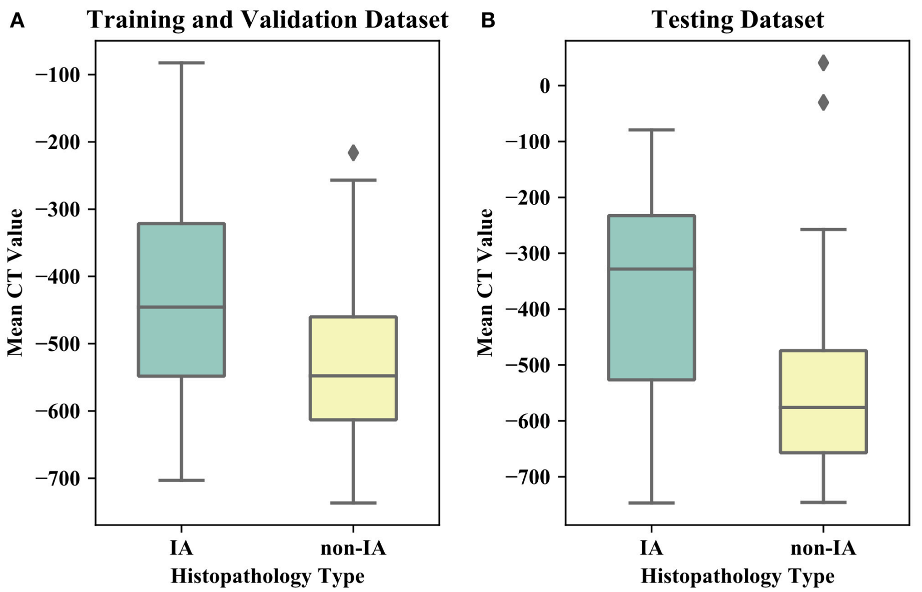


FIGURE 4 | Boxplots of the mean CT value of IA and non-IA GGNs in our dataset. **(A)** Illustrates boxplot of the training and validation dataset. **(B)** Shows boxplot of the testing dataset.

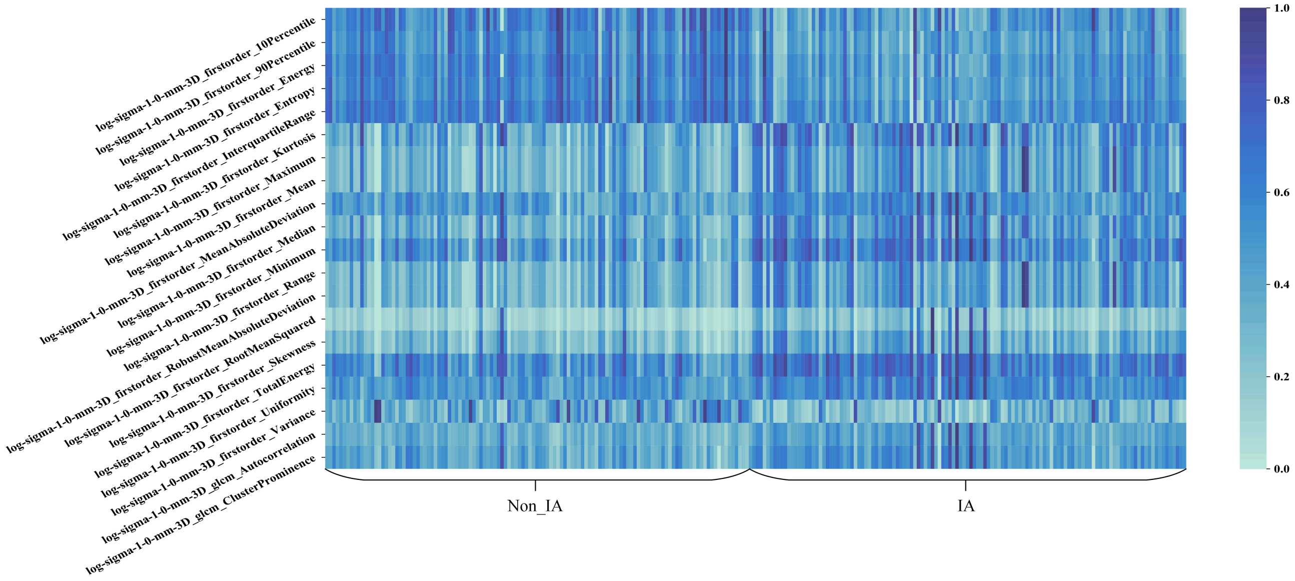


FIGURE 5 | Heat map of the 20 imaging features selected in the radiomics based model.

males, and 206 (63.8%) females, $P > 0.05$] with 373 GGNs were involved in our dataset. Among these GGNs, 107 were AIS (28.7%), 98 were MIA (26.3%), and 168 were IA (45%). Of all 373 GGNs, 228 (61.1%) were located in right lobe, and 145 (38.9%) were located in left lobe ($P > 0.05$). In the dataset, the diameters of 189 (50.7%) GGNs were smaller than 10 mm, the diameters of 148 (39.7%) GGNs were in a range of (10 mm, 20 mm), and the diameters of 36 (9.6%) GGNs were larger than 20 mm ($P < 0.05$). Of 373 GGNs, 249 nodules (66.8%) showed pure GGNs without solid components, and 124 nodules (33.2%) showed part-solid GGNs on CT images. **Figure 4** illustrates the boxplots of GGN mean CT values in training and testing dataset. In training and validation dataset, the mean CT value of IA and non-IA GGNs were −439 ± 138 and −533 ± 116, respectively. Meanwhile, in the testing dataset, the mean CT value of IA and non-IA were −381 ± 182 and −553 ± 142.

Figure 5 shows the heat map of the 20 selected imaging features in the radiomics feature based scheme. In **Figure 5**,

TABLE 2 | AUC values and the corresponding 95% CI generated by different methods with 127 GGNs in testing dataset.

Method	AUC	95% CI
Deep learning based scheme	0.83 ± 0.05	[0.75, 0.90]
Radiomics feature based scheme	0.87 ± 0.04	[0.80, 0.93]
Minimum	0.83 ± 0.05	[0.75, 0.90]
Maximum	0.90 ± 0.03	[0.84, 0.95]
0.1 × Radiomics[a]+0.9×DL[b]	0.85 ± 0.04	[0.77, 0.91]
0.2 × Radiomics+0.8×DL	0.86 ± 0.04	[0.78, 0.92]
0.3 × Radiomics+0.7×DL	0.87 ± 0.04	[0.80, 0.93]
0.4 × Radiomics+0.6×DL	0.88 ± 0.04	[0.81, 0.94]
0.5 × Radiomics+0.5×DL	0.89 ± 0.04	[0.83, 0.95]
0.6 × Radiomics+0.4×DL	0.90 ± 0.04	[0.83, 0.95]
0.7 × Radiomics+0.3×DL	0.90 ± 0.04	[0.83, 0.90]
0.8 × Radiomics+0.2×DL	0.90 ± 0.04	[0.83, 0.88]
0.9 × Radiomics+0.1×DL	0.89 ± 0.03	[0.83, 0.94]

[a] *Radiomics: prediction scores generated by radiomics feature based scheme.*
[b] *DL: prediction scores generated by deep learning based scheme.*

these 20 imaging features selected from the initial feature pool were LoG image based features. It can be seen that LoG features play an important role in building the radiomics feature based classification model. Most of the selected imaging features have a different distribution between non-IA and IA GGNs. It indicated that most of these selected features have a potential to differ non-IA from IA GGNs.

Table 2 listed the AUC values and the corresponding 95% confidence interval (CI) of the models proposed in this study. Testing on the independent testing dataset, the DL based scheme and radiomics feature based scheme yielded an AUC value of 0.83 ± 0.05 and 0.87 ± 0.04, respectively. When we applied the information-fusion method, the scheme performance changed with the different fusion strategy. By using a maximum fusion strategy, our scheme yielded a highest AUC value of 0.90 ± 0.03. Comparing with the performance generated individually, the fusion scheme significantly improved the scheme performance ($P < 0.05$). Meanwhile, there is no significant difference between DL based scheme and radiomics feature based scheme ($P = 0.09$).

Figure 6 shows performance comparisons of three models and radiologists. **Figure 6A** shows scatter plot of prediction score distributions of non-IA and IA nodules, and **Figure 6B** shows ROC curves of the three models and the prediction scores of two radiologists. **Figure 6A** showed that a large number of prediction scores generated by DL and radiomics based models were scattered and inconsistent in both non-IA and IA nodules. It indicated DL model and radiomics model might provide different information in classifying between non-IA and IA nodules. ROC curves also showed the trend that fusing the scores of DL based scheme and radiomics feature based scheme can improved the scheme performance. In a comparison with two radiologists, the fusion scheme yielded higher performance. In order to further compare the fusion scheme performance with two radiologists, **Table 3** illustrated and compared the accuracy, F1 score, weighted average F1 score, and Matthews correlation coefficient of each scheme. Evaluating the results showed in **Table 3**, our fusion scheme yielded higher performance than two radiologists in terms of each index. It indicated that our CADx scheme matched or even outperformed radiologist in classifying between non-IA an IA GGNs. To test the interrater reliability of the results of two radiologists, we also calculated the Cohen's kappa value to measure their agreement (27). The Cohen's kappa value of two radiologists was 0.6. It indicated that two radiologists had a moderate agreement in predicting the invasiveness risk of GGN.

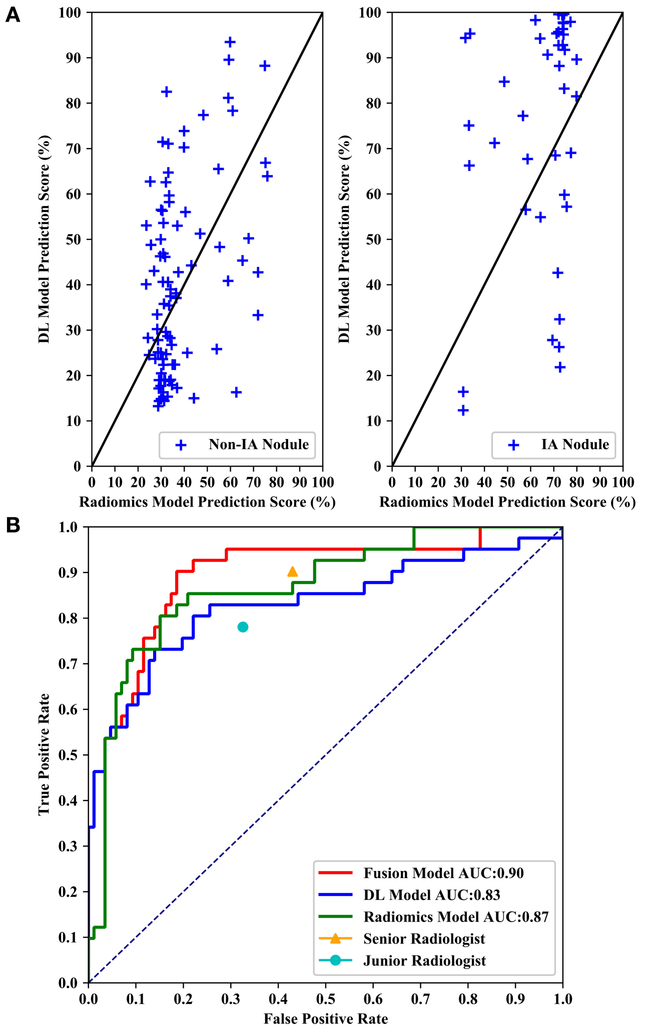


FIGURE 6 | Performance comparisons of three models and radiologists. **(A)** Shows scatter plots of prediction score distributions of non-IA and IA nodules. Left to right: prediction scores generated by DL and radiomics models for non-IA and IA nodules in testing dataset, respectively. **(B)** Shows ROC curves of the three models and the prediction scores of two radiologists.

DISCUSSION

In this study, we developed a CT image based CADx scheme to classify between non-IA and IA GGNs by fusing DL and

TABLE 3 | The comparison of classification performance tested on 127 GGNs in independent testing dataset, in terms of accuracy (ACC), F1 score, weighted average F1 score, and Matthews correlation coefficient (MCC), respectively.

	ACC (%)	F1 (%)	$F1_{weighted}$ (%)	MCC (%)
Senior radiologist	67.7	64.3	68.5	44.8
Junior radiologist	70.9	63.4	71.8	42.6
Our fusion model	80.3	75.2	80.9	62.8

TABLE 4 | Comparison of dataset, methods, and AUC values reported in different studies.

Work	Dataset	Method	AUC
Wang et al. (19)	1,545 nodules	Deep learning	0.892
Zhao et al. (15)	651 nodules	Deep learning	0.880
Gong et al. (28)	828 nodules	Deep learning	0.92 ± 0.03
Our study	373 nodules	Fusion of deep learning and radiomics	0.90 ± 0.03

radiomics features. Our study has a number of characteristics. First, we built an AI model to classify between non-IA and IA GGNs by fusing DL and radiomics features. Since DL based scheme and radiomics feature based scheme used different imaging features to decode the phenotypes of GGN, our fusion model integrated these quantitative and deep features to character the CT features of tumor. Comparing with model built with DL and radiomics features individually, the fusion model has improved the scheme performance significantly (i.e., results showed in **Table 2** and **Figure 6**). It showed that deep feature and radiomics feature may provide complementary information in predicting the invasiveness risk of GGN. To build a robust model, we used the surgery histopathological confirmed GGNs from two centers to train and test the classification scheme. In order to evaluate the performance of our scheme, we compared the scheme prediction scores with two radiologists by testing on an independent dataset. Comparing with two radiologists, our new scheme yielded higher performance in classifying between non-IA and IA GGNs (i.e., results showed in **Figure 6** and **Table 3**). Meanwhile, comparing with previously reported studies (15, 19, 28), our study can yield a rather high classification performance by using a limited dataset (i.e., results showed in **Table 4**). If the robustness of our model was confirmed with more diverse and larger dataset in future studies, the proposed AI scheme would have a high impact on assisting radiologists in their clinical diagnosis of GGNs.

Second, we applied a transfer learning method to build a DL based scheme by training with a limited dataset. Since the DL based scheme was a data-driven model, we should train and build a DL model with a large dataset. To address this issue, we proposed a RRCNN model to segment GGNs, and then used a transfer learning method to fine-tune the segmentation DL model. In this process, our classification DL model shared the same deep features with the segmentation model. As the training images of two model was same, it was easily to transfer the segmentation model to classification task. In a comparison with radiomics feature based model, the DL based scheme yielded equivalent performance ($P > 0.05$). It demonstrated that transferring segmentation DL model to classification task was feasible. Thus, our new scheme may provide a new way to build a DL based classification model with limited dataset.

Third, we built a radiomics feature based scheme to predict the invasiveness risk of GGN. To quantify the imaging phonotypes of GGN, we initially computed 1,218 radiomics features. To remove the redundant imaging features, we applied a univariate feature selection method to select the robust features. Most of the selected imaging features were LoG image based features. It showed that LoG features were essential for classifying between non-IA and IA GGNs. By observing the heat map of 20 selected image features, we found that those features had a different distributions in non-IA and IA group. It indicated that these selected imaging features had a potential to classify between non-IA and IA GGNs.

Fourth, in order to evaluate the performance of our proposed scheme, we conducted an observer study by comparing with two radiologists. Senior radiologist obtained higher sensitivity (90.2 vs. 78.1%) and false positive rate (43.0 vs. 32.6%) in distinguishing between IA and non-IA GGNs. It indicated that senior radiologist was more sensitive to the positive GGNs (i.e., IA GGNs). Meanwhile, the accuracy of senior radiologist was lower than that of junior radiologist. Since the number of non-IA GGNs is larger than that of IA GGNs in our testing dataset, it indicated that the number of negative GGNs (i.e., non-IA GGNs) miscategorized into IA class by senior radiologist was larger. Thus, senior radiologist paid more attention to IA GGNs than non-IA GGNs. Two radiologists had a moderate agreement on diagnosing the invasiveness risk of GGNs. By validating on an independent testing dataset, our AI scheme outperformed two radiologists in classifying between non-IA and IA GGNs (i.e., results showed in **Table 3** and **Figure 6**). It demonstrated that CT image based AI scheme was an effective tool to distinguish between non-IA and IA GGNs. Due to the different ways of surgical management for GGNs with different subtypes of lung adenocarcinoma, our AI scheme may have a potential to assist both radiologists and thoracic surgeons in their decision-making.

Despite of the promising results, this study also had several limitations. First, our dataset was small, and only a total of 373 GGNs were involved in this study. The diversity of GGNs in our dataset cannot sufficiently represent the general GGN population in clinical practice. Since the DL model was data-driven, it may be under-fitting due to lack of training dataset. Thus, large diverse dataset and cross-validation method should be used to validate the reproducibility and generalization of our scheme. Due to the different scanning parameters, the tube current, pixel spacing, and slice thickness of CT image was variety. Whether and how these scanning parameters affect the scheme performance have not been investigated in this study (29).

Second, we only extracted and investigated two type CT image features of lung adenocarcinoma namely, DL image feature and radiomics feature, respectively. Although the scheme performance has been improved by fusing two types of imaging features, CT image features cannot decode the whole phenotypes of lung adenocarcinoma tumor. The clinical data, such as smoking history, family history, carcinogenic exposure history, chronic obstructive pulmonary disease, emphysema, interstitial lung disease, etc., may also provide useful classification information. In future studies, we should also apply and combine other types of features (i.e., clinical information, tumor biomarkers, gene feature) to improve the scheme performance (30).

Third, to improve the scheme performance, we only applied a simple information-fusion method to fuse the prediction scores of DL and radiomics based scheme. Due to the limited dataset, our proposed DL scheme and radiomics model may be over-fitting during training process. By applying different weights to the prediction scores of two models, fusion model can weak the over-fitted model's impacts. The over-fitting can be alleviated to some degree by fusing the prediction scores generated by two models. Although the scheme performance has been improved, it may not be the optimal way to combine two types of image features. Thus, we should investigate and develop new fusion methods to fuse the different types of features in future studies. The weak interpretation of DL based scheme is also a limitation of this study. In addition, we used the positions delineated by radiologist to crop GGN patches and generate the training and testing images. The human intervention may also affect the scheme performance.

Last, in our observer study, two radiologists read CT images with time and information constraints, which is different from real clinical situation. The insufficient diagnosis time and clinical information may result in the low performance of two radiologists. Moreover, this is an only technique development study, and we need to conduct rigorous and valid clinical evaluation before applying the proposed scheme into clinical practice.

CONCLUSION

In this study, we developed an AI scheme to classify between non-IA and IA GGNs in CT images. To improve the scheme performance, we fused the prediction scores generated by DL based scheme and radiomics feature based scheme, respectively. The results shows that fusion of DL and radiomics features can significantly improve the scheme performance. Comparing with two radiologists, our new scheme achieves higher performance. It demonstrates (1) fusing DL and radiomics features can improve the classification performance in distinguishing between non-IA and IA, (2) we can build classification DL model with the limited dataset by transferring segmentation task to classification task, (3) AI scheme matches or even outperform radiologists in predicting invasiveness risk of GGNs. Therefore, to improve the diagnosis performance of GGNs, one should focus on exploring and computing robust imaging features, and developing optimal method to fuse different types of features.

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 Municipal Hospital Affiliated to Medical School of Taizhou University and Fudan University Shanghai Cancer Center. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR'S NOTE

In this study, we investigate and develop CT image based artificial intelligence (AI) schemes to predict the invasiveness risk of lung adenocarcinomas, and incorporate deep learning (DL) and radiomics features to improve the prediction performance. The results show that (1) fusing DL and radiomics features can improve the classification performance in distinguishing between non-IA and IA, (2) we can build classification DL model with limited dataset by transferring segmentation task to classification task, (3) AI scheme matches or even outperform radiologists in predicting invasiveness risk of GGNs.

AUTHOR CONTRIBUTIONS

JG and SW designed this study. JG, XX, TY, YL, and SW performed the search and collected data. JG performed data analysis and wrote the manuscript. WH and SW independently diagnosed all the GGNs in testing dataset. All authors reviewed the manuscript.

FUNDING

This work was partially funded by China Postdoctoral Science Foundation under Grant No. 2019M651372, the Shanghai Science and Technology Funds under Grant No. 13411950107, the Zhejiang Provincial Science and Technology Project of Traditional Chinese Medicine under Grant No. 2016ZA205.

REFERENCES

1. Son JY, Lee HY, Kim J-H, Han J, Jeong JY, Lee KS, et al. Quantitative CT analysis of pulmonary ground-glass opacity nodules for distinguishing invasive adenocarcinoma from non-invasive or minimally invasive adenocarcinoma: the added value of using iodine mapping. *Eur Radiol.* (2016) 26:43–54. doi: 10.1007/s00330-015-3816-y
2. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol.* (2011) 6:244–85. doi: 10.1097/JTO.0b013e318206a221
3. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on

CT images: from the fleischner society 2017. *Radiology*. (2017) 284:228–43. doi: 10.1148/radiol.2017161659

4. Hattori A, Hirayama S, Matsunaga T, Hayashi T, Takamochi K, Oh S, et al. Distinct clinicopathologic characteristics and prognosis based on the presence of ground glass opacity component in clinical stage IA lung adenocarcinoma. *J Thorac Oncol*. (2019) 14:265–75. doi: 10.1016/j.jtho.2018.09.026
5. Pedersen JH, Saghir Z, Wille MMW, Thomsen LHH, Skov BG, Ashraf H. Ground-glass opacity lung nodules in the era of lung cancer CT, screening: radiology, pathology, and clinical management. *Oncology*. (2016) 30:266–74. Available online at: https://www.cancernetwork.com/oncology-journal/ground-glass-opacity-lung-nodules-era-lung-cancer-ct-screening-radiology-pathology-and-clinical
6. Fan L, Fang MJ, Bin LZ, Tu WT, Wang SP, Chen WF, et al. Radiomics signature: a biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground-glass nodule. *Eur Radiol*. (2018) 29:1–9. doi: 10.1007/s00330-018-5530-z
7. Ye T, Deng L, Xiang J, Zhang Y, Hu H, Sun Y, et al. Predictors of pathologic tumor invasion and prognosis for ground glass opacity featured lung adenocarcinoma. *Ann Thorac Surg*. (2018) 106:1682–90. doi: 10.1016/j.athoracsur.2018.06.058
8. Kim H, Park CM, Goo JM, Wildberger JE, Kauczor H-U. Quantitative computed tomography imaging biomarkers in the diagnosis and management of lung cancer. *Invest Radiol*. (2015) 50:571–83. doi: 10.1097/RLI.0000000000000152
9. Gong J, Liu J, Hao W, Nie S, Wang S, Peng W. Computer-aided diagnosis of ground-glass opacity pulmonary nodules using radiomic features analysis. *Phys Med Biol*. (2019) 64:135015. doi: 10.1088/1361-6560/ab2757
10. Beig N, Khorrami M, Alilou M, Prasanna P, Braman N, Orooji M, et al. Perinodular and intranodular radiomic features on lung CT images distinguish adenocarcinomas from granulomas. *Radiology*. (2018) 290:180910. doi: 10.1148/radiol.2018180910
11. Li Q, Fan L, Cao ET, Li QC, Gu YF, Liu SY. Quantitative CT analysis of pulmonary pure ground-glass nodule predicts histological invasiveness. *Eur J Radiol*. (2017) 89:67–71. doi: 10.1016/j.ejrad.2017.01.024
12. Chae H-D, Park CM, Park SJ, Lee SM, Kim KG, Goo JM. Computerized texture analysis of persistent part-solid ground-glass nodules: differentiation of preinvasive lesions from invasive pulmonary adenocarcinomas. *Radiology*. (2014) 273:285–93. doi: 10.1148/radiol.14132187
13. Li M, Narayan V, Gill RR, Jagannathan JP, Barile MF, Gao F, et al. Computer-aided diagnosis of ground-glass opacity nodules using open-source software for quantifying tumor heterogeneity. *Am J Roentgenol*. (2017) 209:1216–27. doi: 10.2214/AJR.17.17857
14. Nemec U, Heidinger BH, Anderson KR, Westmore MS, VanderLaan PA, Bankier AA. Software-based risk stratification of pulmonary adenocarcinomas manifesting as pure ground glass nodules on computed tomography. *Eur Radiol*. (2018) 28:235–42. doi: 10.1007/s00330-017-4937-2
15. Zhao W, Yang J, Sun Y, Li C, Wu W, Jin L, et al. 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer Res*. (2018) 78:6881–9. doi: 10.1158/0008-5472.CAN-18-0696
16. Hao P, You K, Feng H, Xu X, Zhang F, Wu F, et al. Lung adenocarcinoma diagnosis in one stage. *Neurocomputing*. (in press). doi: 10.1016/j.neucom.2018.11.110
17. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
18. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. (2019) 25:954–61. doi: 10.1038/s41591-019-0536-x
19. Wang S, Wang R, Zhang S, Li R, Fu Y, Sun X, et al. 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground- glass nodules with diameters ≤3 cm using HRCT. *Quant Imaging Med Surg*. (2018) 8:491–9. doi: 10.21037/qims.2018.06.03
20. Detterbeck FC, Nicholson AG, Franklin WA, Marom EM, Travis WD, Girard N, et al. The IASLC lung cancer staging project: summary of proposals for revisions of the classification of lung cancers with multiple pulmonary sites of involvement in the forthcoming eighth edition of the TNM classification. *J Thorac Oncol*. (2016) 11:639–50. doi: 10.1016/j.jtho.2016.01.024
21. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. *Recurrent Residual Convolutional Neural Network Based on U-Net. (R2U-Net) for Medical Image Segmentation*. (2018) Available online at: http://arxiv.org/abs/1802.06955 doi: 10.1109/NAECON.2018.8556686
22. Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys*. (2011) 38:915–31. doi: 10.1118/1.3528204
23. Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, et al. Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. *Med Image Anal*. (2017) 40:172–83. doi: 10.1016/j.media.2017.06.014
24. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection. *ACM Comput Surv*. (2017) 50:1–45. doi: 10.1145/3136625
25. Gong J, Liu J, Jiang Y, Sun X, Zheng B, Nie S. Fusion of quantitative imaging features and serum biomarkers to improve performance of computer-aided diagnosis scheme for lung cancer: a preliminary study. *Med Phys*. (2018) 45:5472–81. doi: 10.1002/mp.13237
26. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. (2017) 77:e104–7. doi: 10.1158/0008-5472.CAN-17-0339
27. Ben-David A. About the relationship between ROC curves and Cohen's kappa. *Eng Appl Artif Intell*. (2008) 21:874–82. doi: 10.1016/j.engappai.2007.09.009
28. Gong J, Liu J, Hao W, Nie S, Zheng B, Wang S, et al. A deep residual learning network for predicting lung adenocarcinoma manifesting as ground-glass nodule on CT images. *Eur Radiol*. (2020) 30:1847–55. doi: 10.1007/s00330-019-06533-w
29. Zhang Y, Tang J, Xu J, Cheng J, Wu H. Analysis of pulmonary pure ground-glass nodule in enhanced dual energy CT imaging for predicting invasive adenocarcinoma: comparing with conventional thin-section CT imaging. *J Thorac Dis*. (2017) 9:4967–78. doi: 10.21037/jtd.2017.11.04
30. Hirsch FR, Franklin WA, Gazdar AF, Bunn PA. Early detection of lung cancer: clinical perspectives of recent advances in biology and radiology. *Clin Cancer Res*. (2001) 7:5–22. Available online at: https://clincancerres.aacrjournals.org/content/7/1/5

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.

frontiers in Oncology

ORIGINAL RESEARCH
published: 20 January 2021
doi: 10.3389/fonc.2020.586292

SuperHistopath: A Deep Learning Pipeline for Mapping Tumor Heterogeneity on Low-Resolution Whole-Slide Digital Histopathology Images

Konstantinos Zormpas-Petridis[1]*, *Rosa Noguera*[2,3], *Daniela Kolarevic Ivankovic*[4], *Ioannis Roxanis*[5]†, *Yann Jamin*[1]† and *Yinyin Yuan*[6]*†

[1] Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, United Kingdom, [2] Department of Pathology, Medical School, University of Valencia-INCLIVA Biomedical Health Research Institute, Valencia, Spain, [3] Low Prevalence Tumors, Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain, [4] The Royal Marsden NHS Foundation Trust, London, United Kingdom, [5] Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, United Kingdom, [6] Division of Molecular Pathology, The Institute of Cancer Research, London, United Kingdom

Edited by:
Youyong Kong,
Southeast University, China

Reviewed by:
Li Liu,
Donghua University, China
Mark Hansen,
University of the West of England,
United Kingdom
Eleftheria Panagiotaki,
University College London,
United Kingdom

***Correspondence:**
Konstantinos Zormpas-Petridis
Konstantinos.Zormpas-Petridis@icr.ac.uk
Yinyin Yuan
yinyin.yuan@icr.ac.uk

†These authors share senior authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: 22 July 2020
Accepted: 30 November 2020
Published: 20 January 2021

Citation:
Zormpas-Petridis K, Noguera R, Ivankovic DK, Roxanis I, Jamin Y and Yuan Y (2021) SuperHistopath: A Deep Learning Pipeline for Mapping Tumor Heterogeneity on Low-Resolution Whole-Slide Digital Histopathology Images. Front. Oncol. 10:586292. doi: 10.3389/fonc.2020.586292

High computational cost associated with digital pathology image analysis approaches is a challenge towards their translation in routine pathology clinic. Here, we propose a computationally efficient framework (SuperHistopath), designed to map global context features reflecting the rich tumor morphological heterogeneity. SuperHistopath efficiently combines i) a segmentation approach using the linear iterative clustering (SLIC) superpixels algorithm applied directly on the whole-slide images at low resolution (5x magnification) to adhere to region boundaries and form homogeneous spatial units at tissue-level, followed by ii) classification of superpixels using a convolution neural network (CNN). To demonstrate how versatile SuperHistopath was in accomplishing histopathology tasks, we classified tumor tissue, stroma, necrosis, lymphocytes clusters, differentiating regions, fat, hemorrhage and normal tissue, in 127 melanomas, 23 triple-negative breast cancers, and 73 samples from transgenic mouse models of high-risk childhood neuroblastoma with high accuracy (98.8%, 93.1% and 98.3% respectively). Furthermore, SuperHistopath enabled discovery of significant differences in tumor phenotype of neuroblastoma mouse models emulating genomic variants of high-risk disease, and stratification of melanoma patients (high ratio of lymphocyte-to-tumor superpixels ($p = 0.015$) and low stroma-to-tumor ratio ($p = 0.028$) were associated with a favorable prognosis). Finally, SuperHistopath is efficient for annotation of ground-truth datasets (as there is no need of boundary delineation), training and application (~5 min for classifying a whole-slide image and as low as ~30 min for network training). These attributes make SuperHistopath particularly attractive for research in rich datasets and could also facilitate its adoption in the clinic to accelerate pathologist workflow with the quantification of phenotypes, predictive/prognosis markers.

Keywords: deep learning, machine learning, digital pathology, computational pathology, tumor region classification, melanoma, neuroblastoma, breast cancer

INTRODUCTION

The analysis of histopathological images of surgical tissue specimens stained with hematoxylin and eosin (H&E) remains a critical decision-making tool used for the routine management of patients with cancer and the evaluation of new therapeutic strategies in clinical trials (1–3). In several precision medicine settings, there is an increasing demand for accurate quantification of histological features. However, in their diagnostic practice, pathologists exercise a predominantly qualitative or semi-quantitative assessment with an inherent degree of inter- and intra-observer variability, which occasionally hampers their consistency (4–7). In the new era of digital pathology, advanced computational image analysis techniques are revolutionizing the field of histopathology by providing objective, robust and reproducible quantification of tumor components, thereby assisting pathologists in tasks such as tumor identification and tumor grading (8, 9). Histopathological image analysis can now be performed in high-resolution H&E-stained whole-slide images (WSI) using state-of-the-art deep learning and classical machine learning approaches for single cell segmentation and/or classification. The new ability to map the spatial context of each single cell also opened new avenues for the study of the tumor micro-environment (10–16), which is key to guide the delivery of precision medicine including immunotherapy.

However, computational pathology is still not widely adopted in the oncological setting. One of the challenges lies in the gigabyte sizes of high-resolution WSIs, which result in computationally expensive approaches. WSIs need to be divided into images patches (typical size: 256x256) before being processed by deep networks such as convolutional neural networks (CNNs) (17). Secondly, single-cell approaches provide markers that are often hard-to-be-evaluated or even interpreted by the pathologists and can be prone to the generalization errors when applied in new unseen dataset. As a result, many promising markers eventually fail to reach the clinic due to a lack of cross-validation in new independent datasets. On the other hand, tissue classification approaches, which target multicellular assemblies and paucicellular areas where individual cells are incorporated into the region segmentation, would be accessible for visual validation by pathologists. Such algorithms would enable the characterization of the distribution and interrelationship of global features that are currently detectable by human perception but not quantifiable without artificial intelligence- (AI-)assisted numerical expression.

Current computed pathology tools primarily focus on individual cell analysis at high-resolution (40x/20x magnification) with limited local context features, whereas pathologists frequently employ collateral information, taking into account the overall tissue microarchitecture. Many established clinical markers are actually identified at low or intermediate magnifications, including tumor architecture-based grading systems (18, 19), stroma-tumor ratio (20, 21), infiltrating lymphocytes (TILs) (22, 23) and necrosis (24–26). This has not been yet fully emulated by computational pathology methodologies. However, some methods for the classification of tissue components have been suggested either using image patch classification typically with a CNN or pixel-level classification/segmentation typically with a U-Net-like architecture (27), mainly for tasks such as the dichotomized classification of tissue (e.g. cancerous vs non-cancerous) (28, 29), the segmentation of a feature of interest (e.g. glands) (16, 30) or multi-type tissue classification (9, 31–35). For segmentation purposes, U-Net-like architectures are usually preferred over CNNs, which have established limitations in conforming to object contours. Yet, CNNs have also resulted in promising segmentation approaches (36–38) with the enhanced capability of classifying a large number of categories (39). Multi-scale approaches incorporating information from various image resolutions have also been proposed (40–43). Different approaches have been explored for the classification of epithelium or stroma using superpixels-based segmentation of image patches with either hand-crafted or deep learning features (44, 45). Bejnordi and colleagues used a similar method for their multi-scale approach for the classification of tissue or non-tissue components on low resolution images and stroma and background regions from intermediate and high resolution images (46). However, these methods are typically performed on high-magnifications image patches (20-40x and more rarely 10x) and are associated with a high computational cost.

Here, we propose a framework (SuperHistopath), which can map most of the global context features that contribute to the rich tumor morphological heterogeneity visible to pathologists at low resolution and used for clinical decision making in a computationally efficient manner. We first apply the well-established simple linear iterative clustering (SLIC) superpixels algorithm (47) directly on the WSI at low resolution (5x magnification) and subsequently classify the superpixels into different tumor region categories using a CNN based on pathologists' annotations. SuperHistopath particularly capitalizes on:

i. the use of superpixels which provide visually homogeneous areas of similar size respecting the region boundaries and avoid the potential degradation of classification performance associated with image patches, (no matter how small) spanning over multiple tissue categories.
ii. the use of CNN necessary to accurately classify and map the multiple tissue categories that constitute the rich and complex histological intratumoral heterogeneity.
iii. the computational efficiency, faster processing speed and lower memory requirements associated with processing the WSI at low resolution.

We applied SuperHistopath to H&E-stained images from three different cancer types: clinical cutaneous melanoma, triple-negative breast cancer and tumors arising in genetically-engineered mouse models of high-risk childhood neuroblastoma.

MATERIALS AND METHODS

Datasets

All digitized whole-slide images (WSI) used in this study were H&E-stained, formalin-fixed and paraffin-embedded (FFPE) sections, and scaled to 5x magnification as presented in **Table 1** (image sizes at 5x varied from ~8000x8000 to

TABLE 1 | Summary of the datasets used.

Cancer type	Number of WSIs	Digital scanner	Pixel resolution (5x magnification)	Dataset
Cutaneous melanoma	127	Aperio ImageScope	2.016 µm	The Cancer Genome Atlas (TCGA)
Triple-negative breast cancer	23	NanoZoomer XR	2.3 µm	Internal dataset, Collaboration with The Serbian Institute of Oncology
High-risk neuroblastoma (mouse models)	73	NanoZoomer XR	2.3 µm	Internal dataset Tumors samples coming from established Th-*MYCN* and Th-*ALK*^{F1174L}/*MYCN* transgenic mouse colonies (48, 49) and processed by a clinical histopathological core facility

~12000x12000 pixels). We applied our framework to clinical patient samples of cutaneous melanoma and triple-negative breast cancer, in addition to tumor samples from transgenic mouse models of childhood neuroblastoma. Both the Th-*MYCN* and Th-*ALK*^{F1174L}/*MYCN* mouse models have been shown to spontaneously develop abdominal tumors, which mirror the major histopathological characteristics of childhood high-risk disease (50, 51).

Region Classification

First, each dataset was pre-processed using the Reinhard stain normalization (52) to account for stain variabilities that could affect classification. Then, all images were segmented using the simple linear iterative clustering (SLIC) superpixels algorithm, which groups together similar neighboring pixels. With our pathologist's input, we selected the optimal number of superpixels by visually identifying a superpixel size that capture only homogeneous areas and adhere to image boundaries. This is a critical step for ensuring accurate tissue segmentation, and therefore, classification (**Figure 1**). The number of superpixels was adapted for each image to ensure a homogenous superpixel size across the datasets and was automatically set based on the image size according to *Equation 1* (53).

$$N_i = ceiling\left(\frac{S_i}{U}\right) \tag{1}$$

where *Ni* is the number of superpixels in the i^{th} image, *Si* is the size of image *i* in pixels, and *U* is a constant held across all images that defined the desired superpixels size.

The SLIC algorithm inherently provides a roughly uniform superpixel size. Setting $U = 1500$, *Equation 1* gave a mean superpixels size of 51 × 51 pixels, equivalent to an area of approximately 117 × 117 µm^2. Bilinear interpolation was subsequently use to resize each superpixel to a fixed size of 56 x 56 or 75 x 75 pixels (the minimum input size for inception-like network architectures).

Region annotations were provided by a senior pathologist with over 20 years of experience for the melanoma and breast cancer clinical datasets, and a senior pediatric neuropathologist with over 20 years of experience for the neuroblastoma mouse datasets. For training and testing, superpixels were assigned to each category based on their isocenter location within the annotated regions. Note that region annotations for our algorithm do not need to delineate boundaries as illustrated in **Figure 1B**.

The numbers of clinically relevant tissue categories, number of WSIs and superpixels used for training and testing are summarized for each tumor types in **Table 2**. Standard image

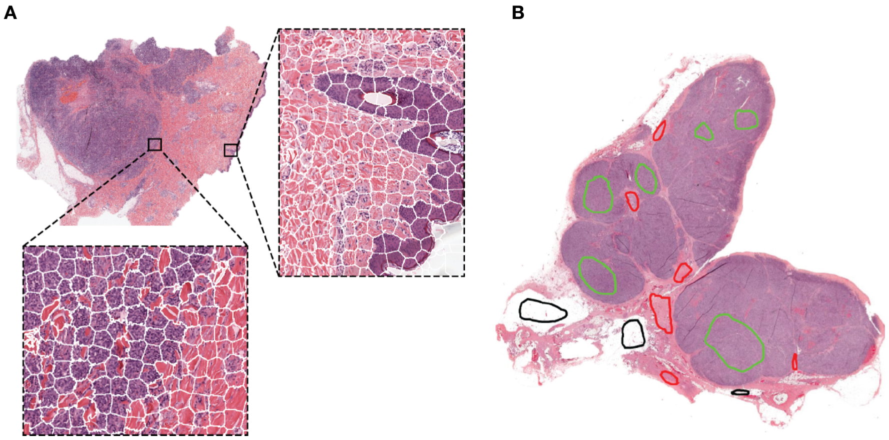


FIGURE 1 | Representative examples of the SLIC superpixels segmentation and ground-truth annotations in TCGA melanoma samples **(A)** Whole-slide image segmentation using the SLIC superpixels algorithm. Note how the superpixels adhere to the boundaries of the different components of the tumor with each superpixel containing a single type of tissue **(B)** Ground-truth annotations are provided by the pathologists by marking samples of the region components (the different colors represent different regions) without the need for delineating the boundaries of the tumor components.

augmentations, such as rotations (90°, -90°, 180°), flips (horizontal and vertical), and contrast (histogram equalization) were performed in each case to capture more variation and even out the training dataset imbalances.

Training of the Convolutional Neural Networks

Our custom-designed CNN for superpixel classification consists of 6 convolutional layers (32, 32, 64, 64, 128, 128 neurons, respectively) of 3 x 3 filter size and 3 max-pooling layers, followed by a "flatten" layer and a dense layer of 256 neurons (**Figure 2**). A superpixel RGB image (post-interpolation) was used as input into the network and normalized from range 0–255 to range 0–1 using the maximum value. The output of the network was a label assigned to each superpixel based on which region category it belonged to. After empirical experimentation, a ReLU activation function was used in all layers except for the last layer where standard softmax was used for classification. The weights incident to each hidden unit were constrained to have a norm value less than or equal to 3 and a dropout unit of 0.2 was used before every max-pooling operation to

TABLE 2 | Summary of the datasets used for training and testing the convolutional neural network.

Cancer type	Number of WSIs used for network training		Regional classification	
Cutaneous melanoma	Total	27	***6 categories***	***Superpixels for training***
	Training	22	Tumor tissue	21940
	Testing	5	Stroma	12419
			Normal epidermis	1646
			Lymphocytes cluster	2367
			Fat	15484
			Empty/white space	3412
Triple-negative breast cancer	Total	23	***6 categories***	***Superpixels for training***
	Training	18	Tumor tissue	18873
	Testing	5	Stroma	24220
			Necrosis	15102
			Lymphocytes cluster	3472
			Fat	10044
			Empty/white space	16473
High-risk neuroblastoma (mouse model)	Total	60	***8 categories***	***Superpixels for training***
	Training	44	Region of undifferentiated neuroblasts	20512
	Testing	16	Tissue damage (necrosis/apoptosis)	17645
			Differentiation region	5740
			Lymphocytes cluster	4009
			Hemorrhage (blood)	6124
			Muscle	6415
			Kidney	14976
			Empty/white space	21470

Note that the testing datasets consisted of whole-slide images from different patients from the training dataset.

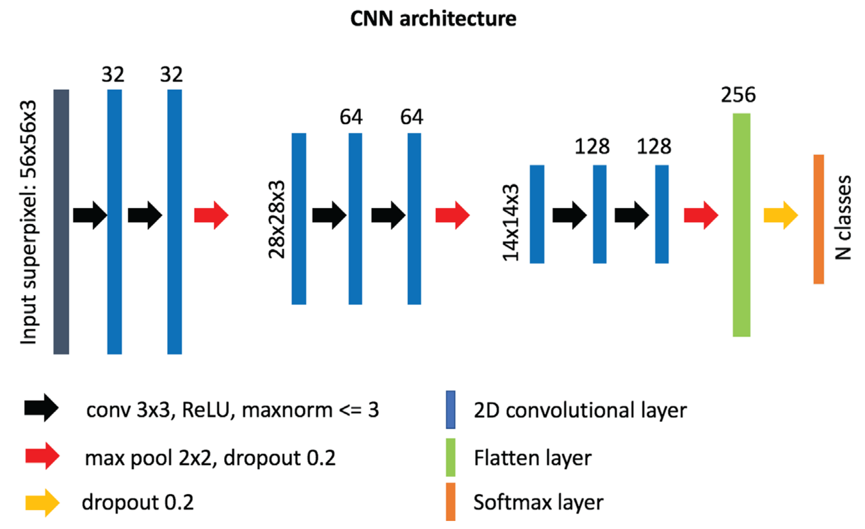


FIGURE 2 | Architecture of our custom-designed convolutional neural network for the classification of superpixels into different tissue-level categories.

avoid overfitting (54). The weights of the layers were randomly initialized using "Glorot uniform" initialization (55), and the network was optimized using the Adam method (56) with a learning rate of 10^{-3} and a categorical cross-entropy cost function. The number of trainable parameters for our custom-made network is ~1.9 M. The network was implemented in python (v. 3.6.5) using the Keras/Tensorflow libraries (v. 2.2.4/1.12.0, respectively).

To choose the best network for our framework, we tested other known neural network architectures as implemented in the Keras framework, including InceptionV3 (57), Xception (58), InceptionResNetV2 (59), and ResNet (60). We initialized the weights using the pre-trained ImageNet weights. To optimize each network, we excluded the final classification layer, and added three additional layers, i) a global average pooling layer, ii) a dense layer of 256 neurons with ReLu activation, constrained to have a norm value less than or equal to 3, and iii) a dense layer tailored to the number of classes of each cancer type using the softmax function for classification.

For inception-like architectures (Inception v3, Inception ResNetV2, Xception) only superpixels of size 75 x 75 were used. We trained all the networks for 50 epochs using batch sizes of 150 and 256 for superpixels of sizes 75 x 75 and 56 x 56, respectively, and kept the models with the highest validation accuracy.

The Xception and custom-made networks were re-trained from the beginning for each cancer type, without applying any further changes.

Application of SuperHistopath for the Quantification of Clinical Features of Interest

In the melanoma dataset, we calculated the number of pixels belonging to each classified category. For each patient we derived i) the ratio of pixels classified as stroma region to all pixels in tumor compartments, and ii) the ratio of pixels classified as clusters of lymphocytes to all pixels in tumor compartments; we evaluated the prognostic value of these quantitative indices using survival analysis. Patients were divided into high- and low-risk groups based on split at the median value of all scores to ensure both groups were of similar size. Kaplan-Meier estimation was used to compare overall survival in the 127 patients. Differences between survival estimates were assessed with the log-rank test and hazard ratios were calculated using Cox's proportional-hazard regression.

In the neuroblastoma dataset, we evaluated the differences in phenotype between the Th-*ALK*^{F1174L}/*MYCN* (n=7) and Th-*MYCN* tumors (n=6) by quantifying the proportion of pixels classified by our SuperHistopath as regions rich in undifferentiated neuroblasts, differentiating neuroblasts, tissue damage (necrosis/apoptosis) hemorrhage and clusters of lymphocytes. Note that i) we did not quantify stroma in these tumors as they faithfully mirror the stroma-poor phenotype which define high-risk disease ii) lymphocytes clusters universally correspond to encapsulation of lymph node by the tumor, rather that tumor infiltrates, consistent with the "cold" immune phenotype of high-risk disease. We focus on identifying any significant difference in the ratio of differentiation or the ratio of hemorrhagic regions to all tumor compartments between the two tumor types using the Mann-Whitney U test, with a 5% level of significance.

RESULTS

SuperHistopath Can Accurately Map the Complex Histological Heterogeneity of Tumors

Melanoma

We first developed and evaluated our framework on the H&E-stained, FFPE sections of clinical specimen of cutaneous melanoma scaled to 5x magnification. **Figure 1** shows the results of the segmentation using the simple linear iterative clustering (SLIC) superpixels algorithm, which groups together similar neighboring pixels.

The optimized Xception network achieved the highest score and classified the melanoma sample regions into 6 predefined tissue categories of interest: tumor tissue, stroma, cluster of lymphocytes, normal epidermis, fat, and empty/white space with an overall accuracy of 98.8%, an average precision of 96.9%, and an average recall of 98.5% over 14,092 superpixels in a separate test set of five images (**Tables 3**, **4**). Our custom CNN also achieved comparable performance to the state-of-the-art networks with an overall accuracy of 96.7%, an average precision of 93.6%, and an average recall of 93.6% (**Figure 2**, **Supplementary Table 1**). The confusion matrices for the XCeption and our custom CNN networks are presented in **Table 4** and **Supplementary Table 1**, respectively. **Figure 3** shows qualitative results of our approach's regional classification in representative melanoma WSIs using the optimized Xception network.

Breast Cancer

SuperHistopath classified sample regions into 6 predefined tissue categories of interest: tumor, necrosis, stroma, cluster of

TABLE 3 | Evaluation metrics of the different neural network architectures in the TCGA melanoma test dataset.

Network	Accuracy (%)	Precision (%)	Recall (%)	Parameters (in millions)
InceptionV3	97.5	94.2	96.7	~22.4
InceptionResNetV2	97.7	94.1	97.3	~54.8
ResNet50	93.8	92.2	88.9	~24.2
Xception	**98.8**	**96.9**	**98.5**	~21.4
Our custom-made CNN	96.7	93.6	93.6	**~1.9**

The bold values in the Accuracy (%), Precision (%) and Recall (%) fields indicate the highest value i.e. the best performance achieved amongst the networks under comparison. The bold value in the Parameters (in millions) field indicate the network with the fewer parameters used amongst the networks under comparison.

TABLE 4 | Confusion matrix of the classification of superpixels using the optimized Xception network in melanoma patients in 6 categories: tumor, stroma, normal epidermis, cluster of lymphocytes (Lym), fat and empty/white space (separate test set of 5 whole-slide images).

	Tumor	Stroma	Epidermis	Lym	Fat	Empty space
Tumor	**5286**	10	7	8	0	0
Stroma	9	**986**	0	0	2	0
Epidermis	22	0	**545**	0	1	0
Lym	0	0	1	**821**	0	0
Fat	0	9	0	0	**5603**	3
Empty space	0	0	0	0	98	**681**

Overall accuracy = 98.8%, average precision = 96.9%, average recall = 98.5%.
The bold values indicate the correct predictions of the network.

lymphocytes, fat, and lumen/empty space with an overall accuracy of 93.1%, an average precision of 93.9%, and an average recall of 93.6% using Xception and 91.7%, 92.5%, 91.8% respectively using our custom-made CNN over 10,349 superpixels in the independent test set of five images. The confusion matrices for the XCeption and our custom CNN networks are presented in **Table 5** and **Supplementary Table 2**, respectively. **Figure 4** shows qualitative results our approach's regional classification in representative triple-negative breast cancer WSIs.

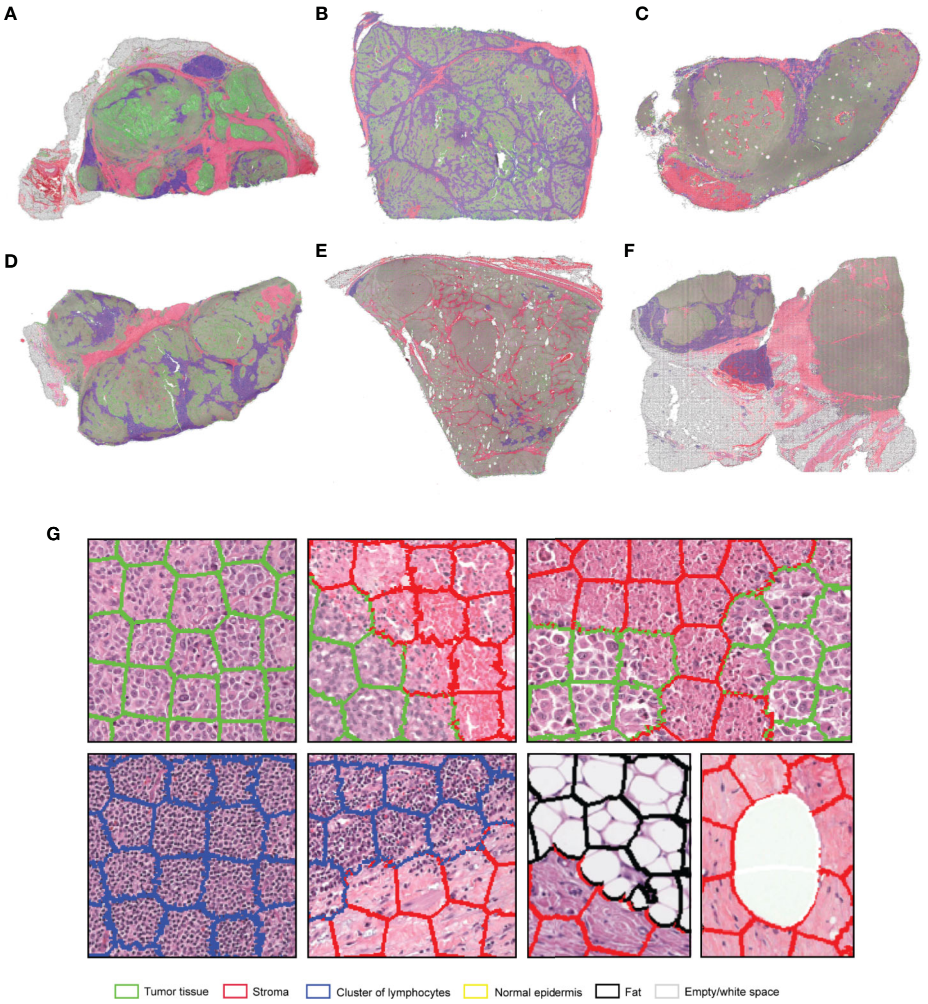


FIGURE 3 | **(A–F)** Representative examples of the results obtained from the application of the SuperHistopath pipeline in whole-slide images of tumors (5x) of the Cancer Genome Atlas (TCGA) melanoma dataset [**(G)** Magnified regions of interest]. Note the important clinically-relevant phenotypes characterized by clusters of lymphocytes infiltrating the tumor in samples **(B, D)**. or the majority of clusters of lymphocytes residing just outside the tumor area (left and central part) with only a few clusters infiltrating the tumor (right part) in sample **(C)**.

TABLE 5 | Confusion matrix of the classification of superpixels using the optimized Xception network in triple-negative breast cancer patients in six categories: tumor, necrosis, cluster of lymphocytes (Lym), stroma, fat, and lumen/empty space (separate test set of five whole-slide images).

	Tumor	Necrosis	Lym	Stroma	Fat	Empty space
Tumor	**1830**	13	15	42	0	0
Necrosis	50	**1446**	2	320	0	0
Lym	4	2	**705**	10	0	0
Stroma	42	120	20	**3836**	0	1
Fat	0	0	0	0	**562**	5
Empty space	0	0	0	0	67	**1257**

Overall accuracy = 93.1%, average precision = 93.9%, average recall = 93.6%.
The bold values indicate the correct predictions of the network.

Neuroblastoma

SuperHistopath classified the tumor regions into eight predefined tissue categories of interest: undifferentiated neuroblasts, tissue damage (necrosis/apoptosis), areas of differentiation, cluster of lymphocytes, hemorrhage, muscle, kidney, and empty/white space with an overall accuracy of 98.3%, an average precision of 98.5%, and an average recall of 98.4% using Xception and 96.8%, 97.1%, 97.2% respectively using our custom-made CNN over 9,868 superpixels in the independent test set of 16 images. The confusion matrices for the XCeption and our custom CNN networks are presented in **Table 6** and **Supplementary Table 3**, respectively. **Figure 5** shows qualitative results of our approach's regional classification in representative WSIs of neuroblastoma arising in the Th-*MYCN* mouse model.

SuperHistopath Pipeline for the Analysis of Low-Resolution WSI Affords Significant Speed Advantages

The average time for the SLIC superpixels algorithm to segment a WSI in 5x magnification was < 2 min using a 3.5 GHz Intel core i7 processor. The average time for both the Xception and our custom-made CNN network to classify every superpixel in the

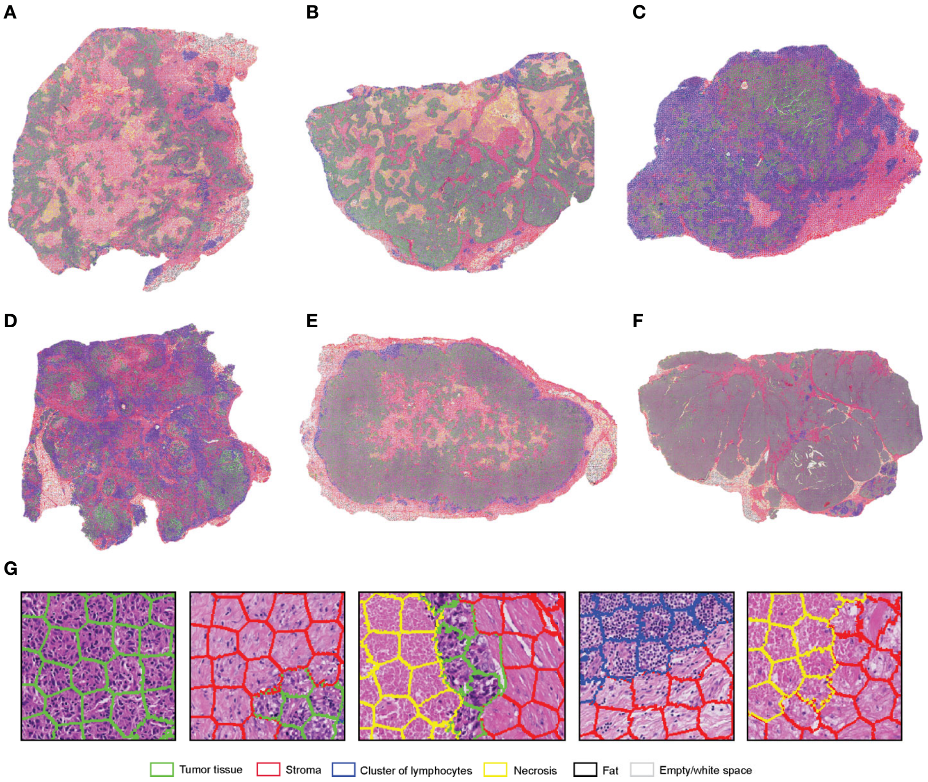


FIGURE 4 | (A–F). Representative examples of the results obtained from the application of the SuperHistopath pipeline in whole-slide images of tumors (5x) of the triple-negative breast cancer **(G)** Magnified regions of interest. Note the important clinically-relevant features, such as the amount of tumor necrosis inside tumors **(A)** and **(B)**, lymphocytes which, are infiltrating the tumor in large number in samples **(C, D)**, but are surrounding the stroma barrier without infiltrating the tumor in samples **(A, B, E, F)**.

TABLE 6 | Confusion matrix of the classification of superpixels using the optimized Xception network in the Th-*MYCN* and *Th*-ALK^{F1174L}/MYCN mouse models in eight categories: region of undifferentiated neuroblasts, necrosis, cluster of lymphocytes (Lym), hemorrhage (blood), empty/white space, muscle tissue and kidney (separate test set of 16 whole-slide images).

	Undifferentiated region	Necrosis	Lym	Differentiation	Blood	Empty space	Muscle	Kidney
Undifferentiated region	**1403**	3	0	14	1	0	0	0
Necrosis	13	**1642**	1	26	49	2	5	18
Lym	6	5	**1150**	0	0	0	0	3
Differentiation	0	0	0	**1261**	0	0	0	0
Blood	1	7	0	0	**1327**	0	9	0
Empty space	0	2	0	0	0	**560**	3	2
Muscle	0	2	0	0	1	0	**1176**	0
Kidney	0	0	0	0	0	0	0	**1176**

Overall accuracy = 98.3%, average precision = 98.5%, average recall = 98.4%.
The bold values indicate the correct predictions of the network.

images was 1–2 min using the same processor. A quick convergence of the networks (around epoch 30) was observed in all cases, which needed ~3 h for Xception and only ~30 min for our custom-made CNN using a Tesla P100-PCIE-16GB GPU card, and therefore the latter was used for experimenting.

SuperHistopath Can Provide Robust Quantification of Clinically Relevant Features

Stroma-to-Tumor Ratio and Clusters of Lymphocytes Abundance as Predictive Markers of Survival in Melanoma

We first use SuperHistopath to quantify both the stroma-to-tumor ratio and the immune infiltrate, which have both shown to provide prognostic and predictive information in patient with solid tumors, including melanoma (20, 21, 23). The important role of immune hotspots has been established based on density analysis of single cell classification of lymphocytes in high-resolution images (61, 62). Here, we demonstrate in our melanoma dataset of 127 WSIs i) that a high stromal ratio as identified in low resolution WSIs is a predictor of poor prognosis (SuperHistopath: p = 0.028, Coxph-Regression [discretized by median]: HR = 2.1, p = 0.0315; **Figure 6A**) and ii) that clusters of lymphocytes hold predictive information in our melanoma dataset, with a high lymphocyte ratio being an indicator of favorable prognosis [SuperHistopath: p = 0.015, Coxph-Regression (discretized by median): HR = 0.4, p = 0.018; **Figure 6B**]. Pearson's correlation showed no significant

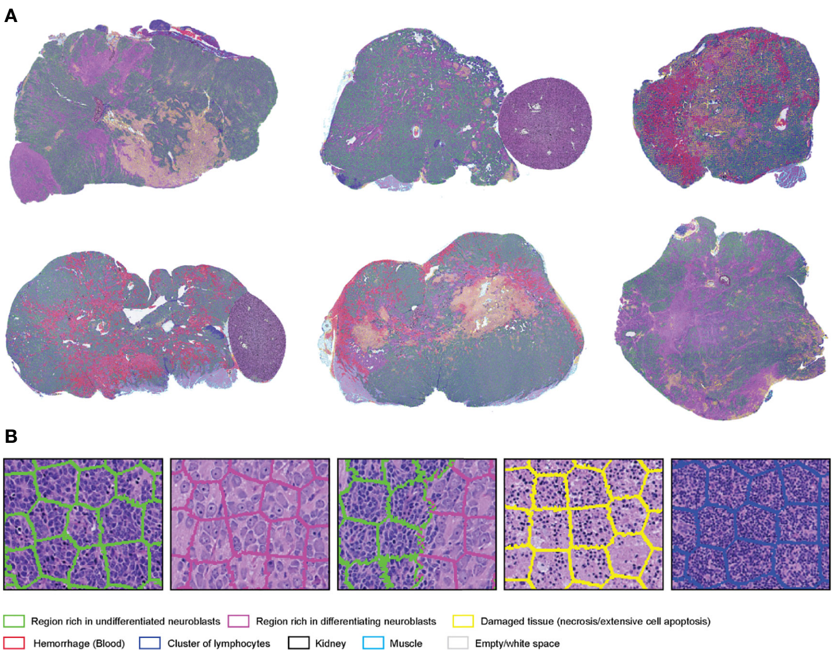

A
B
Region rich in undifferentiated neuroblasts
Region rich in differentiating neuroblasts
Damaged tissue (necrosis/extensive cell apoptosis)
Hemorrhage (Blood)
Cluster of lymphocytes
Kidney
Muscle
Empty/white space

FIGURE 5 | **(A)** Representative examples of the results obtained from the application of the SuperHistopath pipeline in whole-slide images of tumors (5x) arising in genetically-engineered mouse models of high-risk neuroblastoma [**(B)** Magnified region of interest].

correlation between stromal ratio and clusters of lymphocytes ratio (r = -0.13, p = 0.13), and between absolute sizes of stroma and clusters of lymphocytes (r = 0.13, p=0.11). Taken together, our data, captured from low resolution (5x) WSIs, are consistent with those extracted from single-cell analysis in high-resolution WSIs (53).

Necrosis Quantification

We use the SuperHistopath to quantity tumor necrosis in our breast cancer and childhood neuroblastoma preclinical datasets. Tumor necrosis, defined as confluent cell death or large area of tissue damage hold predictive and prognostic information, both at diagnosis and after chemotherapy, in many solid tumors including breast cancer and childhood malignancies (24–26, 63, 64). While visible at 5x objective lens magnification, its quantification can often be a challenging task even for experienced pathologists. Here, we show that SuperHistopath can provide satisfactory quantification of necrosis in clinical breast cancer samples by distinguishing from stroma with high specificity (91.5%) and satisfactory precision (79.5%) and in the high-risk neuroblastoma mouse models with high precision and specificity (93.5% and 98.9% respectively).

Quantification of Neuroblastoma Differentiation

We used SuperHistopath to quantify the phenotype of MYCN-driven transgenic mouse models of high-risk stroma-poor neuroblastoma. We show that SuperHistopath can identify areas of differentiation, a critical feature for the stratification of children neuroblastoma, with both high precision and specificity (100% and 96.9% respectively). SuperHistopath also showed that expression of *ALK*^{F1174L} mutation significantly shift the MYCN-driven phenotype from poorly-differentiated and hemorrhagic phenotype (Th-*MYCN*: 1.8 ± 1.3% differentiating area and 29.2 ± 6.7% hemorrhage, **Figure 6C**) into a differentiating phenotype also characterized by the almost complete abrogation of the hemorrhagic phenotype (Th-*ALK*^{F1174L}/*MYCN*: 20.3 ± 3.1% differentiating area and 0.2 ± 0.1% hemorrhage, p=0.0003 and p=0.0008 respectively, **Figure 6D**) as previously demonstrated (51, 65).

DISCUSSION

In this study, we implemented SuperHistopath: a digital pathology pipeline for the classification of tumor regions and the mapping of tumor heterogeneity from low-resolution H&E-stained WSIs, which we demonstrated to be highly accurate in three types of cancer. Combining the application of the SLIC superpixels algorithm directly on low magnification WSIs (5x) with a CNN architecture for the classification of superpixels, contributes to SuperHistopath computational efficiency allowing for fast processing, whilst affording the quantification of robust and easily interpretable clinically-relevant markers.

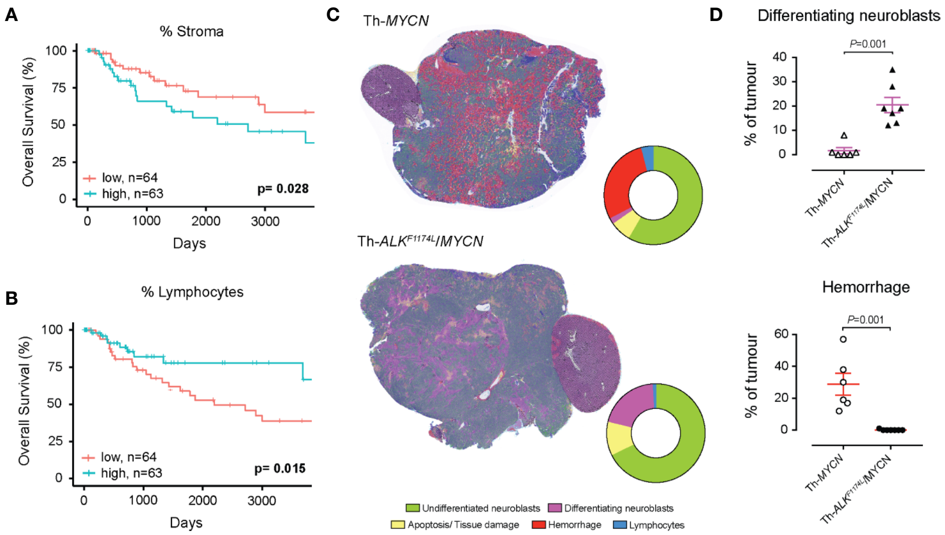


FIGURE 6 | Quantification of clinically relevant features with SuperHistopath. **(A, B)** show associations between survival outcomes and SuperHistopath-defined risk groups in the Cancer Genome Atlas (TCGA) cohorts of patients with melanoma. **(A)** Kaplan-Meier Survival curves for patients in the high-risk group (blue) and low risk group (red) classified by stromal cells ratio derived from SuperHistopath and **(B)** Kaplan-Meier Survival curves for patients in the high-risk group (blue) and low risk group (red) classified by immune infiltrate based on lymphocytes cluster ratio derived from SuperHistopath. **(C, D)** show the SuperHistopath-based quantification of tumor phenotype in genetically-engineered mouse model of high-risk neuroblastoma. **(C)** Representative SuperHistopath-segmented whole-slide images (5x) and pie chart showing the Super-CNN quantified mean composition of the tumors arising in Th-MYCN (n=6) and Th-ALK^{F1174L}/MYCN (n=7) mouse models of high-risk neuroblastoma. Note the marked difference of phenotype induced by the expression of the ALK^{F1174L} mutation characterized by **(D)** a significantly increased neuroblastoma differentiation neuroblasts and the total abrogation of the characteristic hemorrhagic phenotype of Th-MYCN tumors.

Applying our computational approach on low-resolution images leads to markedly increased processing speed, for both the classification of new samples and network training. Here, we chose the (5x) magnification as a compromise between tumor structures visibility and computational cost. Specific metrics such as stroma-to-tumor ratio could potentially be derived from images at even lower magnifications (e.g. 1.25x) as recently shown (53). Digital histology images are conventionally processed at 40x (or 20x) magnification where cell morphology is most visible. At those resolutions, WSIs are large (representative size at 20x: 60000 x 60000 pixels), requiring of a lot of memory and images to be divided into patches (tiles) for processing. Under these conditions, the training of new networks for cell segmentation and classification typically requires days and the application to new WSI samples can take hours prior to code optimization. In contrast, the training of our neural network until acceptable convergence needed as little as ~30 min and application on new samples ~5 min (for both superpixel segmentation and classification) in our study. High-resolution images are essential when studying cell-to-cell interactions, however we show that the processing of low resolution images is appropriate for the extraction of specific global context features.

Furthermore, SuperHistopath combines the main advantages of regional classification and segmentation approaches. On one hand, classification approaches applied on smaller patches resulting from splitting WSIs allow the use of CNN for the robust classification of many categories necessary to capture intratumor heterogeneity (39), yet at the expense of higher risk of misclassification, especially close to regional boundaries where an image patch, regardless of its size, may contain multiple tumor components. Overlapping (sliding) window approaches can improve the issue, yet at an increased computational cost. On the other hand, segmentation approaches such as U-Net-like architectures can resolve the regional boundaries issue but appear to work better for few classes, typically two. SuperHistopath efficiently combines the use of a segmentation approach using superpixels to adhere to region boundaries with CNN classification to cover the rich tumor histological heterogeneity (here 6-8 region categories depending on the cancer type).

Our method also markedly simplifies and accelerates the process of preparing ground-truth (annotations) datasets as *i)* the use of superpixels alleviate the need for careful boundary delineation of the tumor components of interest (**Figure 1B**), a cumbersome and time-consuming process necessary for using U-Net-like architectures and ii) each annotated region contains large numbers of superpixels facilitating the collection of the large datasets traditionally required by deep learning methods.

The appropriate choice of superpixel size is crucial to warrant both accurate tissue segmentation and classification. Equation 1 ensured a uniform superpixel size for every whole-slide image regardless of their original size. The main considerations for choosing superpixels size (i.e. setting the constant U) is to ensure that they only contain a single tissue type, while being large enough to contain sufficient tissue information. In our study, we found that classification is not sensitive to small changes of U. However larger superpixels ($U > 1750$) did not adhere well to the tissue boundaries, whereas smaller superpixels ($U < 1250$) indeed led to a slight decrease in classification performance.

Many promising computational pathology-derived biomarkers ultimately fail to translate in the clinic due to their inherent complexity and the difficulty for pathologists to evaluate them in new datasets. In this proof-of-concept study, we showed that SuperHistopath can quantify well-understood features/markers already used, albeit only qualitatively or semi-quantitatively, by pathologists, including the stroma-to-tumor ratio, lymphocyte infiltration, tumor necrosis, and neuroblastoma differentiation. We also show that SuperHistopath-derived results corroborated those obtained from single-cell analysis on high-resolution samples (53). The computational efficiency of SuperHistopath, combined with the simple superpixels-enabled data collection, could facilitate its adoption in the clinic to accelerate pathologist workflow, could assist in intra-operative pathological diagnosis and should facilitate working with large datasets in clinical research.

Moving forward, we plan to expand the types of global context features extractable from SuperHistopath in more cancer types. We will also evaluate the accuracy of SuperHistopath on digitized frozen tissue sections to demonstrate its potential to assist in the rapid intra-operative pathological diagnostic. We will also update our previous framework (SuperCRF) which incorporates region classification information to improve cell classification (53) using SuperHistopath. Together both SuperHistopath and SuperCRF would provide invaluable tools to study spatial interactions across length scales to provide a deeper understanding of the cancer-immune-stroma interface, key to further unlock the potential of cancer immunotherapy (17).

In this proof-of-concept study, we applied our method to three cancer types with disparate histology without any changes (just retraining). While the approach could thus be virtually extended to any type of cancer, improvement could be made tailored to a specific global feature, cancer type or dataset and could include further exploring i) the use of SVM to combine the CNN-extracted features with handcrafted ones, ii) the use of other image color spaces which has been shown to improve classification in certain cases (66) and iii) alternative superpixel algorithms such as the efficient topology preserving segmentation (ETPS) algorithm (67). Additionally, further improvement of this proof-of-concept framework could be sought *via* experimentation with hyperparameter tuning, or the use of other custom and well-established architectures (59, 68). Since superpixels only capture small homogeneous areas, combination with other approaches such as classification of larger image patches with a deepCNN or U-net-like architectures might be more appropriate for the single purpose of segmenting some large and multi-component tumor structures, e.g. certain types of glands (16).

To conclude, our novel pipeline, SuperHistopath can accurately classify and map the complex tumor heterogeneity from low-resolution H&E-stained histology images. The resulting enhanced speed for both training and application (~5 min for classifying a WSI and as low as ~30 min for network training) and the efficient and simple collection of ground-truth datasets make SuperHistopath particularly attractive for research in rich datasets and would facilitate its adoption in the clinic to accelerate pathologist workflow in the quantification of predictive/prognosis markers derived from global features of interest.

DATA AVAILABILITY STATEMENT

The melanoma dataset comes the publicly available TCGA dataset. The neuroblastoma dataset is available from the corresponding authors upon reasonable request. The images from the triple-negative breast cancer dataset cannot be released yet due to ongoing clinical studies. The codes that support the findings of this study are available from the corresponding authors upon reasonable request.

ETHICS STATEMENT

The breast cancer clinical dataset was generated from diagnostic H&E images provided anonymised to the researchers by the Serbian Institute of Oncology. The neuroblastoma preclinical dataset was built from H&E images collected during previous in vivo studies approved by The Institute of Cancer Research Animal Welfare and Ethical Review Body and performed in accordance with the UK Home Office Animals (Scientific Procedures) Act 1986. The melanoma clinical samples come from the publicly available TCGA dataset (**Table 1**).

AUTHOR CONTRIBUTIONS

Conception and design: KZ-P, IR, YJ, YY. Development of methodology: KZ-P Analysis and interpretation of data: KZ-P, RN, IR, YJ, YY. Administrative and/or material support: RN, DK, IR, YJ. Writing and review of the manuscript: KZ-P, IR, YJ, YY. IR, YJ, and YY are co-senior authors of this study. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge financial support by The Rosetrees Trust (KZ-P, M593). RN acknowledges funding from ISCIII (FIS) and FEDER (European Regional Development Fund) PI17/01558 and CB16/12/00484. YY acknowledges funding from Cancer Research UK Career Establishment Award (C45982/A21808), Breast Cancer Now (2015NovPR638), Children's Cancer and Leukaemia Group (CCLGA201906), NIH U54 CA217376, and R01 CA185138, CDMRP Breast Cancer Research Program Award BC132057, CRUK Brain Tumor Awards (TARGET-GBM), European Commission ITN (H2020-MSCA-ITN-2019), Wellcome Trust (105104/Z/14/Z), and The Royal Marsden/ICR National Institute of Health Research Biomedical Research Centre. YJ is a Children with Cancer UK Research Fellow (2014/176). We thank Breast Cancer Now for funding IR as part of Programme Funding to the Breast Cancer Now Toby Robins Research Centre.

ACKNOWLEDGMENTS

We acknowledge Dr Snezana Susnjar's and Dr Natasa Medic Milijic's contribution to the creation and curation of the triple negative breast cancer dataset at The Serbian Institute of Oncology, which has made this study possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.586292/full#supplementary-material

REFERENCES

1. Tabesh A, Teverovskiy M, Pang H-Y, Kumar VP, Verbel D, Kotsianti A, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imaging* (2007) 26(10):1366–78. doi: 10.1109/TMI.2007.898536
2. Madabhushi A. Digital pathology image analysis: opportunities and challenges. *Imaging Med* (2009) 1(1):7. doi: 10.2217/iim.09.9
3. Kumar R, Srivastava R, Srivastava S. Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features. *J Med Eng* (2015) 2015. doi: 10.1155/2015/457906
4. Allard FD, Goldsmith JD, Ayata G, Challies TL, Najarian RM, Nasser IA, et al. Intraobserver and interobserver variability in the assessment of dysplasia in ampullary mucosal biopsies. *Am J Surg Pathol* (2018) 42(8):1095–100. doi: 10.1097/PAS.0000000000001079
5. Gomes DS, Porto SS, Balabram D, Gobbi H. Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. *Diagn Pathol* (2014) 9(1):121. doi: 10.1186/1746-1596-9-121
6. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, et al. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Hum Pathol* (2006) 37(12):1543–56. doi: 10.1016/j.humpath.2006.08.024
7. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *Am J Surg Pathol* (2018) 42(1):39. doi: 10.1097/PAS.0000000000000948
8. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. *J Am Med Inform Assoc* (2013) 20(6):1099–108. doi: 10.1136/amiajnl-2012-001540
9. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* (2019) 25(8):1301–9. doi: 10.1038/s41591-019-0508-1
10. Jones TR, Kang IH, Wheeler DB, Lindquist RA, Papallo A, Sabatini DM, et al. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinf* (2008) 9(1):482. doi: 10.1186/1471-2105-9-482
11. Yuan Y, Failmezger H, Rueda OM, Ali HR, Gräf S, Chin S-F, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* (2012) 4(157):157ra43–ra43. doi: 10.1126/scitranslmed.3004330

12. Chen CL, Mahjoubfar A, Tai L-C, Blaby IK, Huang A, Niazi KR, et al. Deep learning in label-free cell classification. *Sci Rep* (2016) 6:21471. doi: 10.1038/srep21471
13. Sirinukunwattana K, Raza SEA, Tsang Y-W, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging* (2016) 35 (5):1196–206. doi: 10.1109/TMI.2016.2525803
14. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep* (2017) 7(1):16878. doi: 10.1038/s41598-017-17204-5
15. Khoshdeli M, Cong R, Parvin B eds. "Detection of nuclei in H&E stained sections using convolutional neural networks. Biomedical & Health Informatics (BHI)". In: *2017 IEEE EMBS International Conference on*. New York, US: IEEE. doi: 10.1109/BHI.2017.7897216
16. Raza SEA, Cheung L, Shaban M, Graham S, Epstein D, Pelengaris S, et al. Micro-Net: A unified model for segmentation of various objects in microscopy images. *Med Image Anal* (2019) 52:160–73. doi: 10.1016/j.media.2018.12.003
17. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J* (2018) 16:34–42. doi: 10.1016/j.csbj.2018.01.001
18. Humphrey PA, Moch H, Cubilla AL, Ulbright TM, Reuter VE. The 2016 WHO classification of tumors of the urinary system and male genital organs—part B: prostate and bladder tumors. *Eur Urol* (2016) 70(1):106–19. doi: 10.1016/j.eururo.2016.02.028
19. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* (2010) 12(4):207. doi: 10.1186/bcr2607
20. Ma W, Wang J, Yu L, Zhang X, Wang Z, Tan B, et al. Tumor-stroma ratio is an independent predictor for survival in esophageal squamous cell carcinoma. *J Thorac Oncol* (2012) 7(9):1457–61. doi: 10.1097/JTO.0b013e318260dfe8
21. Chen Y, Zhang L, Liu W, Liu X. Prognostic significance of the tumor-stroma ratio in epithelial ovarian cancer. *BioMed Res Int* (2015) 2015. doi: 10.1155/2015/589301
22. Ruan M, Tian T, Rao J, Xu X, Yu B, Yang W, et al. Predictive value of tumor-infiltrating lymphocytes to pathological complete response in neoadjuvant treated triple-negative breast cancers. *Diagn Pathol* (2018) 13(1):66. doi: 10.1186/s13000-018-0743-7
23. Barnes TA, Amir E. HYPE or HOPE: the prognostic value of infiltrating immune cells in cancer. *Br J Cancer* (2017) 117(4):451–60. doi: 10.1038/bjc.2017.220
24. Renshaw AA, Cheville JC. Quantitative tumor necrosis is an independent predictor of overall survival in clear cell renal cell carcinoma. *Pathology* (2015) 47(1):34–7. doi: 10.1097/PAT.0000000000000193
25. Pichler M, Hutterer GC, Chromecki TF, Jesche J, Kampel-Kettner K, Rehak P, et al. Histologic tumor necrosis is an independent prognostic indicator for clear cell and papillary renal cell carcinoma. *Am J Clin Pathol* (2012) 137 (2):283–9. doi: 10.1309/AJCPLBK9L9KDYQZP
26. Bredholt G, Mannelqvist M, Stefansson IM, Birkeland E, Bø TH, Øyan AM, et al. Tumor necrosis is an important hallmark of aggressive endometrial cancer and associates with hypoxia, angiogenesis and inflammation responses. *Oncotarget* (2015) 6(37):39676. doi: 10.18632/oncotarget.5344
27. Ronneberger O, Fischer P, Brox T eds. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. New York, US: Springer.
28. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. "Deep learning for identifying metastatic breast cancer". arXiv preprint (2016) arXiv:160605718.
29. Nahid A-A, Mehrabi MA, Kong Y. Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *BioMed Res Int* (2018) 2018. doi: 10.1155/2018/2362108
30. Bándi P, van de Loo R, Intezar M, Geijs D, Ciompi F, van Ginneken B, et al. eds. "Comparison of different methods for tissue segmentation in histopathological whole-slide images". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. New York, US: IEEE (2017). doi: 10.1109/ISBI.2017.7950590
31. Bejnordi BE, Zuidhof G, Balkenhol M, Hermsen M, Bult P, van Ginneken B, et al. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *J Med Imaging* (2017) 4(4):044504. doi: 10.1117/1.JMI.4.4.044504
32. Vu QD, Graham S, Kurc T, To MNN, Shaban M, Qaiser T, et al. Methods for segmentation and classification of digital microscopy tissue images. *Front Bioeng Biotechnol* (2019) 7:53. doi: 10.3389/fbioe.2019.00053
33. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS One* (2017) 12(6):e0177544. doi: 10.1371/journal.pone.0177544
34. Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EA eds. "Multiclass tissue classification of whole-slide histological images using convolutional neural networks". In: *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. New York, US: Springer (2019).
35. Xu Z, Moro CF, Kuznyecov D, Bozóky B, Dong L, Zhang Q eds. "Tissue Region Growing for Hispathology Image Segmentation". In: *Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing*. New York, US: Association for Computing Machinery (2018). doi: 10.1145/3288200.3288213
36. Chan L, Hosseini MS, Rowsell C, Plataniotis KN, Damaskinos S eds. "Histosegnet: Semantic segmentation of histological tissue type in whole slide images". In: *Proceedings of the IEEE International Conference on Computer Vision*. New York, US: IEEE (2019).
37. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intelligence* (2018) 40(4):834–48. doi: 10.1109/TPAMI.2017.2699184
38. Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinf* (2017) 18(1):281. doi: 10.1186/s12859-017-1685-x
39. Krizhevsky A, Sutskever I, Hinton GE eds. "Imagenet classification with deep convolutional neural networks". In: *Adv Neural Inf Process Syst*. Cambridge, Massachusetts, US: MIT Press (2012).
40. Song Y, Zhang L, Chen S, Ni D, Lei B, Wang T. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. *IEEE Trans BioMed Eng* (2015) 62(10):2421–33. doi: 10.1109/TBME.2015.2430895
41. Romo D, García-Arteaga JD, Arbeláez P, Romero E eds. "A discriminant multi-scale histopathology descriptor using dictionary learning". In: *Medical Imaging 2014: Digital Pathology; 2014: International Society for Optics and Photonics*. Bellingham, Washington USA: SPIE (2014). doi: 10.1117/12.2043935
42. Qin P, Chen J, Zeng J, Chai R, Wang L. Large-scale tissue histopathology image segmentation based on feature pyramid. *EURASIP J Image Video Processing* (2018) 2018(1):75. doi: 10.1186/s13640-018-0320-8
43. Xu Z, Zhang Q eds. "Multi-scale context-aware networks for quantitative assessment of colorectal liver metastases". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. New York, US: IEEE (2018). doi: 10.1109/BHI.2018.8333445
44. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, Van De Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* (2011) 3(108):108ra13–ra13. doi: 10.1126/scitranslmed.3002564
45. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing* (2016) 191:214–23. doi: 10.1016/j.neucom.2016.01.034
46. Bejnordi BE, Litjens G, Karssemeijer N, van der Laak JA eds. "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images". In: *Medical Imaging 2015: Digital Pathology; 2015: International Society for Optics and Photonics*. Bellingham, Washington USA: SPIE (2015). doi: 10.1117/12.2081768
47. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans Pattern Anal Mach Intelligence* (2012) 34(11):2274–82. doi: 10.1109/TPAMI.2012.120
48. Brockmann M, Poon E, Berry T, Carstensen A, Deubzer HE, Rycak L, et al. Small molecule inhibitors of aurora-a induce proteasomal degradation of N-myc in childhood neuroblastoma. *Cancer Cell* (2013) 24(1):75–89. doi: 10.1016/j.ccr.2013.05.005

49. Berry T, Luther W, Bhatnagar N, Jamin Y, Poon E, Sanda T, et al. The ALK (F1174L) mutation potentiates the oncogenic activity of MYCN in neuroblastoma. *Cancer Cell* (2012) 22(1):117–30. doi: 10.1016/j.ccr.2012.06.001
50. Moore HC, Wood KM, Jackson MS, Lastowska MA, Hall D, Imrie H, et al. Histological profile of tumors from MYCN transgenic mice. *J Clin Pathol* (2008) 61(10):1098–103. doi: 10.1136/jcp.2007.054627
51. Jamin Y, Glass L, Hallsworth A, George R, Koh DM, Pearson AD, et al. Intrinsic susceptibility MRI identifies tumors with ALKF1174L mutation in genetically-engineered murine models of high-risk neuroblastoma. *PLoS One* (2014) 9(3):e92886. doi: 10.1371/journal.pone.0092886
52. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graphics Applications* (2001) 21(5):34–41. doi: 10.1109/38.946629
53. Zormpas-Petridis K, Failmezger H, Raza SEA, Roxanis I, Jamin Y, Yuan Y. Superpixel-based Conditional Random Fields (SuperCRF): Incorporating global and local context for enhanced deep learning in melanoma histopathology. *Front Oncol* (2019) 9:1045. doi: 10.3389/fonc.2019.01045
54. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* (2014) 15(1):1929–58. doi: 10.5555/2627435.2670313
55. Glorot X, Bengio Y eds. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. Proceedings of Machine Learning Research (PMLR) (2010).
56. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint* (2014) arXiv:14126980.
57. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z eds. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. New York, US: IEEE (2015).
58. Chollet F ed. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. New York, US: IEEE (2016).
59. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA eds. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Thirty-First AAAI Conference on Artificial Intelligence*. California, US: AAAI (2016).
60. He K, Zhang X, Ren S, Sun J eds. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. New York, US: IEEE (2015).
61. Heindl A, Sestak I, Naidoo K, Cuzick J, Dowsett M, Yuan Y. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. *JNCI: J Natl Cancer Institute* (2018) 110(2):166–75. doi: 10.1093/jnci/djx137
62. Nawaz S, Heindl A, Koelble K, Yuan Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod Pathol* (2015) 28(6):766. doi: 10.1038/modpathol.2015.37
63. Gilchrist KW, Gray R, Fowble B, Tormey DC, Taylor 4th S. Tumor necrosis is a prognostic predictor for early recurrence and death in lymph node-positive breast cancer: a 10-year follow-up study of 728 Eastern Cooperative Oncology Group patients. *J Clin Oncol* (1993) 11(10):1929–35. doi: 10.1200/JCO.1993.11.10.1929
64. Hanafy E, Al Jabri A, Gadelkarim G, Dasaq A, Nazim F, Al Pakrah M. Tumor histopathological response to neoadjuvant chemotherapy in childhood solid malignancies: is it still impressive? *J Invest Med* (2018) 66(2):289–97. doi: 10.1136/jim-2017-000531
65. Lambertz I, Kumps C, Claeys S, Lindner S, Beckers A, Janssens E, et al. Upregulation of MAPK Negative Feedback Regulators and RET in Mutant ALK Neuroblastoma: Implications for Targeted Treatment. *Clin Cancer Res* (2015) 21(14):3327–39. doi: 10.1158/1078-0432.CCR-14-2024
66. Gowda SN, Yuan C eds. "ColorNet: Investigating the importance of color spaces for image classification". In: *Asian Conference on Computer Vision*. New York, US: Springer.
67. Yao J, Boben M, Fidler S, Urtasun R eds. "Real-time coarse-to-fine topologically preserving segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York, US: IEEE (2015).
68. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint* (2019) arXiv:190511946.

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.

frontiers in Oncology

ORIGINAL RESEARCH
published: 27 April 2021
doi: 10.3389/fonc.2021.575166

Deep Learning Based on ACR TI-RADS Can Improve the Differential Diagnosis of Thyroid Nodules

Ge-Ge Wu[1], Wen-Zhi Lv[2], Rui Yin[3], Jian-Wei Xu[4], Yu-Jing Yan[1], Rui-Xue Chen[5], Jia-Yu Wang[1], Bo Zhang[6]*, Xin-Wu Cui[1]* and Christoph F. Dietrich[7]

[1] Sino-German Tongji-Caritas Research Center of Ultrasound in Medicine, Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2] Department of Artificial Intelligence, Julei Technology Company, Wuhan, China, [3] Department of Ultrasound, Affiliated Renhe Hospital of China Three Gorges University, Yichang, China, [4] Department of Ultrasound, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [5] Department of Ultrasound, Wuchang Hospital, Wuhan University of Science and Technology, Wuhan, China, [6] Department of Ultrasonic Imaging, Xiangya Hospital, Central South University, Changsha, China, [7] Department of General Internal Medicine, Kliniken Hirslanden Beau-Site, Bern, Switzerland

OPEN ACCESS

Edited by:
Po-Hsiang Tsui,
Chang Gung University, Taiwan

Reviewed by:
Chandramohan Anuradha,
Christian Medical College & Hospital, India
Jacob Jaremko,
University of Alberta, Canada

Correspondence:
Xin-Wu Cui
cuixinwu@live.cn
Bo Zhang
zhangbo8095@csu.edu.cn

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *07 December 2020*
Accepted: *07 April 2021*
Published: *27 April 2021*

Citation:
Wu G-G, Lv W-Z, Yin R, Xu J-W, Yan Y-J, Chen R-X, Wang J-Y, Zhang B, Cui X-W and Dietrich CF (2021) Deep Learning Based on ACR TI-RADS Can Improve the Differential Diagnosis of Thyroid Nodules. Front. Oncol. 11:575166. doi: 10.3389/fonc.2021.575166

Objective: The purpose of this study was to improve the differentiation between malignant and benign thyroid nodules using deep learning (DL) in category 4 and 5 based on the Thyroid Imaging Reporting and Data System (TI-RADS, TR) from the American College of Radiology (ACR).

Design and Methods: From June 2, 2017 to April 23, 2019, 2082 thyroid ultrasound images from 1396 consecutive patients with confirmed pathology were retrospectively collected, of which 1289 nodules were category 4 (TR4) and 793 nodules were category 5 (TR5). Ninety percent of the B-mode ultrasound images were applied for training and validation, and the residual 10% and an independent external dataset for testing purpose by three different deep learning algorithms.

Results: In the independent test set, the DL algorithm of best performance got an AUC of 0.904, 0.845, 0.829 in TR4, TR5, and TR4&5, respectively. The sensitivity and specificity of the optimal model was 0.829, 0.831 on TR4, 0.846, 0.778 on TR5, 0.790, 0.779 on TR4&5, versus the radiologists of 0.686 (P=0.108), 0.766 (P=0.101), 0.677 (P=0.211), 0.750 (P=0.128), and 0.680 (P=0.023), 0.761 (P=0.530), respectively.

Conclusions: The study demonstrated that DL could improve the differentiation of malignant from benign thyroid nodules and had significant potential for clinical application on TR4 and TR5.

Keywords: artificial intelligence, thyroid imaging reporting and data system (TI-RADS), ultrasound, thyroid cancer, deep learning

INTRODUCTION

With the utilization of high-frequency ultrasound in clinical practice and the gradual enhancement of public health awareness especially on physical examination, the detection of thyroid nodules (TN) has increased, with a prevalence ranging from 19% to 68% in the general unselected population (1, 2). Moreover, the incidence rate of thyroid cancer has continued to increase and is now the highest cause of cancer in women under 30 years old in China (3, 4). Ultrasound has an irreplaceable role in early detection of thyroid cancer due to its accessibility, high resolution, safety, using no radiation, and provision of real-time imaging with multi-dimensions. Experience and skills of different operators influence the accurate differential diagnosis of TN, and thus, a precise and independent method is needed.

To implement standardized management of the thyroid nodules, the Thyroid Imaging Reporting and Data System (TI-RADS) Committee of American College of Radiology (ACR) published a white paper in 2017 that presented a new risk stratification system from TR1 to TR5 for classifying thyroid nodules by adding scores of the five characteristics on ultrasound, composition, echogenicity, shape, margin, and echogenic foci (5). Recommendations for biopsy or ultrasound follow-up are determined on the nodule's ACR TI-RADS categories and its maximum diameter (6), which provides clarity for the further diagnosis and treatment measures. The guidance of ACR TI-RADS has been proven to be a reliable tool to assist doctors to differentiate between malignant and benign thyroid nodules (7–11), with a pooled sensitivity of 0.79 (95% confidence interval [CI] = 0.77-0.81) and a pooled specificity of 0.71 (95% CI = 0.70-0.72) (12, 13).

Artificial Intelligence (AI) is of unique value for its time-saving and non-dependence on radiologist's experience, and performs extremely well on the tasks of detection, extraction and classification of the TN on ultrasound images (14–18). Recently, AI has accomplished many complex tasks on thyroid ultrasound, such as the differentiation of malignant from benign thyroid nodules using ultrasound images from multiple cohorts (19), developing a deep learning (DL) algorithm to decide whether a TN should undergo a biopsy (16), using ultrasound elastography to improve thyroid nodule discrimination (20) and applying ultrasound images to predict metastasis in the cervical lymph nodes (21, 22).

However, there are still some flaws in these studies. First, pathological results of some nodules are missing in almost all of the published studies (19). Second, all types of thyroid nodules were included, but some nodules are easily diagnosed by doctors and AI is not that necessary. For example, cystic nodules are usually echoless with clear boundaries and it is not surprising that AI performs diagnosing them as benign.

ACR TI-RADS is popularly used in routine clinical practice, and has proven value. It is still an open question if the combination of DL and TI-RADS can improve the differential diagnosis of TNs. TR1, TR2, TR3 have a very low (less than 5%) chance of malignancy (6) and the necessity for them to proceed AI analysis seem less sufficient. Adversely, malignant thyroid nodules were most distributed in TR4 and TR5. However, it is difficult for radiologists to differentiate benign from malignant nodules in the same category causing that they have same ultrasound descriptive features (23). A non-invasive method such as DL is needed to avoid the need for unnecessary biopsy.

The purpose of this study was to evaluate whether DL based on ACR TI-RADS category 4 and 5 could improve the differentiation of malignant from benign thyroid nodules, and explore the clinical application potential for it.

MATERIALS AND METHODS

Source of the Data

This study was approved by the Ethics Committee of Tongji Medical College of Huazhong University of Science and Technology. Informed consent from the patients was exempted (2019S1233). All ultrasound images included were consecutively acquired from 11 operators with more than 5 years of experience from Tongji hospital, Wuhan, China (internal cohort), and Xiangya Hospital of Central South University, Changsha, China (external cohort) from June 2017 to April 2019. Ultrasound equipment manufactured by GE Healthcare (LOGIQ E9, LOGIQ S7), Samsung (RS80A), and Philips (EPIQ5, EPIQ7 and IU22), was used to generate the thyroid ultrasound images. Ultrasound images were derived from the picture archiving and communication system (PACS) workstations.

Images Enrolments and Grouping

The inclusion criteria for thyroid nodules in this study were patients who 1) underwent total or nearly total thyroidectomy or lobectomy; 2) had pathological specimens examined within one month after US examination; 3) had complete medical information including preoperative ultrasound of the thyroid nodules; 4) had no previous surgical treatment or FNA performed on the nodules.

Exclusion criteria were lesions 1) with unsatisfactory ultrasound image quality; 2) where the finding on ultrasound did not match with the pathological results in position or size; 3) received chemotherapy and/or radiotherapy such as iodine 131 treatment before ultrasound examination.

From June 2nd, 2017 to April 23th, 2019, 4910 thyroid images from 2779 consecutive patients and 213 thyroid images from 195 consecutive patients with confirmed postoperative pathological results were retrospectively collected in Tongji hospital and Xiangya Hospital of Central South University. Three doctors (C.R, Y.R, and W.G) scored these images on the five features according to ACR TI-RADS lexicon (6). The opinion of the third was referred to for cases where the first opinions differed. Only nodules of TI-RADS category 4 (dataset I) and category 5 (dataset II) were enrolled, and they were merged together as new dataset III (i.e. combination of ACR TI-RADS 4 and 5). In accordance with the pathological results, images of each category were sorted out into a benign group and a malignant group.

Establishment of Training Set and Test Set

Each inner dataset (I, II, III) was randomly divided into two sets, 90% for training and validation, and the residual 10% (test set A) for testing. In addition, another independent outer test set (test set B) was obtained for testing as well. Three convolutional neutral Network (CNN) models named ResNet-50, Inception-Resnet v2, Desnet-121 were used for analysis. The workflow of the selection and construction is shown in **Figure 1**.

Three independent experienced radiologists (X.J and Y.Y and Z.B) with 8 years, 9 years and 24 years of experience, respectively, read the images and gave their judgments according to the ACR TI-RADS lexicon (5, 6) and their own clinical experience. If their opinions did not agree, the opinion of the most senior radiologist was used.

Processing of Ultrasound Images

Nodules were manually marked, and the region of interest (ROI) of the thyroid nodules was cut out using rectangular boxes by Image J (version 1.48, National Institutes of Health, USA) by a radiologist, in which the cropped images include the entire thyroid nodule. All the images were resized to 299 × 299 pixels to standardize the distance scale. Due to the limited quantity of the dataset, augmentation strategy was introduced to process the images. All preprocessing steps were conducted using the Keras Image Data Generator and then fed into the input.

Construction of CNNs

The tasks on three sets (datasets I, II, and III) were trained on three pre-trained convolutional neural networks, named ResNet50, Inception-ResNet v2, Desnet 121, respectively. The initialization set of the parameters of these models was referred to ImageNet and obtained from Keras Team (https://github.com/keras-team/keras-applications/releases). The learning rate was set to 0.03 and decelerated by a factor of 0.1 for each 50 epochs when the accuracy had no further improvement in the training and validation set. Model learning continued until the least loss of the validation set appeared and the final model was determined accordingly. Optimizer of Stochastic Gradient Descent (SGD) and binary cross entropy technique were used to decrease loss in the process in CNNs. All models were trained in Python 3.6.2 (https://www.python.org) by using a computer with a GeForce GTX 2080 Ti graphics processing unit (NVIDIA, Santa Clara, California, America), a Core i9-9900K central processing unit (Intel, Santa Clara, California, America).

The class activation mapping (CAM) technique was also used to produce the heated maps which indicated the focus of the CNN model's prediction (24, 25). The CAM can be regarded as

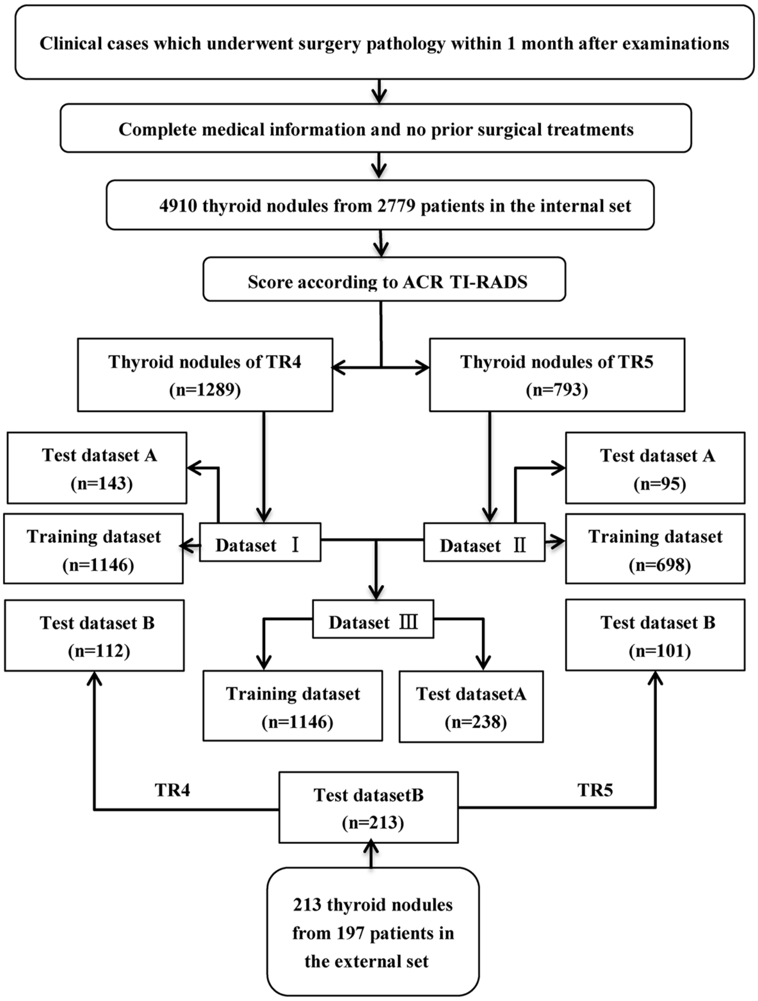

Clinical cases which underwent surgery pathology within 1 month after examinations
Complete medical information and no prior surgical treatments
4910 thyroid nodules from 2779 patients in the internal set
Score according to ACR TI-RADS
Thyroid nodules of TR4 (n=1289)
Thyroid nodules of TR5 (n=793)
Test dataset A (n=143)
Training dataset (n=1146)
Dataset Ⅰ
Dataset Ⅱ
Test dataset A (n=95)
Training dataset (n=698)
Test dataset B (n=112)
Dataset Ⅲ
Test dataset B (n=101)
Training dataset (n=1146)
Test datasetA (n=238)
TR4
Test datasetB (n=213)
TR5
213 thyroid nodules from 197 patients in the external set

FIGURE 1 | Workflow of the construction of the training and test dataset.

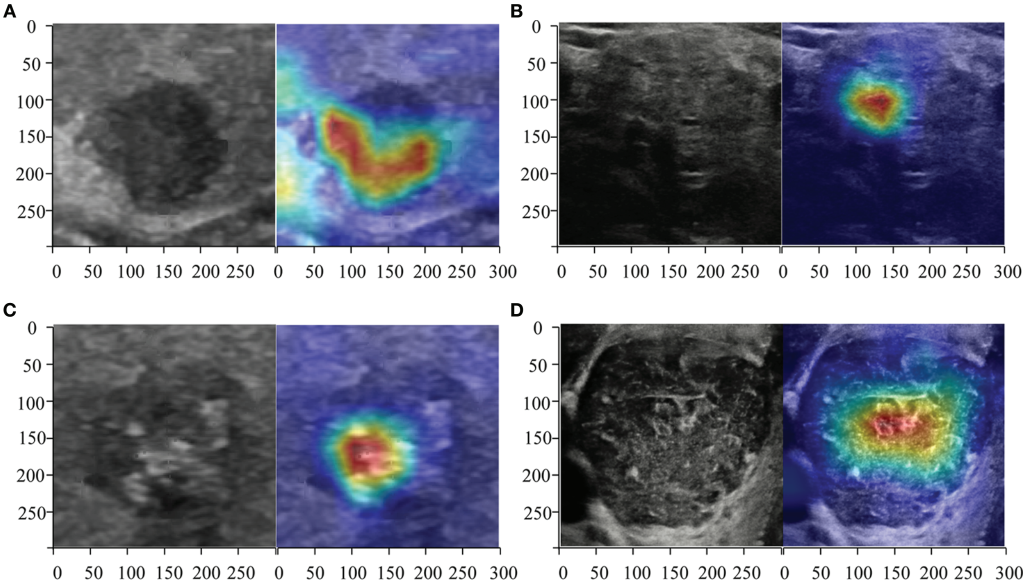


FIGURE 2 | Heatmaps of the region of interest (ROI) of the thyroid nodules using class activation mapping (CAM). The red color showed the prediction regions the CNNs focused which estimated to be determined as the thyroid cancer. Three radiologists and DL correctly predicted a malignant **(A)** thyroid nodule diagnosed as micro papillary carcinoma TR4 and a benign **(B)** one diagnosed as non-toxic nodular goiter of TR4. ResNet50, Desnet121, and the radiologists deemed a malignant nodule **(C)** diagnosed as papillary carcinoma of TR5 as malignance but a DL algorithm named Inception-ResNet version 2 judged it as benign. All CNNs correctly predicted a benign **(D)** thyroid nodule diagnosed as Hashimoto's thyroiditis of TR5 but the radiologists all predicted wrongly.

the multiplication of the feature maps of the pooling layers and weight of the fully connected layer, which prevented loss of the special information when feature maps were transferred to eigenvector. It highlighted the specific discriminative regions demonstrated as thyroid cancer by CNN. Packages Matplotlib 3.1.1 (https://matplotlib.org) and Open cv-Python 3.4.4.19 (https://github.com/skvark/opencv-python) was employed to generate heatmaps (**Figure 3**).

Statistical Analysis

The performance of the three algorithms was measured by the area under the receiver operating characteristic curve (AUROC) of the training and test dataset. The cut-off value was obtained as the threshold value when the Youden index reached its maximum. Then, the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of each method were calculated to judge the performance of the experts and the CNNs. Delong test was introduced to evaluate the statistical difference between different AUCs. Ninety-five percent confidence interval (CI) was utilized to estimate the range of these evaluation values. P-value less than 0.05 with two tailed was considered statistically significant. Interobserver agreements on thyroid nodules were assessed using Kruskal–Wallis test. Kappa values were interpreted as follows. Less than 0.20 mean poor agreement, from 0.20 to 0.40 mean fair agreement, from 0.40 to 0.60 imply moderate agreement, between 0.60 and 0.80 imply substantial agreement, and excellent agreement tend to be over 0.80. F score was introduced to measure the efficiency of the CNNs while taking both Precision and Recall into account, the formula is as follows. When β = 1, the F1 score improves Precision and Recall as much as possible, and makes the difference between the two as small as possible.

$$F \ \ score = \left(1+\beta^2\right) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

The curve of ROC was performed and portraited using the pROC package of R software (version 1.8) and MedCalc (version 11.2, Ostend, Belgium). Outcome of evaluation values was also obtained by SPSS (version 22.0, IBM, Chicago) and R software.

RESULTS

Characteristics of the Thyroid Nodules

A total of 2295 thyroid images from 1593 patients were used in this research (**Table 1**). In the internal cohort, the mean age of all patients was 45.48 ± 10.33, of which 1059 were woman, 337 were men. In the external cohort, the mean age of all patients was 45.54 ± 11.82, of which 150 were woman, 47 were men. 1146 thyroid images of TR4 and 698 thyroid images of TR5 were enrolled in training set in this research, which consisted of 637 benign images and 509 malignant images in the former, 297 benign images and 401 malignant images in the latter. 143 thyroid images of TR4 and 95 thyroid images of TR5 were predicted for the internal test in this

TABLE 1 | Basic information of the patients.

	Internal dataset (n=1396)	External dataset (n=197)
Age (year)	45.48 ± 10.33 (8-71)	45.54 ± 11.82 (16-77)
≤20	13(0.9)	1(0.5)
20-30	85(6.1)	27(13.7)
30-40	281(20.1)	37(18.8)
40-50	549(39.3)	62(31.5)
≥50	468(33.5)	70(35.5)
Gender		
Male	337(24.1)	47(23.9)
Female	1059(75.9)	150(76.1)

research, while 112 of TR4 and 101 of TR5 for the external test. The characteristics of the thyroid nodules in five ACR TI-RADS features were summarized in **Table 2**.

DL Performance Compared With Radiologists

The performance of DL was better compared to the radiologists in three tasks. In the internal test set, the AUROC of the best algorithm in differentiation of thyroid nodules was 0.936 (95%CI 0.898-0.973) in TR4, 0.915 (95%CI 0.857-0.973) in TR5 and 0.892 (95%CI 0.850-0.933) in TR 4&5 respectively, which overwhelmingly exceeded the radiologists respectively ($P < 0.001$). In the external test set, the AUROC of the optimal algorithm was 0.904 (95%CI 0.833-0.951) in TR4, 0.845 (95%CI 0.759-0.909) in TR5 and 0.829 (95%CI 0.772-0.877) in TR 4&5 respectively, which again was better than the radiologists ($P < 0.001$).

Evaluation of the performance on differentiation of malignant from benign thyroid nodules in TR4, TR 5 and TR 4&5 were

TABLE 2 | Characteristics of the thyroid nodules in internal set enrolled in this survey.

	Task1			Task2			Task3		
	Training dataset (n=1146)	Test dataset A (n=143)	Test dataset B (n=112)	Training dataset (n=698)	Test dataset A (n=95)	Test dataset B (n=101)	Training dataset (n=1844)	Test dataset A (n=238)	Test dataset B (n=213)
Pathology									
benign	637(55.6)	70(49.0)	77(68.8)	297(42.6)	32(33.7)	36(35.6)	934(50.7)	102(42.9)	113(53.1)
malignant	509(44.4)	73(51.0)	35(31.2)	401(57.4)	63(66.3)	65(64.4)	910(49.3)	136(57.1)	100(46.9)
Diameter (mm)									
≤ 0.5	221(19.3)	26(18.1)	19(17.0)	93(13.3)	14(14.7)	9(8.9)	314(17.0)	40(16.8)	28(13.1)
0.5-1.0	431(37.6)	57(39.9)	55(49.0)	295(42.3)	41(43.2)	35(34.7)	726(39.4)	98(41.2)	90(42.3)
1.0-2.0	176(15.4)	39(27.3)	28(25.0)	125(17.9)	25(26.3)	29(28.7)	301(16.3)	64(26.9)	57(26.8)
> 2.0	318(27.7)	21(14.7)	10(9.0)	185(36.5)	15(15.8)	28(27.7)	503(23.3)	36(15.1)	38(17.8)
Internal Composition									
Cystic/partially cystic/spongifom	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mixed	48(4.2)	7(4.9)	6(5.3)	3(0.4)	1(1.1)	1(1.0)	51(2.8)	8(3.4)	7(3.3)
Solid/almost solid	1098(95.8)	136(95.1)	106(94.6)	695(99.6)	94(98.9)	100(99.0)	1793(97.2)	230(96.6)	206(96.7)
Echogenicity									
Anechoic	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Hyperechoic/ isoechoic	30(2.6)	6(4.2)	4(3.6)	5(0.7)	1(1.1)	0(45)	35(1.9)	7(2.9)	4(1.9)
Hypoechoic	1113(97.1)	137(95.8)	107(95.5)	681(95.6)	92(96.8)	100(99.0)	1814(98.3)	229(96.2)	207(97.2)
Very hypoechoic	3(0.3)	0(0)	1(0.9)	12(1.7)	2(2.1)	1(1.0)	15(0.8)	2(0.8)	2(0.9)
Shape									
Wider-than-tall	1143(99.7)	142(99.3)	112(100.0)	478(68.5)	65(68.4)	70(69.3)	1621(87.9)	207(87.0)	182(85.4)
Taller-than-wide	3(0.3)	1(0.7)	0(0)	220(31.5)	30(31.6)	32(31.7)	223(12.1)	31(13.0)	31(14.6)
Margins									
Smooth/ Ill-defined	992(86.6)	108(75.5)	85(78.9)	477(68.3)	60(63.2)	62(61.4)	1469(79.7)	168(70.6)	147(69.0)
Lobulated/ irregular	153(13.3)	35(37.5)	27(24.1)	210(30.1)	30(31.5)	32(31.7)	363(19.7)	65(27.3)	59(27.7)
Extra-thyroid extension	1(0.1)	0(0)	0(0)	11(1.6)	5(5.3)	7(6.9)	12(0.6)	5(2.1)	7(3.3)
Echogenic foci									
None/large comet-tail artifacts	991(86.5)	122(85.3)	93(83.0)	92(13.2)	16(16.8)	19(18.8)	1083(58.7)	138(58.0)	112(52.6)
	133(11.6)	16(11.2)	10(8.9)	17(2.4)	5(5.3)	3(3.0)	150(8.1)	21(8.8)	13(6.1)
Macrocalcifications									
Peripheral calcifications	22(1.9)	6(4.2)	6(5.4)	3(0.4)	0(0)	1(1.0)	25(1.4)	6(2.5)	7(3.3)
Punctate echogenic foci	22(1.9)	7(4.9)	5(4.5)	601(86.1)	76(80.0)	78(77.2)	623(33.8)	83(34.9)	83(39.0)

Test set A and test set B referred to the internal test set and external test set. Data in parentheses are percentages.

TABLE 3 | Performance of deep learning containing three CNNs compared with the radiologists in differentiating benign and malignant thyroid nodules classified into ACR TI-RADS category 4.

	ResNet-50	Inception-Resnet-v2	Desnet-121	Radiologists	*P* value
Internal dataset (n=143)					
Accuracy	0.874 (0.810-0.919)	0.846 (0.778-0.896)	0.846 (0.778-0.896)	0.734 (0.656-0.800)	0.010
Sensitivity	0.836 (0.727-0.909)	0.918 (0.824-0.966)	0.863 (0.758-0.929)	0.684 (0.564-0.786)	0.004
Specificity	0.914 (0.816-0.965)	0.771 (0.653-0.860)	0.871 (0.765-0.936)	0.786 (0.668-0.871)	0.066
PPV	0.910 (0.809-0.963)	0.807 (0.703-0.883)	0.875 (0.771-0.938)	0.769 (0.645-0.861)	0.115
NPV	0.842 (0.736-0.912)	0.900 (0.788-0.959)	0.859 (0.752-0.927)	0.705 (0.590-0.800)	0.024
Kappa value	0.749	0.691	0.693	0.470	
F_1	0.846	0.775	0.846	0.649	
AUROC	0.936 (0.898-0.973)	0.902 (0.853-0.952)	0.911 (0.865-0.958)	0.735 (0.652-0.819)	
External dataset (n=112)					
Accuracy	0.830 (0.749-0.890)	0.821 (0.739-0.882)	0.795 (0.710-0.860)	0.741 (0.653 -0.814)	0.033
Sensitivity	0.829 (0.657-0.928)	0.657 (0.477-0.803)	0.800 (0.625-0.909)	0.686 (0.506-0.826)	0.108
Specificity	0.831 (0.725-0.904)	0.896 (0.800-0.951)	0.792 (0.682-0.873)	0.766 (0.653-0.852)	0.101
PPV	0.690 (0.528-0.819)	0.742 (0.551-0.875)	0.636 (0.477-0.772)	0.571 (0.410-0.719)	0.037
NPV	0.914 (0.816-0.965)	0.852 (0.752-0.918)	0.897 (0.793-0.954)	0.843 (0.732-0.915)	0.226
Kappa value	0.626	0.571	0.553	0.429	
F_1	0.812	0.785	0.775	0.713	
AUROC	0.904 (0.833-0.951)	0.845 (0.765-0.907)	0.842 (0.761-0.904)	0.726 (0.634-0.806)	

recorded in **Tables 3–5**, respectively. ResNet-50 performed best in the certain classification in both TR4 and TR5 dataset. Meanwhile, performance in two datasets was also excellent with a stable repeatability, of which the kappa value was all over 0.50.

Heatmaps Generated by CAM

Heatmaps were generated to present the recognition pattern of the deep learning model as demonstrated in **Figure 2**. The greatest predictive regions of the tumor CNNs concentrated were shown as red and yellow; whereas the areas green and blue regions were of less predictive significance. This shows that the DL algorithms focuses on the most predictive image features of thyroid nodules malignance risk.

TABLE 4 | Performance of deep learning containing three CNNs compared with the radiologists in differentiating benign and malignant thyroid nodules classified into ACR TI-RADS category 5.

	ResNet-50	Inception-Resnet-v2	Desnet-121	Radiologists	*P* value
Internal dataset (n=95)					
Accuracy	0.863 (0.780-0.918)	0.811 (0.720-0.877)	0.832 (0.744-0.894)	0.695 (0.596-0.778)	0.022
Sensitivity	0.841 (0.723-0.917)	0.841 (0.723-0.917)	0.952 (0.858-0.988)	0.635 (0.504-0.750)	<0.001
Specificity	0.906 (0.738-0.975)	0.750 (0.562-0.879)	0.594 (0.408-0.758)	0.813 (0.630-0.921)	0.026
PPV	0.946 (0.842-0.986)	0.869 (0.752-0.938)	0.822 (0.711-0.898)	0.870 (0.730-0.946)	0.055
NPV	0.744 (0.576-0.864)	0.706 (0.523-0.843)	0.864 (0.640-0.964)	0.531 (0.384-0.672)	0.026
Kappa value	0.709	0.592	0.582	0.396	
F_1	0.854	0.791	0.793	0.688	
AUROC	0.915 (0.857-0.973)	0.838 (0.756-0.919)	0.906 (0.846-0.966)	0.724 (0.617-0.831)	
External dataset (n=101)					
Accuracy	0.822 (0.735-0.885)	0.713 (0.618 to 0.792)	0.802 (0.713-0.869)	0.703 (0.607-0.784)	0.080
Sensitivity	0.846 (0.731-0.920)	0.615 (0.486-0.731)	0.754 (0.629-0.849)	0.677 (0.548-0.785)	0.211
Specificity	0.778 (0.604-0.893)	0.889 (0.730-0.964)	0.889 (0.730-0.964)	0.750 (0.575-0.873)	0.128
PPV	0.873 (0.760-0.940)	0.909 (0.774-0.970)	0.925 (0.809-0.976)	0.830 (0.697-0.915)	0.132
NPV	0.737 (0.566-0.860)	0.561 (0.424-0.690)	0.667 (0.515-0.792)	0.563 (0.413-0.702)	0.203
Kappa value	0.616	0.446	0.598	0.397	
F_1	0.808	0.711	0.796	0.694	
AUROC	0.845 (0.759-0.909)	0.770 (0.676-0.848)	0.842 (0.756-0.907)	0.713 (0.615-0.799)	

DISCUSSION

In this study, we combined ACR TI-RADS with DL by training three commonly used deep learning algorithms to discriminate between benign and malignant in TR4 and TR5 thyroid nodules with available pathology. As shown in **Figure 3**, no matter which type of TI-RADS was used for the classification competition, DL algorithms performed better than radiologists. The accuracy in all models was higher in TR4 and TR5 for test set A and test set B, which was parallel to the performance of the radiologists. However, in the case of mixing different feature sets containing TR4 and TR5, DL still had good performance but slightly weaker than the two separated sets, which might be related to more complex tasks.

TABLE 5 | Performance of deep learning containing three CNNs compared with the radiologists in differentiating benign and malignant thyroid nodules classified into ACR TI-RADS category 4 and 5.

	ResNet-50	Inception-Resnet-v2	Desnet-121	Radiologists	*P* value
Internal dataset (n=238)					
Accuracy	0.832 (0.779-0.874)	0.811 (0.756-0.856)	0.824 (0.770-0.867)	0.718 (0.658-0.772)	0.007
Sensitivity	0.882 (0.813-0.929)	0.794 (0.715-0.857)	0.824 (0.747-0.882)	0.662 (0.711-0.898)	<0.001
Specificity	0.745 (0.647-0.824)	0.833 (0.744-0.897)	0.843 (0.755-0.905)	0.794 (0.700-0.865)	0.227
PPV	0.822 (0.748-0.878)	0.864 (0.788-0.916)	0.875 (0.802-0.925)	0.811 (0.723-0.877)	0.429
NPV	0.826 (0.730-0.894)	0.752 (0.660-0.826)	0.782 (0.691-0.853)	0.638 (0.547-0.720)	0.009
Kappa value	0.635	0.619	0.660	0.442	
F_1	0.852	0.784	0.836	0.668	
AUROC	0.879 (0.835-0.922)	0.883 (0.841-0.926)	0.892 (0.850-0.933)	0.728 (0.663-0.793)	
External dataset (n=213)					
Accuracy	0.784 (0.724-0.834)	0.770 (0.709-0.822)	0.761 (0.699-0.813)	0.723 (0.659-0.779)	0.009
Sensitivity	0.790 (0.695-0.862)	0.860 (0.773-0.919)	0.710 (0.609-0.794)	0.680 (0.578-0.768)	0.023
Specificity	0.779 (0.689-0.849)	0.690 (0.595-0.772)	0.805 (0.718-0.871)	0.761 (0.670-0.834)	0.530
PPV	0.760 (0.664-0.836)	0.711 (0.620-0.788)	0.763 (0.662-0.843)	0.716 (0.613-0.801)	0.055
NPV	0.807 (0.718-0.874)	0.848 (0.754-0.911)	0.758 (0.670-0.830)	0.729 (0.638-0.805)	0.071
Kappa value	0.567	0.544	0.517	0.442	
F_1	0.784	0.770	0.758	0.722	
AUROC	0.829 (0.772-0.877)	0.807 (0.748-0.858)	0.793 (0.733-0.845)	0.721 (0.655-0.780)	

Patients with suspected thyroid nodules, nodular goiter, nodules accidentally discovered by radiological examination such as computed tomography (CT), magnetic resonance imaging (MRI), or 18F-flurodeoxyglucose positron emission computed tomography (FDP18-PET) scan showing thyroid uptake should undergo diagnostic thyroid ultrasound examination as recommended by ATA Guidelines 2015 (26). The benign and malignant ultrasound results of nodules will determine whether FNA and follow-up are to be carried out (27), and the choice of treatment methods will be influenced by ultrasound opinions and cervical lymph node conditions (28). In ultrasound diagnosis, malignant nodules have various manifestations and particularly those with atypical appearances and fuzzy boundaries lead to diagnostic difficulties (29, 30). Radiologists frequently disagree over the interpretation of these malignant tumors. DL may provide assistance for radiologists with good accuracy and consistency.

The performance of DL is often better than that of radiologists and even machine learning, in the diagnosis of thyroid nodules. Xia and colleagues (31) achieved an accuracy of 87.7% in differentiating malignant and benign nodules by constructing extreme machine learning based on collected features obtained from 203 ultrasound images of 187 patients with thyroid cancer. Li and colleagues (19) got an accuracy of 89.8% (95% CI 86.8–92.3) in internal validation set with the DCNN model versus 78.8% with the radiologists and 85.7% (95% CI 79.2–90.8) versus 72.7% (65.0–79.6%) in external validation set. Machine learning gives opinions by extracting computational features and calculating statistically significant finite features and modeling. The modeling process of machine learning requires the segmentation of images to be more accurate, while the commonly manual work is difficult to control. Limited quantities of features and smaller sample size also resulted in inferior performance and narrow application range.

Moreover, the DL result in thyroid nodules of all TR categories was not that impressive because it contained some tasks that even radiological beginners can do such as recognizing and selecting the TR1 nodules and labelling them as benign (5). Limiting the work to differentiation between subtype TR4 and TR5 is difficult for radiologists because they had similar visible features (20). As recent studies have reported, DL had achieved great success on the classification on thyroid cancer (32), when all types of thyroid nodules were included. In these studies, pathological results of some nodules were not available (19), while in our study all the nodules correlated with surgical pathology. Limitations of the TR categories on ultrasound images avoid heterogeneity of the dataset to a degree. In specific classification, our study revealed that a precise set of certain categories contributed to the higher accuracy compared with former studies (19, 32).

The result of this study may potentially be of clinical value. TI-RADS is already widely applied worldwide and combining the TI-RADS and DL provides more accurate results and should be easily accepted clinically. Previous studies had reported that interobserver agreement in the lexicon was also substantial thus the pre-classification was easily performed and credible wherever used (33). Application of the DL based on ACR TI-RADS will supply useful suggestions when there is doubt over the diagnosis and will support services where medical resources were unbalanced.

Our study also had limitations. First, this was a retrospective study with limited categories of data. The performance of our DL system is expected to increase by including more data and expanding several sets from other hospitals. And exclusion of TR3 thyroid nodules decrease clinical application to some extent. Second, ultrasound systems of different manufactures and heterogeneity of operators may give rise to the variability in the training process. The inter-reader reliability of nodule extraction was not assessed. Third, the images reviewed were static in this study that features from multi-sections were not considered.

To be summarized, the study demonstrated that DL based on ACR TI-RADS could improve the differentiation of malignant from benign thyroid nodules with great clinical application potential. With a stable repeatability, DL algorithms showed

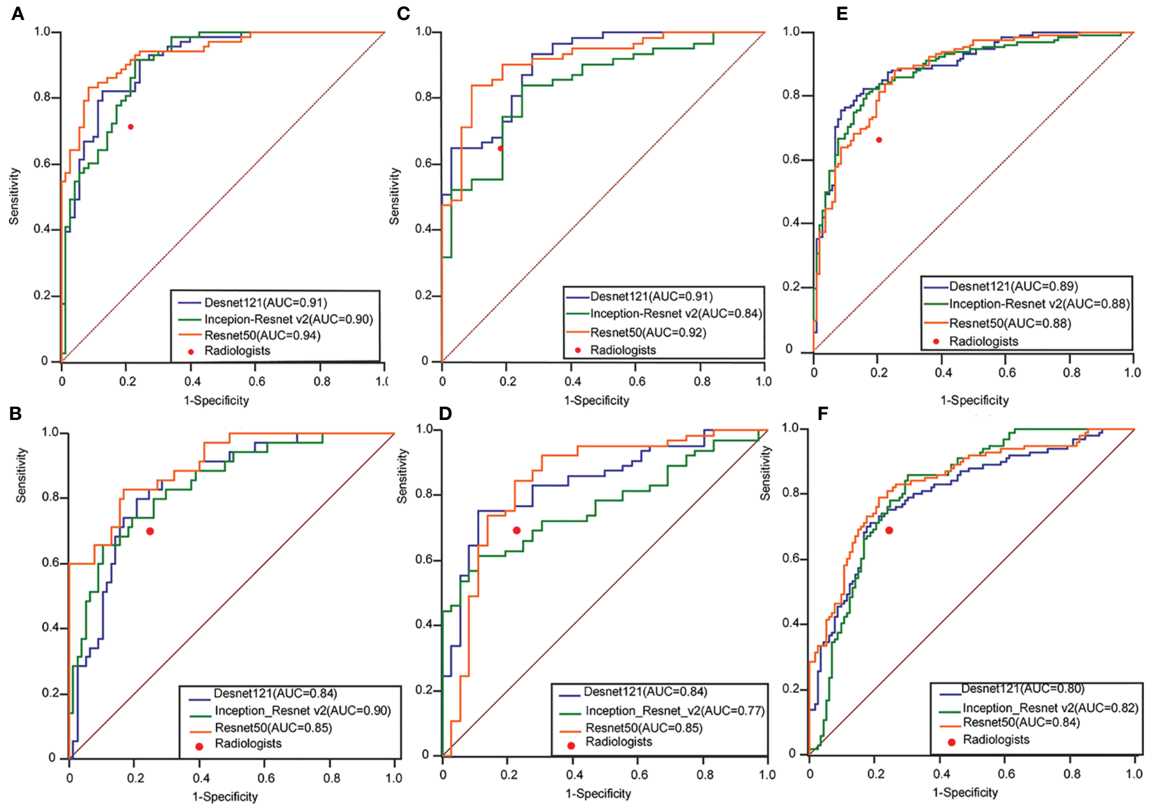


FIGURE 3 | Performance of the ensemble D-CNN models in identifying patients with thyroid cancer in TR4 **(A)**, TR5 **(C)**, and TR4&5 **(E)** on three inner test datasets and TR4 **(B)**, TR5 **(D)**, and TR4&5 **(F)** on three outer test datasets. The red dots on each ROC curve demonstrate the performance of the radiologists. AUC, area under the curve; DCNN, deep convolutional neural network; ROC, receiver operating characteristics curve.

better performance than radiologists for TNs of TR4 and TR5 categories, which are the most difficult categories for diagnosis in clinical practice. Prospective studies with long-term follow-up will be needed to examine the utility of the system and assess its effectiveness in routine clinical practice.

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 studies involving human participants were reviewed and approved by Ethics Committee of Tongji Medical College of Huazhong University of Science and Technology. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Guarantors of integrity of entire study: G-GW, W-ZL, RY, X-WC, and BZ. Literature research: G-GW, W-ZL, RY, J-YW, X-WC, and BZ. Study concepts/study design: all authors. Contributed to acquisition of data: G-GW, W-ZL, RY,Y-JY, and BZ. Clinical studies: G-GW, RY, J-WX, Y-JY, R-XC, X-WC, and BZ. Contributed reagents/materials/analysis tools: G-GW, W-ZL. Manuscript drafting or manuscript revision: all authors. Statistical analysis: G-GW, W-ZL, RY, J-WX, Y-JY, R-XC, X-WC, and BZ. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported Natural Science Foundation of Hubei Province (2019CFB286), Natural Science Foundation of Hunan Province (2017JJ2394), Corps Science and Technology Key Project (2019DB012).

ACKNOWLEDGMENTS

The authors thank ultrasound doctors in Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College for data collection. The authors also thank Shu-Kun He and Yu-Qian Jia, MD, Tongji Medical College, for assisting with preprocessing of the imaging data used and Qiu-Yu Cheng MD, Tongji Medical College for assisting with calculating the statistical data.

REFERENCES

1. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very High Prevalence of Thyroid Nodules Detected by High Frequency (13 MHz) Ultrasound Examination. *Eur J Clin Invest* (2009) 39(8):699–706. doi: 10.1111/j.1365-2362.2009.02162.x
2. Lew JI, Solorzano CC. Use of Ultrasound in the Management of Thyroid Cancer. *Oncol* (2010) 15(3):253–8. doi: 10.1634/theoncologist.2009-0324
3. Burman KD, Wartofsky L. Clinical Practice. Thyroid Nodules. *New Engl J Med* (2015) 373(24):2347–56. doi: 10.1056/NEJMcp1415786
4. Siegel RL, Miller KD. Cancer Statistics, 2019. *Cancer Stat* (2019) 69: (1):7–34. doi: 10.3322/caac.21551
5. Tessler FN, Middleton WD, Grant EG. Thyroid Imaging Reporting and Data System (Ti-Rads): A User's Guide. *Radiology* (2018) 287(3):1082. doi: 10.1148/radiol.2018184008
6. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. Acr Thyroid Imaging, Reporting and Data System (Ti-Rads): White Paper of the ACR Ti-RADS Committee. *J Am Coll Radiol JACR* (2017) 14 (5):587–95. doi: 10.1016/j.jacr.2017.01.046
7. Barbosa TLM, Junior COM, Graf H, Cavalvanti T, Trippia MA, da Silveira Ugino RT, et al. Acr TI-RADS and ATA US Scores are Helpful for the Management of Thyroid Nodules With Indeterminate Cytology. *BMC Endocrine Disord* (2019) 19(1):112. doi: 10.1186/s12902-019-0429-5
8. Gao L, Xi X, Jiang Y, Yang X, Wang Y, Zhu S, et al. Comparison Among TIRADS (Acr TI-RADS and KWAK- Ti-RADS) and 2015 ATA Guidelines in the Diagnostic Efficiency of Thyroid Nodules. *Endocrine* (2019) 64(1):90–6. doi: 10.1007/s12020-019-01843-x
9. Hong HS, Lee JY. Diagnostic Performance of Ultrasound Patterns by K-TIRADS and 2015 ATA Guidelines in Risk Stratification of Thyroid Nodules and Follicular Lesions of Undetermined Significance. *AJR Am J Roentgenol* (2019) 213(2):444–50. doi: 10.2214/AJR.18.20961
10. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, et al. Comparison of Performance Characteristics of American College of Radiology Ti-Rads, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association Guidelines. *AJR Am J Roentgenol* (2018) 210 (5):1148–54. doi: 10.2214/AJR.17.18822
11. Lauria Pantano A, Maddaloni E, Briganti SI, Beretta Anguissola G, Perrella E, Taffon C, et al. Differences Between ATA, AACE/ACE/AME and ACR Ti-RADS Ultrasound Classifications Performance in Identifying Cytological High-Risk Thyroid Nodules. *Eur J Endocrinol* (2018) 178(6):595–603. doi: 10.1530/EJE-18-0083
12. Wei X, Li Y, Zhang S, Gao M. Meta-Analysis of Thyroid Imaging Reporting and Data System in the Ultrasonographic Diagnosis of 10,437 Thyroid Nodules. *Head Neck* (2016) 38(2):309–15. doi: 10.1002/hed.23878
13. Xu T, Wu Y, Wu RX, Zhang YZ, Gu JY, Ye XH, et al. Validation and Comparison of Three Newly-Released Thyroid Imaging Reporting and Data Systems for Cancer Risk Determination. *Endocrine* (2019) 64: (2):299–307. doi: 10.1007/s12020-018-1817-8
14. Liu T, Guo Q, Lian C, Ren X, Liang S, Yu J, et al. Automated Detection and Classification of Thyroid Nodules in Ultrasound Images Using Clinical-Knowledge-Guided Convolutional Neural Networks. *Med Image Anal* (2019) 58:101555. doi: 10.1016/j.media.2019.101555
15. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. *J Am Coll Radiol JACR* (2019) 16(9 Pt B):1318–28. doi: 10.1016/j.jacr.2019.06.004
16. Buda M, Wildman-Tobriner B. Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists. *Radiology* (2019) 292: (3):695–701. doi: 10.1148/radiol.2019181343
17. Song W, Li S, Liu J, Qin H, Zhang B, Zhang S, et al. Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition. *IEEE J Biomed Health Inf* (2019) 23(3):1215–24. doi: 10.1109/JBHI.2018.2852718
18. Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, et al. An Improved Deep Learning Approach for Detection of Thyroid Papillary Cancer in Ultrasound Images. *Sci Rep* (2018) 8(1):6600. doi: 10.1038/s41598-018-25005-7
19. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, et al. Diagnosis of Thyroid Cancer Using Deep Convolutional Neural Network Models Applied to Sonographic Images: A Retrospective, Multicohort, Diagnostic Study. *Lancet Oncol* (2019) 20(2):193–201. doi: 10.1016/S1470-2045(18)30762-9
20. Cantisani V, David E, Grazhdani H, Rubini A, Radzina M, Dietrich CF, et al. Prospective Evaluation of Semiquantitative Strain Ratio and Quantitative 2d Ultrasound Shear Wave Elastography (SWE) in Association With TIRADS Classification for Thyroid Nodule Characterization. *Ultraschall der Med (Stuttgart Germany 1980)* (2019) 40(4):495–503. doi: 10.1055/a-0853-1821
21. Lee JH, Baek JH, Kim JH, Shim WH, Chung SR, Choi YJ, et al. Deep Learning-Based Computer-Aided Diagnosis System for Localization and Diagnosis of Metastatic Lymph Nodes on Ultrasound: A Pilot Study. *Thyroid Off J Am Thyroid Assoc* (2018) 28(10):1332–8. doi: 10.1089/thy.2018.0082
22. Jiang M, Li C, Tang S, Lv W, Yi A, Wang B, et al. Nomogram Based on Shear-Wave Elastography Radiomics can Improve Preoperative Cervical Lymph Node Staging for Papillary Thyroid Carcinoma. *Thyroid Off J Am Thyroid Assoc* (2020) 30(6):885–97. doi: 10.1089/thy.2019.0780
23. Chaigneau E, Russ G, Royer B, Bigorgne C, Bienvenu-Perrard M, Rouxel A, et al. TIRADS Score is of Limited Clinical Value for Risk Stratification of Indeterminate Cytological Results. *Eur J Endocrinol* (2018) 179(1):13–20. doi: 10.1530/EJE-18-0078
24. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. (2016). Learning Deep Features for Discriminative Localization. 2016 Ieee Conference on Computer Vision and Pattern Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA. pp. 2921–9. doi: 10.1109/CVPR.2016.319
25. Zhou LQ, Wu XL. Lymph Node Metastasis Prediction From Primary Breast Cancer Us Images Using Deep Learning. *Radiology* (2020) 294: (1):19–28. doi: 10.1148/radiol.2019190372
26. Pitoia F, Miyauchi A. 2015 American Thyroid Association Guidelines for Thyroid Nodules and Differentiated Thyroid Cancer and Their Implementation in Various Care Settings. *Thyroid Off J Am Thyroid Assoc* (2016) 26(2):319–21. doi: 10.1089/thy.2015.0530
27. Dighe M, Barr R, Bojunga J, Cantisani V, Chammas MC, Cosgrove D, et al. Thyroid Ultrasound: State of the Art Part 1 - Thyroid Ultrasound Reporting and Diffuse Thyroid Diseases. *Med Ultrasonography* (2017) 19(1):79–93. doi: 10.11152/mu-980
28. Dietrich CF, Müller T, Bojunga J, Dong Y, Mauri G, Radzina M, et al. Statement and Recommendations on Interventional Ultrasound as a Thyroid Diagnostic and Treatment Procedure. *Ultrasound Med Biol* (2018) 44(1):14–36. doi: 10.1016/j.ultrasmedbio.2017.08.1889
29. Dighe M, Barr R, Bojunga J, Cantisani V, Chammas MC, Cosgrove D, et al. Thyroid Ultrasound: State of the Art. Part 2 - Focal Thyroid Lesions. *Med Ultrasonography* (2017) 19(2):195–210. doi: 10.11152/mu-999
30. Trimboli P, Dietrich CF, David E, Mastroeni G, Ventura Spagnolo O, Sidhu PS, et al. Ultrasound and Ultrasound-Related Techniques in Endocrine Diseases. *Minerva Endocrinologica* (2018) 43(3):333–40. doi: 10.1016/j.ultrasmedbio.2017.08.1500
31. Xia J, Chen H, Li Q, Zhou M, Chen L, Cai Z, et al. Ultrasound-Based Differentiation of Malignant and Benign Thyroid Nodules: An Extreme Learning Machine Approach. *Comput Methods Programs Biomed* (2017) 147:37–49. doi: 10.1016/j.cmpb.2017.06.005
32. Gao L, Liu R, Jiang Y, Song W, Wang Y, Liu J, et al. Computer-Aided System for Diagnosing Thyroid Nodules on Ultrasound: A Comparison With Radiologist-Based Clinical Assessments. *Head Neck* (2018) 40: (4):778–83. doi: 10.1002/hed.25049
33. Seifert P. Interobserver Agreement and Efficacy of Consensus Reading in Kwak-, EU-, and ACR-thyroid Imaging Recording and Data Systems and

ATA Guidelines for the Ultrasound Risk Stratification of Thyroid Nodules. *Cancer Cytopathol* (2019) 67(1):143–54. doi: 10.1055/s-0039-1683623

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.

frontiers in Oncology

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fonc.2021.674579

Identifying Periampullary Regions in MRI Images Using Deep Learning

Yong Tang[1†], *Yingjun Zheng*[2†], *Xinpei Chen*[3], *Weijia Wang*[4], *Qingxi Guo*[5], *Jian Shu*[6*], *Jiali Wu*[7*] *and Song Su*[2*]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2] Department of General Surgery (Hepatobiliary Surgery), The Affiliated Hospital of Southwest Medical University, Luzhou, China, [3] Department of Hepatobiliary Surgery, Deyang People's Hospital, Deyang, China, [4] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, [5] Department of Pathology, The Affiliated Hospital of Southwest Medical University, Luzhou, China, [6] Department of Radiology, The Affiliated Hospital of Southwest Medical University, Luzhou, China, [7] Department of Anesthesiology, The Affiliated Hospital of Southwest Medical University, Luzhou, China

OPEN ACCESS

Edited by:
Naranamangalam Raghunathan Jagannathan,
Chettinad University, India

Reviewed by:
Amit Mehndiratta,
Indian Institute of Technology Delhi, India
Yee Kai Tee,
Tunku Abdul Rahman University, Malaysia

***Correspondence:**
Jian Shu
shujiannc@163.com
Jiali Wu
13982759964@163.com
Song Su
13882778554@163.com

[†]These authors share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *02 March 2021*
Accepted: *20 April 2021*
Published: *28 May 2021*

Citation:
Tang Y, Zheng Y, Chen X, Wang W, Guo Q, Shu J, Wu J and Su S (2021) Identifying Periampullary Regions in MRI Images Using Deep Learning. Front. Oncol. 11:674579. doi: 10.3389/fonc.2021.674579

Background: Development and validation of a deep learning method to automatically segment the peri-ampullary (PA) region in magnetic resonance imaging (MRI) images.

Methods: A group of patients with or without periampullary carcinoma (PAC) was included. The PA regions were manually annotated in MRI images by experts. Patients were randomly divided into one training set, one validation set, and one test set. Deep learning methods were developed to automatically segment the PA region in MRI images. The segmentation performance of the methods was compared in the validation set. The model with the highest intersection over union (IoU) was evaluated in the test set.

Results: The deep learning algorithm achieved optimal accuracies in the segmentation of the PA regions in both T1 and T2 MRI images. The value of the IoU was 0.68, 0.68, and 0.64 for T1, T2, and combination of T1 and T2 images, respectively.

Conclusions: Deep learning algorithm is promising with accuracies of concordance with manual human assessment in segmentation of the PA region in MRI images. This automated non-invasive method helps clinicians to identify and locate the PA region using preoperative MRI scanning.

Keywords: peri-ampullary cancer, periampullary regions, MRI, deep learning, segmentation

INTRODUCTION

The peri-ampulla (PA) region refers to the area within 2cm of the main papilla of the duodenum, including Vater ampulla, lower segment of common bile duct, opening of pancreatic duct, duodenal papilla and duodenal mucosa nearby (1–4). This region was deep and narrow in the abdomen and has many adjacent organs and blood vessels, so it is difficult to identify this area using conventional imaging examinations. At the same time, the PA region was prone to a series of diseases, including malignant tumors such as periampullary carcinoma (PAC) and benign lesions such as chronic mass pancreatitis, the inflammatory stricture of the lower of common bile duct, or the lower of common bile duct stone etc. (5, 6). The treatment and prognosis of these diseases vary differently, so accurate diagnosis of these disease has important clinical significance. However, the imaging diagnosis of this kind of disease is based on the determination of the specific location of PA region.

So far, among all these modern imaging techniques, magnetic resonance imaging (MRI) is a preferable choice to detect the diseases of the PA region for its advantages of excellent soft-tissue contrast and fewer radiation exposures (5, 7). However, the accuracy and specificity of MRI are still unsatisfying in the diagnosis of the diseases. A study has reported that the specificity of MRI was only 78.26%, while the accuracy was 89.89% in the diagnosis of PAC (5). Similarly, our previous study also found that MRI had only 87% accuracy in detecting PAC (8). For the disease in PA region, misdiagnose will lead to many adverse factors for the follow-up treatment of patients (8, 9). Therefore, it is necessary to further improve the preoperative diagnostic accuracy of the diseases in this special region. Meanwhile, the precise segmentation of PA region is the first and foundation for the accurate diagnosis.

Deep learning is an emerging sub-branch of artificial intelligence that has demonstrated transformative capabilities in many domains (10). Technically, deep learning is a type of neural network with multiple neural layers that is capable of extracting abstract representations of input data like images, videos, time series, natural languages, and texts. Recently, there is a remarkable research advance of applying deep learning in healthcare and clinical medicine (11–13). Deep learning has applications in the analysis of electronic health records, physiological data, and especially in the diagnosis of diseases using medical imaging (14). In the analysis of medical images of MRI, computed tomography (CT), X-ray, microscopy, and other images, deep learning shows promising performance in tasks like classification, segmentation, detection, and registration (15). Recently, considerable literature has grown up in analyzing image segmentation of different human organs using deep learning, such as pancreas (16), liver (17, 18), heart (19), brain (20, 21), etc. However, the PA region remains largely under-explored in medical image analysis based on advanced deep learning algorithms. Though the neural networks have been applied to classify ampullary tumors, the images were taken by endoscopic during operations rather than preoperative and non-invasive MRI or CT scanning (22). To our best knowledge, there is no reported work has been devoted to develop and evaluate deep learning methods to segment the PA region in MRI images.

Therefore, in this study, we presented a deep learning method to automatically segment the PA region in MRI images. We retrospectively collected an MRI image dataset from different types of PA region diseases to train, including PAC and non-PAC patients, so that the PA region could be accurately identified on the MRI image information of different cases. In a training-validation approach, we developed the deep learning method in the training set and validated the performance in the validation set. This would provide a basis for further research on the diagnosis of PAC.

MATERIALS AND METHODS

The overall workflow of this study was illustrated in **Figure 1**. First, patients were included, and the MRI images were obtained. Next, the PA regions were annotated in the MRI images by experts. Based on the raw images and annotation information, the deep learning segmentation algorithms were trained and evaluated in training and validation datasets, respectively. Finally, the performance was summarized and reported.

Patients Characteristics

This was a retrospective study approved by the Ethics Committee of the Affiliated Hospital of Southwest Medical University (No.KY2020157). A total of 504 patients who underwent MRI examinations in the Department of Hepatobiliary and Pancreatic Surgery of the Affiliated Hospital of Southwest Medical University were included from June 1, 2018 to May 1, 2019. In these people, 86 persons were diagnosed as peri-ampullary carcinoma through pathology after surgery or endoscopy, and the other 418 persons show no peri-ampullary lesion determined by radiologist. All patients underwent MRI examinations. The demographic and clinical characteristics of PAC and non-PAC patients were shown in **Table S1** and **Table S2**, respectively.

MRI Techniques

After 3-8 hours of fasting, patients were asked to practice their breathing techniques. MRI was performed in all patients with a 3.0-T MR equipment (Philips Achieva, Holland, Netherlands) with a quasar dual gradient system and a 16.0-channel phased-array Torso coil in the supine position. Drinking water or conventional oral medicines were not restricted. The MR scan started with the localization scan, followed by a sensitivity-encoding (SENSE) reference scan. The scanning sequences were as follows: breath-hold axial dual fast field echo (dual FFE) and high spatial resolution isotropic volume exam (THRIVE) T1-weighted imaging (T1WI), respiratory triggered coronal turbo spin echo (TSE) T2-weighted imaging (T2WI), axial fat-suppressed TSE-T2WI, single-shot TSE echo-planar imaging (EPI) diffusion-weighted imaging (DWI), and MR cholangiopancreatography (MRCP). For the dynamic contrast enhancement (DCE)-MRI, axial-THRIVE-T1WI were used. 15mL of contrast agent Gd-DTPA was injected through the antecubital vein at a speed of 2mL/s. DCE-MRI was performed in three phases, including arterial, portal, and delayed phase, and images were collected after 20s, 60s, and 180s, respectively (10). In result, among the 504 patients, 485 patients had THRIVE-T1W images (n = 5,861), and 495 patients had T2 W images (n = 2,558).

MRI Imaging Analysis

Post-processing of MRI images was performed using the Extended MR Workspace R2.6.3.1 (Philips Healthcare) with the FuncTool package. MRI showed typical PAC imaging manifestations: (1) the mass was nodular or invasive; (2) Tumour parenchyma on T1WI was equal or marginally lower signals; (3) Tumour parenchyma on T2WI was equally or slightly stronger signal; (4) DWI showed high signal intensity; (5) the mass was mild or moderate enhancement after contrast and (6) when MRCP was performed, the bile duct suddenly terminated asymmetrically and expanded proportionally (double-duct signs may occur when the lesion obstructed the ducts (8).

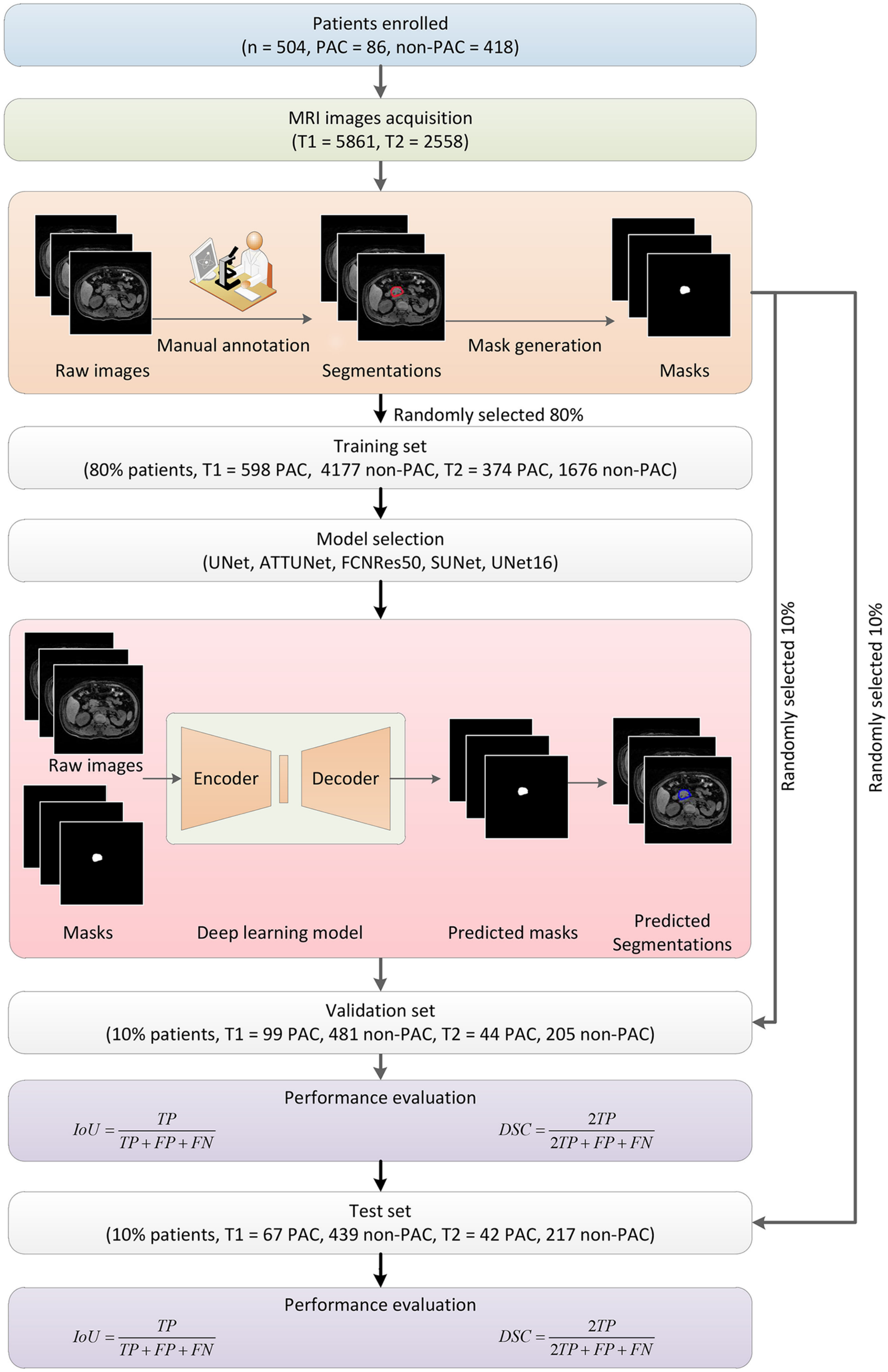

Patients enrolled
(n = 504, PAC = 86, non-PAC = 418)
MRI images acquisition
(T1 = 5861, T2 = 2558)
Manual annotation
Raw images
Segmentations
Mask generation
Masks
Randomly selected 80%
Training set
(80% patients, T1 = 598 PAC, 4177 non-PAC, T2 = 374 PAC, 1676 non-PAC)
Model selection
(UNet, ATTUNet, FCNRes50, SUNet, UNet16)
Raw images
Encoder
Decoder
Masks
Deep learning model
Predicted masks
Predicted Segmentations
Randomly selected 10%
Randomly selected 10%
Validation set
(10% patients, T1 = 99 PAC, 481 non-PAC, T2 = 44 PAC, 205 non-PAC)
Performance evaluation
$IoU = \frac{TP}{TP + FP + FN}$
$DSC = \frac{2TP}{2TP + FP + FN}$
Test set
(10% patients, T1 = 67 PAC, 439 non-PAC, T2 = 42 PAC, 217 non-PAC)
Performance evaluation
$IoU = \frac{TP}{TP + FP + FN}$
$DSC = \frac{2TP}{2TP + FP + FN}$

FIGURE 1 | Overall flowchart of this study. First, MRI images were obtained from enrolled patients and manually annotated by experts to obtain the masks for later deep learning algorithm development. The dataset was randomly divided into subsets for algorithm training, validation, and testing, respectively. Five models were developed and evaluated, and the UNet16 and FCNRes50 achieved the best performance.

Pathological Examination

The pathological data from all of the cases were analyzed by two pathologists with more than 15 years of diagnostic experience. The pathologists were blinded to the clinical and imaging findings.

Image Annotation

First, all MRI images were annotated by two experienced radiologists using in-house software. In the annotation, one radiologist was required to manually draw the outlines of the PA regions in the MRI images. The outline information was used to generate a corresponding mask image in the same size to indicate the segmentation and of the PA region. An expert radiologist reviewed all manual annotations to ensure the quality of the annotations, which served as ground truths to develop and validate deep learning algorithms (23–26).

Among the 504 patients, 485 patients had T1 images (n = 5,861), and 495 patients had T2 images (n = 2,558) were processed separately. We developed algorithms for three cases, namely using only T1, only T2, and combination of T1 and T2. In a cross-validation approach, we first randomly divided the patients into three independent cohorts, namely one training cohort (80%), one validation cohort (10%), and one test cohort (10%). Their images and corresponding annotated mask images were also accordingly grouped into one training set, one validation set, and one test set, respectively. In other words, the MRI images and the corresponding mask images of the training cohort were used to train deep learning algorithms, and those images of the validation and test cohort were later used to select and evaluate the performance of deep learning algorithms.

Deep Learning Methods

In this study, we developed deep learning algorithms using multiple layers of convolutional neural network (CNN) to automatically segment PA regions in MRI images. CNN is usually utilized to extract hierarchical patterns from images in a feedforward manner. CNN-based deep learning algorithms have achieved remarkable performance in many computer vision applications surpassing human experts (10). In medical image analysis, UNet adopted a two-block structure utilizing multiple layers of CNN (27). More specifically, the architecture consisted of two components. Namely, one encoder transformed the high dimensional input images into low dimensional abstract representations, and one following decoder projected the low dimensional abstract representations back to the high dimensional space by reversing the encoding. Finally, generated images were output with pixel-level label information indicating the PA region. The detailed structures were illustrated in **Figure S1A** for UNet16 and **Figure S1B** for FCNRes50, respectively. In order to systematically investigate the performance of the deep learning approach, in this study, we also considered another four structure variations, namely ATTUNet using the attention gate approach in UNet (27), FCNRes50 using ResNet50 as the downsampling approach FCNRes50 combine residual network and fully convolutional network structures to extract pixel-level information and generate segmentation (28), UNet16 use VGG16 as the downsampling approach (29), and SUNet using SeLu as the nonlinear activation function instead of ReLu.

In the deep learning algorithm training stage, the MRI images of the training cohort were input into the encoder one by one. The output masks generated by the decoder were compared against the corresponding ground truth to calculate the loss function, which indicated the deviations of predicted segmentation. By using the back-propagation technique of stochastic gradient descent optimization, the encoder-decoder structure was continuously optimized to minimize the loss. More technically, the weights between neural network layers were adjusted to improve the capability of segmentations. Once the training started, both the encoder and decoder were all trained together. In this manner, a satisfying deep learning neural network could hopefully be obtained after training with enough training samples. Meanwhile, since the input and output were both images, this deep learning approach enjoyed significant advantages over the conventional image analysis methods by eliminating the exhausting feature engineering or troublesome manual interferences. After the training stage, the trained encoder-decoder structure was used in passive inferences to predict PA regions in MRI images. In inferences, the weights were kept unchanged. In the validation stage, the MRI images of the validation set were input into the neural network, and the corresponding mask images were obtained. The images of the test cohort were used in evaluating the performance of the selected best model. We systematically considered four different variations of the UNet structures and one FCNRes50 structure to seek the best performing deep learning structure. Deep learning algorithms were trained, validated, and tested separately using respective images. The five models were trained, validated, and tested in the dataset contained both T1 and T2 images.

All programs were implemented in Python programming language (version 3.7) with freely available open-source packages, including Opencv-Python (version 4.1.0.25) for image and data processing, Scipy (version 1.2.1) and Numpy (version 1.16.2) for data management, Pytorch (version 1.1) for deep learning framework, Cuda (version 10.1) for graphics processing unit (GPU) support. The training and validation were conducted in a computer installed with an NVIDIA 3090Ti deep learning GPU, 24GB main memory, and Intel(R) Xeon(R) 2.10GHz central processing unit (CPU). It is worth mentioning that the validation task could be done using a conventional personal computer within an acceptable time since the passive inference requires fewer computations.

Statistical Evaluation of Segmentation

The performance of the segmentation task for the PA region in MRI images was quantitatively evaluated using intersection over union (IoU) and Dice similarity coefficient (DSC). For one PA region instance in an MRI image, the manually annotated ground truth and the deep learning predicted segmentation were compared at pixel-level to see how the two regions overlapped. In general, larger values of IoU and DSC indicated better segmentation accuracies. The average IoU and DSC were calculated based on predictions for all images in the validation set. For simplicity, we used IoU as the main measurement, and the performance of five deep learning structures was ranked according to IoU. The predictions of T1 and T2 MRI images were conducted separately in the same manner.

RESULTS

MRI Images

In preparing the training, validation, test datasets, we divided the initial dataset based on patients to ensure that images from a given patient would only appear in one dataset. In result, for T1 images (n = 5,861), the training set included 598 images from 67 PAC patients, and 4,177 images from 322 patients without PAC. The validation set included 99 images from 8 PAC patients, and 418 images from 40 patients without PAC. The test set included 67 images from 8 PAC patients, and 439 images from 40 patients without PAC. For T2 images (n = 2,558), the training set included 374 images from 68 PAC patients, and 1,676 images from 329 patients without PAC. The validation set included 44 images from 8 PAC patients, and 205 images from 41 patients without PAC. The validation set included 42 images from 8 PAC patients, and 217 images from 41 patients without PAC. For the dataset combined T1 and T2 MRI images (n = 8,419). The training set included 959 images from 69 PAC patients, and 5,701 images from 335 patients without PAC. The validation set included 176 images from 9 PAC patients, and 806 images from 42 patients without PAC. The test set included 89 images from 8 PAC patients, and 668 images from 41 patients without PAC.

Segmentation Performance

For the five segmentation deep learning structures, we followed the same training approach in separated training, validation, and testing. Specifically, each image formed a batch (batch size = 1), and ten rounds were repeated (epoch = 10) to ensure the convergence of the loss. The optimizer of all models is Adam, with a learning rate of 0.0001. The final segmentation performance of all five structures was presented in **Table 1** for T1 images, **Table 2** for T2 images, and **Table 3** for T1 and T2 images, respectively. We found that UNet16 outperformed all the rest structures with the best performance for both of only T1 (IoU = 0.68, DSC = 0.79) and combined T1 and T2 (IoU = 0.64, DSC = 0.74), respectively.

TABLE 1 | Segmentation performance of deep learning structures in the test T1 images ranked by mean IoU.

Model	IoU			DSC		
	Total	PAC	non-PAC	Total	PAC	non-PAC
UNet16	**0.68 ± 0.21**	**0.67 ± 0.18**	**0.69 ± 0.21**	**0.79 ± 0.21**	**0.78 ± 0.17**	**0.79 ± 0.21**
FCNRes50	0.67 ± 0.24	0.65 ± 0.22	0.67 ± 0.24	0.77 ± 0.26	0.76 ± 0.23	0.77 ± 0.26
UNet	0.53 ± 0.33	0.37 ± 0.34	0.55 ± 0.32	0.62 ± 0.36	0.44 ± 0.39	0.64 ± 0.35
SUnet	0.49 ± 0.30	0.40 ± 0.31	0.50 ± 0.30	0.59 ± 0.34	0.50 ± 0.35	0.60 ± 0.33
ATTUnet	0.44 ± 0.32	0.31 ± 0.32	0.46 ± 0.32	0.53 ± 0.37	0.37 ± 0.38	0.55 ± 0.36

UNet16 achieved the best performance.

TABLE 2 | Segmentation performance of deep learning structures in the test T2 images ranked by mean IoU.

Model	IoU			DSC		
	Total	PAC	non-PAC	Total	PAC	non-PAC
FCNRes50	**0.68 ± 0.20**	**0.66 ± 0.18**	**0.69 ± 0.21**	**0.79 ± 0.21**	**0.78 ± 0.16**	**0.79 ± 0.21**
UNet16	0.67 ± 0.19	0.60 ± 0.21	0.68 ± 0.18	0.78 ± 0.19	0.72 ± 0.21	0.79 ± 0.18
ATTUnet	0.58 ± 0.26	0.51 ±0.29	0.60 ± 0.25	0.69 ± 0.27	0.61 ± 0.32	0.71 ± 0.26
SUnet	0.48 ± 0.25	0.52 ± 0.25	0.47 ± 0.25	0.60 ± 0.28	0.64 ± 0.28	0.59 ± 0.28
UNet	0.40 ± 0.30	0.35 ± 0.29	0.42 ± 0.30	0.50 ± 0.35	0.44 ± 0.34	0.51 ± 0.34

FCNRes50 achieved the best performance.

TABLE 3 | Segmentation performance of deep learning structures in the test T1 and T2 images ranked by mean IoU.

Model	IoU			DSC		
	Total	PAC	non-PAC	Total	PAC	non-PAC
UNet16	**0.64 ± 0.25**	**0.61 ± 0.18**	**0.65 ± 0.25**	**0.74 ± 0.26**	**0.74 ± 0.18**	**0.74 ± 0.27**
FCNRES50	0.55 ± 0.30	0.47 ± 0.27	0.56 ± 0.30	0.64 ± 0.33	0.59 ± 0.30	0.65 ± 0.33
ATTUnet	0.45 ± 0.34	0.34 ± 0.32	0.46 ± 0.34	0.53 ± 0.38	0.42 ± 0.36	0.54 ± 0.38
SUnet	0.40 ± 0.33	0.28 ± 0.31	0.41 ± 0.33	0.48 ± 0.37	0.34 ± 0.36	0.50 ± 0.37
UNet	0.35 ± 0.35	0.21 ± 0.29	0.37 ± 0.36	0.42 ± 0.40	0.27 ± 0.34	0.43 ± 0.40

UNet16 achieved the best performance.

The performance of FCNRes50 is better than UNet16 in only T2 (IoU = 0.68, DSC = 0.79) images segmentation. As shown in the tables, the performance of patients with PAC and patients without PAC is calculated, respectively. **Figure 2** demonstrated the segmentation samples obtained by UNet16 for T1 images, FCNRes50 for T2 images, and UNet16 for combined T1 and T2 images. In terms of speed, the algorithms could output the segmentation for a given image within two seconds, which significantly improved the efficiency of image analysis.

DISCUSSION

PAC occurs in 5% of gastrointestinal tumors, and pancreatic cancer is the most common, followed by distal cholangiocarcinoma (2, 30). Pancreatoduodenectomy (PD) was the standard treatment for patients with PAC (31). However, complications such as pancreatic fistula, biliary fistula, infection, and hemorrhage often occur after PD surgery. A previous study has shown that the incidence of postoperative complications of PD may be as high as 30-65% (32). For patients with benign lesions, unnecessary PD surgery could lead to the occurrence of these surgical complications in patients, or even death in some patients. Meanwhile, if malignant lesions are misdiagnosed as benign lesions, it will undoubtedly delay the treatment of patients, resulting in poor prognosis. Due to the anatomical complexity of the periampullary region and less of particular serum markers, the early-accurate diagnose of PAC still remains challenging. Currently, non-invasive diagnostic methods, including ultrasound scan, CT imaging as well as MRI, have been successfully applied to the detection and diagnosis of PAC. One study has reported that the specificity of ultrasound scan was only 52.1%, while the accuracy was 61.61% in the diagnosis of PAC (6). Another study has reported that the specificity of CT was only 16.7%, while the accuracy was 84.4% in the diagnosis of PAC (33). So far, among all these modern imaging techniques, MRI has been reported to be an optimal choice for allowing assessment of periampullary lesions (32). However, there are still limiting factors in the evaluation of the disease using MRI because the PA region is small and the relatively complicated anatomy. Moreover, the tapered area of the distal biliary and pancreatic ducts contain little or no fluid. Physiologic contraction of the sphincter of Oddi also makes it difficult to evaluate the PA region (34). Recently, with the significant development in deep learning and increasing medical needs, artificial intelligence technology has significant advantages in improving the diagnosis of diseases. Therefore, we proposed and developed a deep learning method to automatically segment the PA region in MRI, which could be further extended to future AI-based diagnosis of the disease in PA region using AI, and also facilitate the plan of surgery and endoscopic treatment for clinicians.

In this work, we developed deep learning structures to automatically segment the PA region using MRI T1 and T2 images. Recently, there were abundant reported studies developing AI algorithms for segmentation of abdominal organs or structures including pancreas (16), liver (17, 18), spleen (35, 36), gallbladder (37), kidney (38, 39), the local lesions of stomach (40),

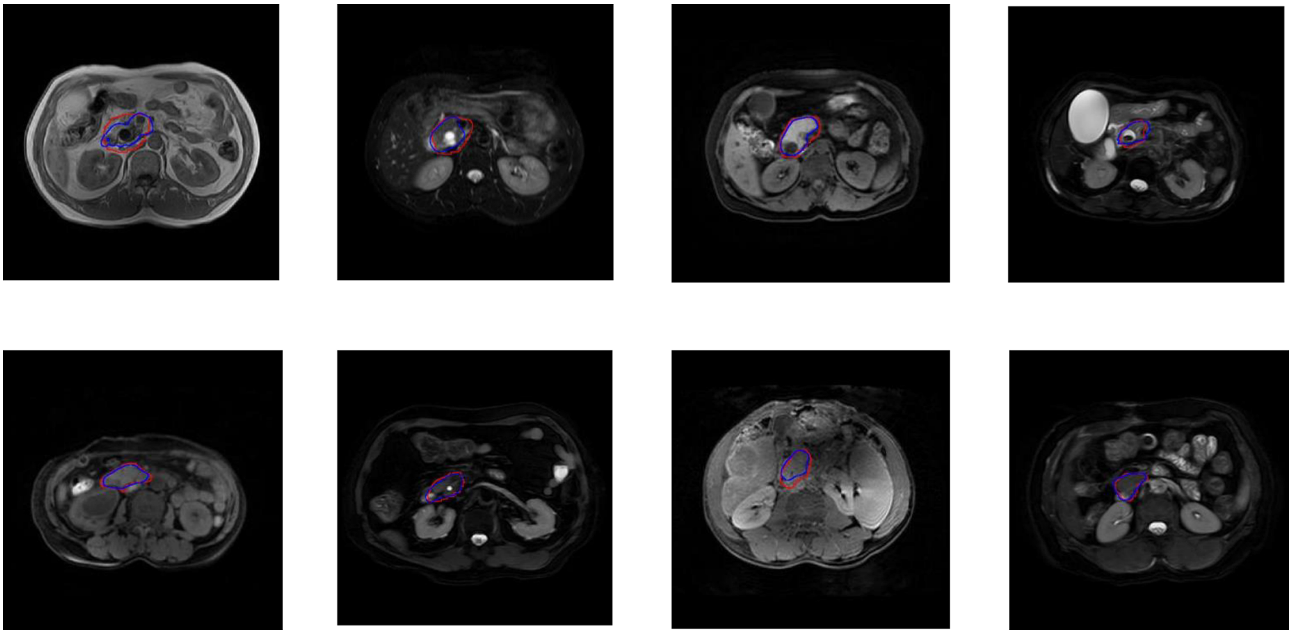


FIGURE 2 | Examples of PA regions of PAC patients (top panel) and PA regions of patients without PAC (bottom panel). The first column were examples of T1 MRI image obtained by UNet16 trained using only T1 images, the second column were examples of T2 MRI image obtained by FCNRes50 trained using only T2 images, the third column were examples of T1 MRI image obtained by UNet16 trained using both T1 and T2 images, and the fourth column were examples of T2 MRI image obtained by UNet16 trained using both T1 and T2 images. Blue, algorithm; red, expert.

etc. However, there is no report of PA region segmentation using AI algorithms. To our best knowledge, this work is the first systematic study of developing and evaluating deep learning approaches for the segmentation of the PA regions in MRI. To evaluate the performance of various deep learning structures, we implemented five algorithms that appeared in deep learning literature, including UNet (27), ATTUNet (41), FCNRes50 (28), UNet16 (29), and SUNet. UNet was the most used deep learning structure in medical image analysis using the encoder and decoder components based on CNN (42). The rest variations improve the UNet structures with attention or replace nonlinear activation functions. This study considered these structures and compared their performance in the same datasets.

In total, 504 patients were included in this study and 5,861 T1 images and 2,558 T2 images were collected. All images were manually annotated by experts to delineate the PA regions in the MRI images. By dividing patients into training and validation cohorts, their images were split into a training set for algorithms training and a validation set for final performance evaluation. As a result, UNet16 achieved the best performance among the five structures with the highest IoU of 0.68 and DSC of 0.79 for T1 images. The model with the best performance for T2 images segmentation is FCNRes50 with an IoU of 0.68 and DSC of 0.79. UNet16 achieved the best performance in the dataset of combined T1 and T2. The IoU is 0.64 and the highest DSC is 0.74 which are not better than the results obtained in the independent T1 or T2 datasets. Therefore, the results showed that UNet16 and FCNRes50 were able to accurately identify the PA region in MRI images.

However, there are still several limitations in this study. First, we only focused on developing an AI to automated localize and segment the PA regions in MRI of PA cancer, but did not make a diagnosis. In the future, we would collect more data and extend the present deep learning framework to classify and diagnose PA cancer. Second, this is a retrospective study from a single hospital, which may inevitably lead to selective bias for the patients. The results need to be validated by prospective and external cohorts. Third, the applied AI technologies in this study are still in rapid evolution with more emerging advanced deep learning algorithms. In the future, it's necessary to evaluate new deep learning algorithms in PA cancer image analysis to achieve better performance.

In conclusion, we established an MRI image dataset, developed an MRI image data annotation system, established an automatic deep learning the PA region image segmentation model, and realized the location of the PA region.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

This was a retrospective study approved by the Ethics Committee of the Affiliated Hospital of Southwest Medical University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

YT, YZ, XC, WW, QG, JS, JW, and SS conceived and designed the study, and were responsible for the final decision to submit for publication. All authors contributed to the article and approved the submitted version.

FUNDING

This study is supported by the Innovation Method Program of the Ministry of Science and Technology of the People's Republic of China (M112017IM010700), The Key Research and Development Project of Science & Technology Department of Sichuan Province (20ZDYF1129), The Applied Basic Research Project of Science & Technology Department of Luzhou city (2018-JYJ-45).

ACKNOWLEDGMENTS

The authors thank the Department of Radiology and the Department of Pathology of the Affiliated Hospital of Southwest Medical University for the support of this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.674579/full#supplementary-material

Supplementary Figure 1 | Schematic diagram of the proposed deep learning algorithms UNet16 **(A)** and FCNRes50 **(B)**. UNet16 is based on an Encoder-Decoder architecture. The encoder was a down-sampling stage, while the decoder was an up-sampling stage. FCNRes50 combine residual network and fully convolutional network structures to extract pixel-level information and generate segmentation. Images and ground truth masks were input into the network to obtain the predicted segmentation.

REFERENCES

1. Berberat PO, Künzli BM, Gulbinas A, Ramanauskas T, Kleeff J, Müller MW, et al. An Audit of Outcomes of a Series of Periampullary Carcinomas. *Eur J Surg Oncol* (2009) 35(2):187–91. doi: 10.1016/j.ejso.2008.01.030
2. Heinrich S, Clavien PA. Ampullary Cancer. *Curr Opin Gastroen* (2010) 26 (3):280–5. doi: 10.1097/MOG.0b013e3283378eb0
3. Bronsert P, Kohler I, Werner M, Makowiec F, Kuesters S, Hoeppner J, et al. Intestinal-Type of Differentiation Predicts Favourable Overall Survival: Confirmatory Clinicopathological Analysis of 198 Periampullary

Adenocarcinomas of Pancreatic, Biliary, Ampullary and Duodenal Origin. *BMC Cancer* (2013) 13:428. doi: 10.1186/1471-2407-13-428

4. Baghmar S, Agrawal N, Kumar G, Bihari C, Patidar Y, Kumar S, et al. Prognostic Factors and the Role of Adjuvant Treatment in Periampullary Carcinoma: A Single-Centre Experience of 95 Patients. *J Gastrointest Cancer* (2019) 50(3):361–9. doi: 10.1007/s12029-018-0058-7
5. Zhang T, Su ZZ, Wang P, Wu T, Tang W, Xu EJ, et al. Double Contrast-Enhanced Ultrasonography in the Detection of Periampullary Cancer: Comparison With B-Mode Ultrasonography and MR Imaging. *Eur J Radiol* (2016) 85(11):1993–2000. doi: 10.1016/j.ejrad.2016.08.021
6. Hester CA, Dogeas E, Augustine MM, Mansour JC, Polanco PM, Porembka MR, et al. Incidence and Comparative Outcomes of Periampullary Cancer: A Population-Based Analysis Demonstrating Improved Outcomes and Increased Use of Adjuvant Therapy From 2004 to 2012. *J Surg Oncol* (2019) 119(3):303–17. doi: 10.1002/jso.25336
7. Sugita R, Furuta A, Ito K, Fujita N, Ichinohasama R, Takahashi S. Periampullary Tumors: High-Spatial-Resolution MR Imaging and Histopathologic Findings in Ampullary Region Specimens. *Radiology* (2004) 231(3):767–74. doi: 10.1148/radiol.2313030797
8. Chen XP, Liu J, Zhou J, Zhou PC, Shu J, Xu LL, et al. Combination of CEUS and MRI for the Diagnosis of Periampullary Space-Occupying Lesions: A Retrospective Analysis. *BMC Med Imaging* (2019) 19(1):77. doi: 10.1186/s12880-019-0376-7
9. Schmidt CM, Powell ES, Yiannoutsos CT, Howard TJ, Wiebke EA, Wiesenauer CA, et al. Pancreaticoduodenectomy: A 20-Year Experience in 516 Patients. *Arch Surg (Chicago Ill 1960)* (2004) 139(7):718–25, 725-7. doi: 10.1001/archsurg.139.7.718
10. Lecun Y, Bengio Y, Hinton G. Deep Learning. *NATURE* (2015) 521 (7553):436–44. doi: 10.1038/nature14539
11. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *N Engl J Med* (2017) 376(26):2507–9. doi: 10.1056/NEJMp1702071
12. Hinton G. Deep Learning—a Technology With the Potential to Transform Health Care. *JAMA* (2018) 320(11):1101–2. doi: 10.1001/jama.2018.11100
13. Stead WW. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. *JAMA* (2018) 320(11):1107–8. doi: 10.1001/jama.2018.11029
14. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A Guide to Deep Learning in Healthcare. *Nat Med* (2019) 25(1):24–9. doi: 10.1038/s41591-018-0316-z
15. Litjens G, Kooi T, Bejnordi B, Arindra A, Setio A, Ciompi F, et al. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
16. Roth HR, Lu L, Lay N, Harrison AP, Farag A, Sohn A, et al. Spatial Aggregation of Holistically-Nested Convolutional Neural Networks for Automated Pancreas Localization and Segmentation. *Med Image Anal* (2018) 45:94–107. doi: 10.1016/j.media.2018.01.006
17. Chung M, Lee J, Park S, Lee CE, Lee J, Shin YG. Liver Segmentation in Abdominal CT Images Via Auto-Context Neural Network and Self-Supervised Contour Attention. *Artif Intell Med* (2021) 113:102023. doi: 10.1016/j.artmed.2021.102023
18. Kushnure DT, Talbar SN. MS-Unet: A Multi-Scale Unet With Feature Recalibration Approach for Automatic Liver and Tumor Segmentation in CT Images. *Computerized Med Imaging Graphics* (2021) 89:101885. doi: 10.1016/j.compmedimag.2021.101885
19. Bai W, Shi W, O'Regan D, Tong T, Wang H, Jamil-Copley S, et al. A Probabilistic Patch-Based Label Fusion Model for Multi-Atlas Segmentation With Registration Refifinement: Application to Cardiac Mr Images. *IEEE TMI* (2013) 32(7):1302–15. doi: 10.1109/TMI.2013.2256922
20. Wang H, Suh J, Das S, Pluta J, Craige C, Yushkevich P. Multiatlas Segmentation With Joint Label Fusion. *IEEE PAMI* (2013) 35:611–23. doi: 10.1109/TPAMI.2012.143
21. Wang L, Shi F, Li G, Gao Y, Lin W, Gilmore J, et al. Segmentation of Neonatal Brain Mr Images Using Patch-Driven Level Sets. *NeuroImage* (2014) 84 (1):141–58. doi: 10.1016/j.neuroimage.2013.08.008
22. Seo JD, Seo DW, Alirezaie J. Simple Net: Convolutional Neural Network to Perform Differential Diagnosis of Ampullary Tumors, in: *2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME); 2018 2018-01-01*. IEEE (2018). pp. 187–92.
23. Vorontsov E, Cerny M, Régnier P, Jorio L, Pal C, Lapointe R, et al. Deep Learning for Automated Segmentation of Liver Lesions At CT in Patients With Colorectal Cancer Liver Metastases. *Radiol: Artif Intell* (2019) 1:180014. doi: 10.1148/ryai.2019180014
24. Ahn Y, Yoon JS, Lee SS, Suk HII, Son JH, Sung YS, et al. Deep Learning Algorithm for Automated Segmentation and Volume Measurement of the Liver and Spleen Using Portal Venous Phase Computed Tomography Images. *Korean J Radiol* (2020) 21(8):987–97. doi: 10.3348/kjr.2020.0237
25. Chen Y, Dan R, Xiao J, Wang L, Sun B, Saouaf R, et al. Fully Automated Multi-Organ Segmentation in Abdominal Magnetic Resonance Imaging With Deep Neural Networks. *Med Phys* (2020). doi: 10.1002/MP.14429
26. Song D, Wang Y, Wang W, Wang Y, Cai J, Zhu K, et al. Using Deep Learning to Predict Microvascular Invasion in Hepatocellular Carcinoma Based on Dynamic Contrast-Enhanced MRI Combined With Clinical Parameters. *J Cancer Res Clin Oncol* (2021). doi: 10.1007/s00432-021-03617-3
27. Ronneberger O, Fischer P, Brox T. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; 2015 2015-01-01*. N Navab, J Hornegger, W Wells, A Frangi, editors. Springer International Publishing (2015). p. 234–41.
28. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE T Pattern Anal* (2017) 39(4):640–51. doi: 10.1109/TPAMI.2016.2572683
29. Pravitasari AA, Iriawan N, Almuhayar M, Azmi T, Fithriasari K, Purnami SW, et al. UNet-VGG16 With Transfer Learning for MRI-Based Brain Tumor Segmentation. *Telkomnika* (2020) 18:1310–8. doi: 10.12928/TELKOMNIKA.v18i3.14753
30. Albores-saavedra J, Schwartz AM, Batich K, Henson DE. Cancers of the Ampulla of Vater: Demographics, Morphology, and Survival Based on 5,625 Cases From the SEER Program. *J Surg Oncol* (2009) 100(7):598–605. doi: 10.1002/jso.21374
31. Winter JM, Brennan MF, Tang LH, D'Angelica MI, Dematteo RP, Fong Y, et al. Survival After Resection of Pancreatic Adenocarcinoma: Results From a Single Institution Over Three Decades. *Ann Surg Oncol* (2012) 19(1):169–75. doi: 10.1245/s10434-011-1900-3
32. Hill JS, Zhou Z, Simons JP, Ng SC, McDade TP, Whalen GF, et al. A Simple Risk Score to Predict in-Hospital Mortality After Pancreatic Resection for Cancer. *Ann Surg Oncol* (2010) 17(7):1802–7. doi: 10.1245/s10434-010-0947-x
33. Hashemzadeh S, Mehrafsa B, Kakaei F, Javadrashid R, Golshan R, Seifar F, et al. Diagnostic Accuracy of a 64-Slice Multi-Detector CT Scan in the Preoperative Evaluation of Periampullary Neoplasms. *J Clin Med* (2018) 7 (5):7. doi: 10.3390/jcm7050091
34. Kim JH, Kim M-J, Chung J-J, Lee WJ, Yoo HS, Lee JT. Differential Diagnosis of Periampullary Carcinomas at MR Imaging. *Radiographics* (2002) 22 (6):1335–52. doi: 10.1148/rg.226025060
35. Moon H, Huo Y, Abramson RG, Peters RA, Assad A, Moyo TK, et al. Acceleration of Spleen Segmentation With End-to-End Deep Learning Method and Automated Pipeline. *Comput Biol Med* (2019) 107:109–17. doi: 10.1016/j.compbiomed.2019.01.018
36. Yucheng T, Yuankai H, Yunxi X, Hyeonsoo M, Albert A, Tamara KM, et al. Improving Splenomegaly Segmentation by Learning From Heterogeneous Multi-Source Labels. *Proc SPIE Int Soc Opt Eng* (2019) 10949. doi: 10.1117/12.2512842
37. Lian J, Ma Y, Ma Y, Shi B, Liu J, Yang Z, et al. Automatic Gallbladder and Gallstone Regions Segmentation in Ultrasound Image. *Int J Comput Assist Radiol Surg* (2017) 12(4):553–68. doi: 10.1007/s11548-016-1515-z
38. Park J, Bae S, Seo S, Park S, Bang J-I, Han JH, et al. Measurement of Glomerular Filtration Rate Using Quantitative SPECT/CT and Deep-Learning-Based Kidney Segmentation. *Sci Rep* (2019) 9(1):4223. doi: 10.1038/s41598-019-40710-7
39. Onthoni DD, Sheng T-W, Sahoo PK, Wang L-J, Gupta P. Deep Learning Assisted Localization of Polycystic Kidney on Contrast-Enhanced CT Images. *Diagn (Basel)* (2020) 10(12):25. doi: 10.3390/diagnostics10121113
40. Huang L, Li M, Gou S, Zhang X, Jiang K. Automated Segmentation Method for Low Field 3D Stomach MRI Using Transferred Learning Image Enhancement Network. *BioMed Res Int* (2021) 2021:6679603. doi: 10.1155/2021/6679603
41. Lian S, Luo Z, Zhong Z, Lin X, Su S, Li S. Attention guided U-Net for accurate iris segmentation. *J Vis Commun Image Represent* (2018) 56:296–304. doi: 10.1016/j.jvcir.2018.10.001.

42. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation* (1989) 1:541–51. doi: 10.1162/neco.1989.1.4.541

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.

frontiers in Oncology

ORIGINAL RESEARCH
published: 11 June 2021
doi: 10.3389/fonc.2021.654685

Deep Learning on Enhanced CT Images Can Predict the Muscular Invasiveness of Bladder Cancer

Gumuyang Zhang[1†], Zhe Wu[2†], Lili Xu[1], Xiaoxiao Zhang[1], Daming Zhang[1], Li Mao[3], Xiuli Li[3], Yu Xiao[4], Jun Guo[2], Zhigang Ji[5], Hao Sun[1‡] and Zhengyu Jin[1‡]*

[1] Department of Radiology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China, [2] Department of Radiology, Fushun Central Hospital of Liaoning Province, Fushun, China, [3] Deepwise Artificial Intelligence (AI) Lab, Deepwise Inc., Beijing, China, [4] Department of Pathology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China, [5] Department of Urology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

OPEN ACCESS

Edited by:
Sebastian Cerdan,
Autonomous University of Madrid,
Spain

Reviewed by:
Ying-Shi Sun,
Peking University Cancer Hospital,
China
Zhen Li,
Huazhong University of Science and
Technology, China

****Correspondence:***
Hao Sun
sunhao_robert@126.com

†These authors have contributed equally to this work and share first authorship

‡These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *17 January 2021*
Accepted: *27 May 2021*
Published: *11 June 2021*

Citation:
Zhang G, Wu Z, Xu L, Zhang X, Zhang D, Mao L, Li X, Xiao Y, Guo J, Ji Z, Sun H and Jin Z (2021) Deep Learning on Enhanced CT Images Can Predict the Muscular Invasiveness of Bladder Cancer. Front. Oncol. 11:654685. doi: 10.3389/fonc.2021.654685

Background: Clinical treatment decision making of bladder cancer (BCa) relies on the absence or presence of muscle invasion and tumor staging. Deep learning (DL) is a novel technique in image analysis, but its potential for evaluating the muscular invasiveness of bladder cancer remains unclear. The purpose of this study was to develop and validate a DL model based on computed tomography (CT) images for prediction of muscle-invasive status of BCa.

Methods: A total of 441 BCa patients were retrospectively enrolled from two centers and were divided into development (n=183), tuning (n=110), internal validation (n=73) and external validation (n=75) cohorts. The model was built based on nephrographic phase images of preoperative CT urography. Receiver operating characteristic (ROC) curves were performed and the area under the ROC curve (AUC) for discrimination between muscle-invasive BCa and non-muscle-invasive BCa was calculated. The performance of the model was evaluated and compared with that of the subjective assessment by two radiologists.

Results: The DL model exhibited relatively good performance in all cohorts [AUC: 0.861 in the internal validation cohort, 0.791 in the external validation cohort] and outperformed the two radiologists. The model yielded a sensitivity of 0.733, a specificity of 0.810 in the internal validation cohort and a sensitivity of 0.710 and a specificity of 0.773 in the external validation cohort.

Conclusion: The proposed DL model based on CT images exhibited relatively good prediction ability of muscle-invasive status of BCa preoperatively, which may improve individual treatment of BCa.

Keywords: bladder cancer, deep learning, computed tomography, diagnosis, computed-assisted, artificial intelligence

INTRODUCTION

Bladder cancer (BCa) is one of the most common and lethal malignancies worldwide (1, 2). Clinical treatment decision making primarily relies on the absence or presence of muscle invasion and tumor staging (3). Nonmuscle-invasive BCa (NMIBC) and muscle-invasive BCa (MIBC) exhibit significant differences in prognosis, management and therapeutic aims (3, 4). Accurate preoperative assessment of the muscular invasiveness of BCa is crucial for selecting the optimal therapy for individual patients.

Cystoscopy examination together with histological evaluation of the resected tissues is the mainstay of diagnosis and clinical staging of BCa. As biopsy is operator dependent and unlikely to sample every part of the tumor, incorrect staging occurs, and up to 25% of MIBC cases are initially misdiagnosed as NIMBC (5, 6). Repeated examinations could improve the diagnostic accuracy, but the invasive nature has made this process undesirable. Developing a noninvasive method for preoperative evaluation would greatly benefit BCa patients. Computed tomography (CT) imaging has been widely used to preoperatively evaluate BCa patients and assist in tumor staging, especially for T3 and T4 tumors (7). Given its inability to differentiate among layers of the bladder wall, the role of traditional CT in the classification of NIMBC and MIBC is limited. Thus, developing a technique that could provide additional information about the status of muscular invasion of BCa would enable traditional CT to play a larger role in BCa evaluation and assist in patient management.

Deep learning (DL) is a novel and promising technique that has demonstrated great potential in disease diagnosis (8–10). DL can extract and combine features from images to construct a model that reveals the relationship between images and diseases. It has been reported that the DL model could facilitate imaging diagnosis in various diseases with high accuracy, including liver fibrosis, pancreatic cancer and pulmonary nodules (9, 11, 12). For BCa, the DL model based on CT images has demonstrated the potential to assist in therapy evaluation (13). However, the use of the DL based on CT images to discriminate between MIBC and NIBC has not yet been reported.

Therefore, the aim of this study was to develop and validate a DL model based on CT images for individualized prediction of the muscle-invasive status of BCa preoperatively.

MATERIALS AND METHODS

Study Population

This retrospective study was approved by the Institutional Review Board of the two medical centers, and the requirement of informed consent was waived. The inclusion criteria were as follows: (i) patients who underwent transurethral resection of bladder tumor (TURBT) or radical cystectomy in the two centers with pathologically confirmed urothelial carcinoma and (ii) availability of preoperative CT urography (CTU) within 20 days before surgery. Patients were excluded if (i) they had preoperative therapy, including chemotherapy or radiotherapy; (ii) they had other tumors simultaneously; (iii) their TURBT specimens had no muscle after resection; or (iv) no visible tumor was detected on preoperative enhanced pelvic CT images. Two radiologists (H.S. in Center 1 and Z.W. in Center 2) identified patients according to the above criteria, and 366 patients were recruited from May 2014 to July 2018 from Center 1 (91 patients with MIBC, 275 patients with NIMBC) and 75 patients from April 2018 to May 2020 in Center 2 (31 patients with MIBC, 44 patients with NIMBC). We divided the patients into three cohorts: 293 patients treated between May 2014 and September 2017 in Center 1 were allocated to the training cohort, 73 patients treated between October 2017 and July 2018 in Center 1 were allocated to the internal validation cohort, and all 75 patients treated in Center 2 constituted the external cohort. The training cohort was further randomly assigned into a development set (n=183) for model training and a tuning set (n=110) for model selection. The study flow and recruitment pathway are presented in **Figure 1**.

Clinical-pathologic information, including age, sex and pathologic T stage, was obtained from medical records. Two experienced radiologists (6 and 14 years of experience in in urogenital imaging) reviewed all the CT images together and recorded data, including the number of tumors, the size and the CT attenuation of the largest tumor. Any disagreement was resolved by consensus.

CT Imaging

All the enrolled patients in both centers underwent preoperative CTU with a similar protocol setup with different systems. The CT image acquisition settings are provided in **Supplementary Table S1**. Patients fasted for 4-6 hours, and then were asked to drink about 1000 ml water about 45 minutes before the scan and not to urinate until the scan was finished. Patients were scanned from the hemidiaphragm to the pelvic floor. For the contrast scans, patients were injected with 100 ml of nonionic contrast material (Ultravist 370, Bayer Schering Pharma AG, Germany) followed by a 100-ml saline chaser intravenously at a rate of 4–4.5 mL/s after the unenhanced scan. Renal corticomedullary-phase, nephrographic-phase and excretory-phase images were acquired at 25 s, 75 s and 300 s after the bolus-triggering threshold of 120 HU was achieved in the thoracoabdominal aorta junction. To show BCa lesions better, coronal and sagittal reformations were reconstructed besides axial images. But only the axial nephrographic-phase images were used for subsequent analysis.

Tumor Region Segmentation

Regions of interest (ROIs) were delineated semiautomatically on thin-slice CT images of the nephrographic phase by an experienced radiologist (G.Z., 6 years of experience in urogenital imaging and 5 years of experience in tumor segmentation) who was blinded to the pathological status of muscular invasion of lesions. For patients with multiple lesions, only the largest lesion was chosen for segmentation. A three-dimensional ROI of the whole tumor was delineated semiautomatically using the Deepwise Research Platform (Deepwise Inc., Beijing, China, http://label.deepwise.com). On the platform, a level-set-based

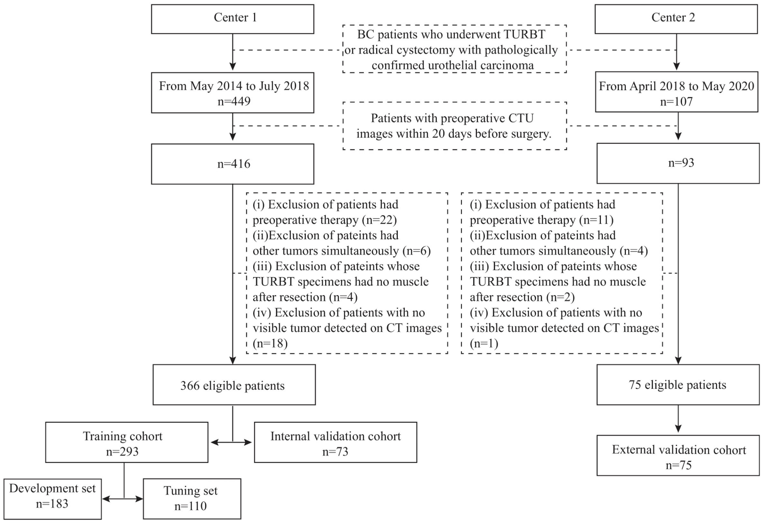

Center 1
Center 2
BC patients who underwent TURBT or radical cystectomy with pathologically confirmed urothelial carcinoma
From May 2014 to July 2018 n=449
From April 2018 to May 2020 n=107
Patients with preoperative CTU images within 20 days before surgery.
n=416
n=93
(i) Exclusion of patients had preoperative therapy (n=22)
(ii)Exclusion of pateints had other tumors simultaneously (n=6)
(iii) Exclusion of pateints whose TURBT specimens had no muscle after resection (n=4)
(iv) Exclusion of patients with no visible tumor detected on CT images (n=18)
(i) Exclusion of patients had preoperative therapy (n=11)
(ii)Exclusion of pateints had other tumors simultaneously (n=4)
(iii) Exclusion of pateints whose TURBT specimens had no muscle after resection (n=2)
(iv) Exclusion of patients with no visible tumor detected on CT images (n=1)
366 eligible patients
75 eligible patients
Training cohort n=293
Internal validation cohort n=73
External validation cohort n=75
Development set n=183
Tuning set n=110

FIGURE 1 | The study flow and the recruitment pathway. BC, bladder cancer; TURBT, transurethral resection of bladder tumor; CTU, computed tomography urography.

segmentation algorithm was initially used to outline the tumor margin automatically, and then the radiologist manually corrected the tumor margin where it was not accurate. After 8 weeks, 93 patients in the development set were selected randomly, and their tumors were segmented again by the same radiologist and another radiologist (X.Z., 1 year of experience in urogenital imaging and tumor segmentation) to evaluate intra- and interobserver reproducibility by calculating intra- and interclass dice coefficients.

Development and Validation of the Model

The pipeline of DL modeling is presented in **Figure 2**. Before the training of the model, the images were preprocessed. The voxel size was normalized to 1.0 x 1.0 x 1.0 mm3, and the pixel values were rescaled to (0,1). To further utilize the segmentation and focus the model's attention on the tumor area, the masked tumor region and the original tumor region were stacked vertically, then cropped it according to the tumor center to form an input volume of 2 x 64 x 64 x 64 for channel, depth, height and width, respectively. Our model was constructed on the basis of Filter-guided Pyramid Network (FGP-Net), a novel 3D convolutional network structure that was designed to capture the global feature and the local features simultaneously in our previous study (14). To avoid overfitting, the growth rate of the dense block was reduced to 8, and a dropout layer with a drop rate of 0.5 was added. In addition, the input patches were augmented by random cropping and rotation during the training process. The output of our model was the probability of the MIBC. Focal loss with a gamma of 1.5 and a class weight of 3 were used to manage the unbalanced amount of MIBC and NMIBC tumors. The Adam optimizer was used to minimize the focal loss with an initial learning rate of 0.001 (15). The output of our model was the probability of the MIBC, the model that achieved the highest area under the receiver operating characteristic curve (AUC) on the tuning set during the training procedure was selected, and the cut-off value was selected at the points that maximized the Youden index value on the tuning set. The AUC, accuracy, sensitivity, and specificity of all sets were calculated. The calibration curve with LOESS smoother was generated to assess the calibration of the DL model (16).

Two methods that visualizing the feature extraction process by the convolutional neural network were used to demonstrate whether the DL model learned valuable features from meaningful CT areas. First, the feature maps before discriminative filter learning modules in our model were extracted to show the target area of the model. The value of the area on the feature map indicated its contribution to the final result. The higher the value, the larger the contribution of the area. Using gamma correction (γ=2.0), the feature maps were transformed, mapped to a colored scheme and overlaid on the original images. Second, t-distributed stochastic neighborhood embedding (t-SNE), which is an unsupervised dimension-reduction algorithm to visualize high-dimensional data, was used to test the effectiveness of the learned features. In this study, t-SNE was used to reduce the dimension of features (the output of the layer before the final fully connected layer) from 150 to 2 with a learning rate of 450 and a perplexity of 30.

Subjective Image Evaluation

For subjective assessment of muscular invasion of BC based on CT images, a tumor was defined as MIBC if it invaded perivesical fat with the tumor bulging out or based on the presence of abnormal enhancement of bladder wall; otherwise, it was considered NIMBC. Examples of these imaging features were demonstrated to two radiologists (Reader 1, L.X., Reader 2, D.Z., with 3 and 9 years of experience in CTU, respectively) before they started the review process. The two radiologists reviewed all the images in validation cohorts (n=148) and determined whether the tumor was MIBC or NMIBC independently,

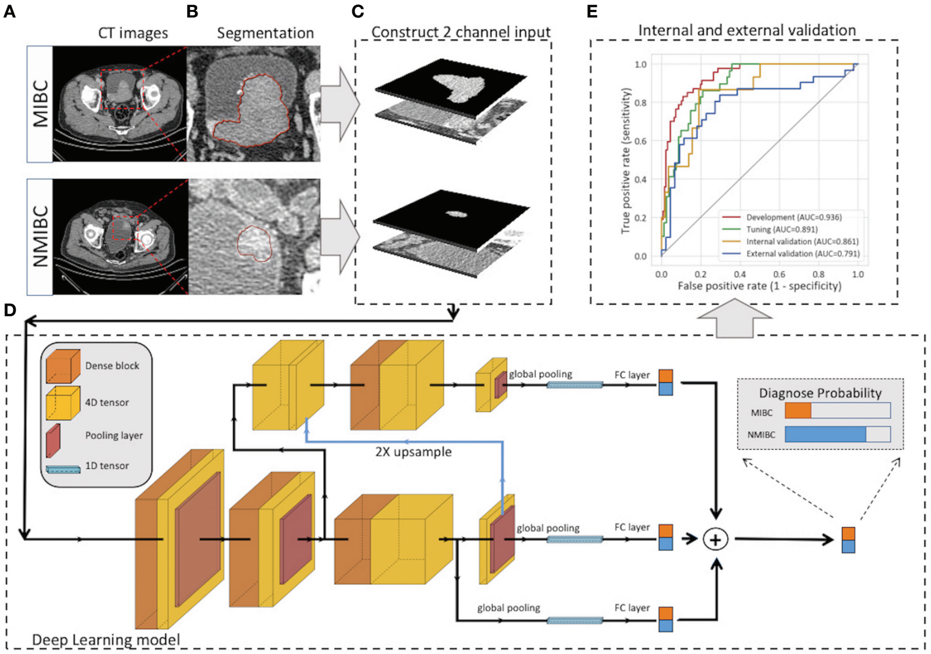


FIGURE 2 | Workflow of the deep learning model for the prediction of muscle invasiveness status in bladder cancer patients. **(A)** Collection of the CT images of MIBC and NMIBC. **(B)** Semiautomatic segmentation of the tumor region. **(C)** The masked tumor region and the original tumor region were stacked vertically to form the input volume, and the cropped 2-channel input was constructed. **(D)** The structure of our deep-learning model. The model was constructed on the basis of Filter-guided Pyramid Network (FGP-Net), a novel 3D convolutional network structure that is designed to capture the global feature and the local features simultaneously. **(E)** Internal and external validation of our model. CT, computed tomography; FC, fully connected layer.

without knowledge of pathological information (including the status of muscular invasiveness of tumors). For patients with multiple tumors, only the largest tumor was evaluated. The performance of the two radiologists for diagnosing MIBC was evaluated by calculating accuracy, sensitivity and specificity.

Statistical Analysis

A two-sided $P<0.05$ indicated statistically significant differences. Analysis of variance or Kruskal-Wallis H test was used to compare clinical characteristics among development, tuning, internal and external validation cohorts. These statistical analyses were performed by using SPSS version 25.0 (IBM, SPSS; Chicago, IL, USA). The comparison of the AUC was calculated by the DeLong test (17) which was performed by using R (version 3.6.0). The ROC curves, decision curve analysis (DCA) and calibration curves were calculated using scikit-learn (version 0.22.1) and matplotlib (version 3.1.3).

RESULTS

Patient Clinical Characteristics

Patient characteristics in all the cohorts are shown in **Table 1**. No significant differences in gender or CT-reported largest lesion diameter (P > 0.05) were noted among the training, internal validation and external validation cohorts. Patient age, CT-reported number of lesions, CT attenuation of the largest lesion and pT stage were significantly different. The proportion

TABLE 1 | Clinical characteristics of patients with bladder cancer.

Characteristics	Training cohort* (n=293)	Internal validation cohort (n=73)	External validation cohort (n=75)	p-value
Age				0.038
Median (IQR)	65 (56,72)	68 (61,74)	65 (59,77)	
Gender				0.166
Female	75 (25.6)	13 (17.8)	13 (17.3)	
Male	218 (74.4)	60 (82.2)	62 (82.7)	
CT-reported number of lesions				0.016
Unifocal	229 (78.2)	66 (90.4)	54 (72.0)	
Multifocal	64 (21.8)	7(9.6)	21 (28.0)	
CT-reported largest lesion diameter (cm)				0.063
Mean ± SD	2.71 ± 1.67	2.33 ± 1.62	2.78 ± 1.70	
≤3	188 (64.2)	57 (78.1)	52 (69.3)	
>3	105 (35.8)	16 (21.9)	23 (30.7)	
CT attenuation of the largest lesion (HU)				0.030
Mean ± SD	67.1 ± 14.0	56.3 ± 20.9	70.5 ± 13.0	
Pathologic T stage				0.010
≤T1	217 (74.1)	58 (79.5)	44 (58.7)	
≥T2	76 (25.9)	15 (20.5)	31 (41.3)	

**The training cohort (n=293) is the combination of the development (n=183) and tuning (n=110) cohorts. IQR, interquartile; SD, standard deviation.*

of MIBC was significantly increased in the external validation cohort ($P = 0.010$).

The Performance Assessment and the Clinical Usefulness of the Model

For the semiautomatic segmented ROI, the intraclass dice coefficient (0.800 ± 0.201) indicating favorable reproducibility, while the interclass dice was relatively low (0.706 ± 0.253). The ROC curves of the DL model are presented in **Figure 3A**. The model produced satisfactory performance in the development (AUC 0.936) and tuning (AUC 0.891) cohorts. The AUC in the internal validation cohort and the external validation cohort reached 0.861 (95% CI: 0.765, 0.957) and 0.791 (95% CI: 0.678, 0.904), respectively, demonstrating good differentiating ability between MIBC and NMIBC and good model robustness. The cut-off value that maximized the Youden index was 0.337. The performance of our model for differentiating between MIBC and NMIBC on development and tuning sets is also summarized in **Table 2**.

The calibration curves of the model exhibited good agreement between the model predicted outcome and the real status of muscular invasiveness (**Figure 3C**). The DCA indicated that the DL model could add more benefit to patients than the "treat all" or "treat none" strategies when the threshold probability was ranged from 0 to 0.74 in the internal validation cohort and 0.21 to 0.79 in the external validation cohort (**Figures 3D, E**).

The Comparison With Radiologists

In the subjective assessment of muscular invasion of BCa, the two radiologists generally performed slightly worse compared

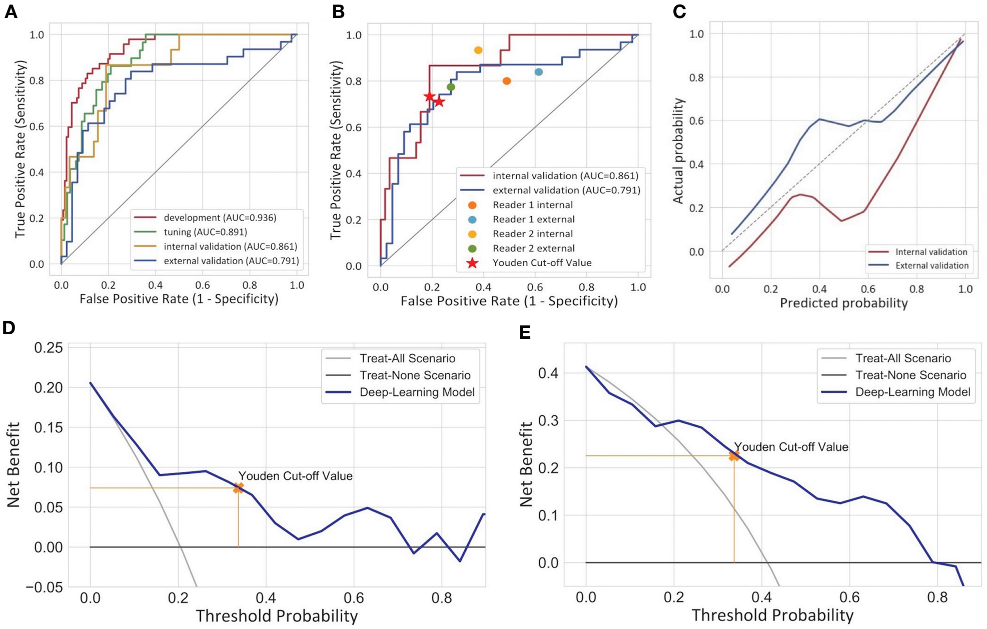


FIGURE 3 | Performance of the deep learning model for the differentiation of MIBC and NMIBC. **(A)** Receiver operator characteristic curves of the model in four different cohorts. **(B)** Comparison of the performance between the model and two radiologists. **(C)** Calibration curves of the model in internal and external validation cohorts. The calibration curve showed that the predicted probabilities generally agreed with the observed probabilities. The predictive performance of the model in the external validation cohort exhibited a closer fit to the perfect calibration. **(D, E)** showed decision curve analyses (DCA) in the internal and external validation cohorts respectively. DCA compared the net benefit of the deep learning model versus treat all or treat none are shown. The net benefit was plotted versus the threshold probability. The net benefits of the deep learning model (blue line) were superior to the benefits of treating all or treating none.

TABLE 2 | Performance of the model in development, tuning and validation cohorts.

	AUC (95%CI)	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
Development cohort (n=183)	0.936 (0.901, 0.971)	0.836 (0.773, 0.885)	0.872 (0.736, 0.947)	0.824 (0.747, 0.882)
Tuning cohort (n=110)	0.891 (0.832, 0.950)	0.800 (0.711, 0.868)	0.828 (0.635, 0.935)	0.790 (0.683, 0.87)
Internal validation cohort (n=73)	0.861 (0.765, 0.957)	0.795 (0.681, 0.877)	0.733 (0.448, 0.911)	0.810 (0.682, 0.897)
External validation cohort (n=75)	0.791 (0.678, 0.904)	0.747 (0.631, 0.837)	0.710 (0.518, 0.851)	0.773 (0.618, 0.880)

AUC, area under the receiver operating characteristics curve; CI, confidence interval.

with the DL model (**Table 3** and **Figure 3B**). In the internal validation cohort, the accuracy and specificity of Reader 1 (0.685 and 0.621) and Reader 2 (0.585 and 0.517) were lower than those of the model (0.795 and 0.810), while the sensitivity of them (0.933 and 0.800) exceeded that of the model (0.733). In the external validation cohort, Reader 1 demonstrated comparable performance compared to the model with the same accuracy (0.747) and similar specificity (0.727 *vs* 0.773) and sensitivity (0.774 *vs* 0.710). However, the performance of Reader 2 was inferior to the model in general with a lower accuracy (0.573 *vs* 0.747) and specificity (0.386 *vs* 0.773) but higher sensitivity (0.839 *vs* 0.710).

Additional Analysis

Violin plots of the predicted score for muscle invasion in the development, tuning, internal and external cohorts are shown in **Figure 4A**. NMIBC patients had significantly lower predicted scores than those with MIBC in the development (median 0.214 [interquartile range 0.136-0.291] *vs* 0.813 [0.607, 0.938], *P*< 0.001), tuning (0.225 [0.161, 0.304] *vs* 0.539 [0.385, 0.846], *P* < 0.001), internal validation (0.216 [0.172, 0.288] *vs* 0.422 [0.327, 0.843], *P* < 0.001) and external validation (0.184, [0.124, 0.305] *vs* 0.759 [0.307, 0.889], *P* < 0.001) cohorts. The waterfall plots in **Figures 4B, C** illustrate the distribution of the predicted score and the status of muscular invasion of individual patients in the internal and external validation cohorts, respectively.

We used feature maps and t-SNE to visualize the learned features. **Figure 5A** demonstrates feature maps of four examples (two for MIBC and two for NMIBC) from the external validation cohort. The focus area of the model or the active area is illustrated by bright colors. These regions represent different characteristics of lesions and were in accord with human observations, and the models would aid in the classification of lesions. T-SNE visualization demonstrated that the learned features of the DL model can distinguish MIBC and NMIBC. The locations of BC lesions depended on the similarity of their features. They were close to each other if they had similar features; otherwise, they were far apart. As shown in **Figure 5B**, MIBC and NMIBC clusters were basically separated except for several outliners, demonstrating that the developed model has captured effective features for differentiation.

TABLE 3 | Performance of two radiologists and the deep learning model on validation cohorts.

Validation cohort	Reader	Accuracy (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
Internal	Reader 1	0.685 (0.564, 0.786)	0.933 (0.660, 0,.997)	0.621 (0.483, 0.742)
	Reader 2	0.585 (0.454, 0.688)	0.800 (0.514, 0.947)	0.517 (0.383, 0.649)
	Model	0.795 (0.681, 0.877)	0.733 (0.448, 0.911)	0.810 (0.682, 0.897)
External	Reader 1	0.747 (0.631, 0.837)	0.774 (0.585, 0.897)	0.727 (0.570, 0.845)
	Reader 2	0.573 (0.454, 0.685)	0.839 (0.655, 0.939)	0.386 (0.247, 0.545)
	Model	0.747 (0.631, 0.837)	0.710 (0.518, 0.851)	0.773 (0.618, 0.880)

AUC, area under the receiver operating characteristics curve; CI, confidence interval.

DISCUSSION

The aim of this double-center study was to predict the muscular invasiveness of bladder cancer based on enhanced CT images. Our DL model exhibited relatively good performance to discriminate NMIBC from MIBC. The AUC was 0.861 in the internal validation cohort and 0.791 in the external validation cohort.

Preoperative evaluation of muscle invasion in bladder cancer is important for patient management. Currently, transurethral resection of bladder tumor is the standard for preoperative T staging evaluation (3, 7, 18–21). As the procedure highly depends on surgeon experience and biopsy quality, its diagnostic accuracy for MIBCs varies. MRI is also recommended, and the Vesical Imaging-Reporting and Data System based on multiparametric MRI has been proposed for the diagnosis of MIBCs (22). But it is still a subjective evaluation process based on the experience of radiologists. In recent years, researchers have investigated alternative techniques to assist muscle invasiveness evaluation. Garapati et al. (23) explored machine learning methods to discriminate between MIBC and NMIBC in 84 BC lesions from 76 CTU cases retrospectively. They found that morphological and texture features achieved comparable performance with AUCs of about 0.90. Some other studies developed MRI-based radiomic models for preoperative prediction of the muscle-invasive status of BCa with AUCs ranging from 0.87 to 0.98 (24–27). These studies revealed encouraging results for avoiding subjectivity in the preoperative assessment of BCa, but external validation in larger cohorts is required to verify the clinical validity of these new techniques. In contrast to the above studies, we investigated the feasibility of using DL on CT images to differentiate between MIBC and NMIBC. We used a well-designed deep learning structure, which utilizes the dense block and the pyramid structure to extract the features effectively and integrate the global features and the local features (14). Considering the relatively small sample size, several methods were utilized to alleviate the problem of overfitting, including reducing the growth rate of the dense block and data augmentation. The focal loss, which is designed to handle the imbalance of the data amount and the difficulty, was employed (15). Regarding diagnostic performance, the AUCs in this study were slightly lower than those in other studies. When we analyzed what went wrong and why, we found that most true NMIBC cases that were mistakenly identified as MIBC were large (typically >4 cm), and almost all the MIBC cases falsely recognized as NIMBC were small (typically < 1 cm). These findings suggest that the DL model considers the tumor size as one of the key features to determine the muscle-invasive status of BCa.

In general, the DL model outperformed the two radiologists in terms of accuracy, and the DL model also demonstrated increased specificity. But the DL model exhibited reduced sensitivity. This finding may be explained by the fact that radiologists are more prone to suspect a tumor to be muscularly invasive due to their fear of the negative consequences of missing MIBC. Moreover,

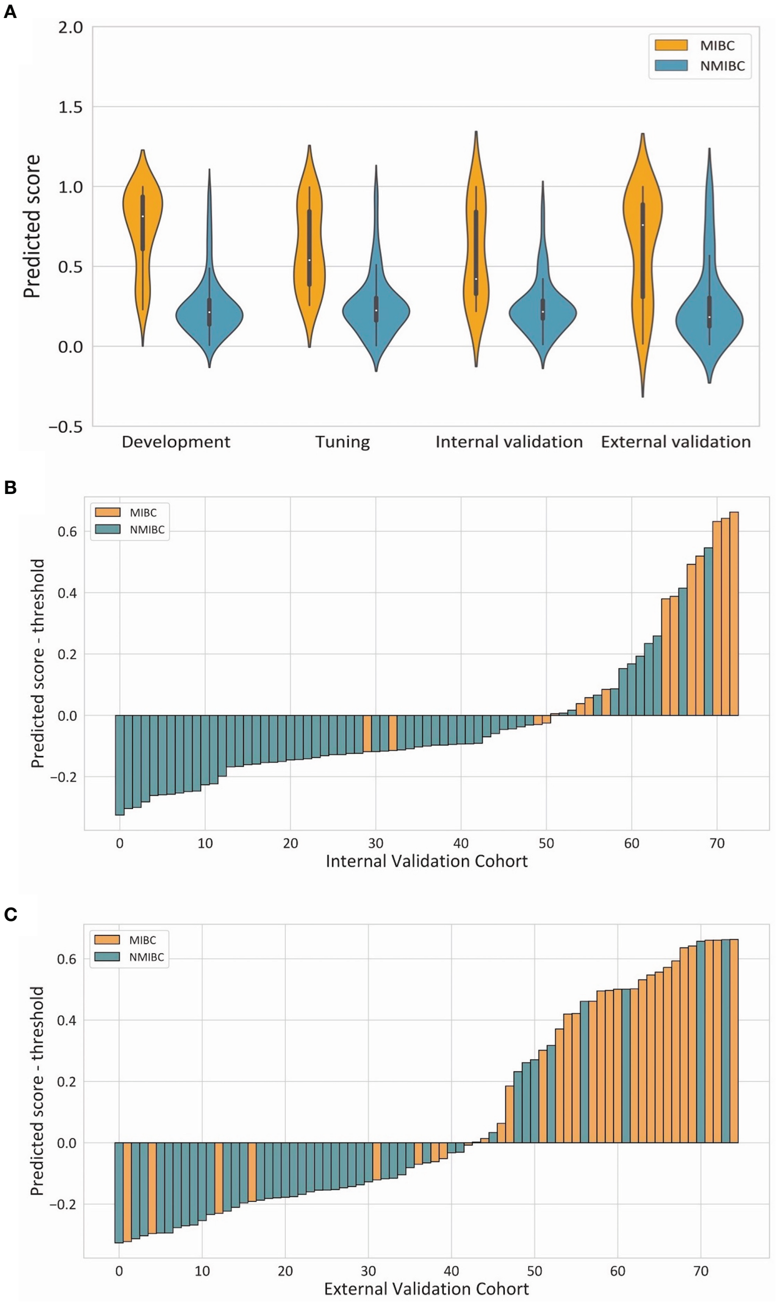


FIGURE 4 | Illustrations of the performance of the deep learning model. **(A)** Violin plots of predictive scores in the development, tuning, internal validation and external validation cohorts. **(B, C)** showed waterfall plots of the distribution of predictive scores and muscle invasive status of each patient in the internal and external validation cohorts respectively.

surprisingly, Reader 1, who had less experience in urogenital imaging, demonstrated better performance than Reader 2. Thus, a radiologist's experience may not necessarily have a positive correlation with prediction accuracy. On the other hand, our results also indicated that the DL model could produce a more stable, objective and balanced outcome for discrimination between MIBC and NMIB compared to subjective assessment by radiologists.

ROI segmentation is an essential part of the research process. Currently, 3D segmentation of the whole tumor is widely adopted by researchers because it is thought to provide a more comprehensive evaluation compared to one ROI from the largest cross-sectional area of the tumor. Researchers typically need to manually draw the outline of the tumor on each image slice, which is time consuming, especially when the study population is large. Automated segmentation has been proposed, but the

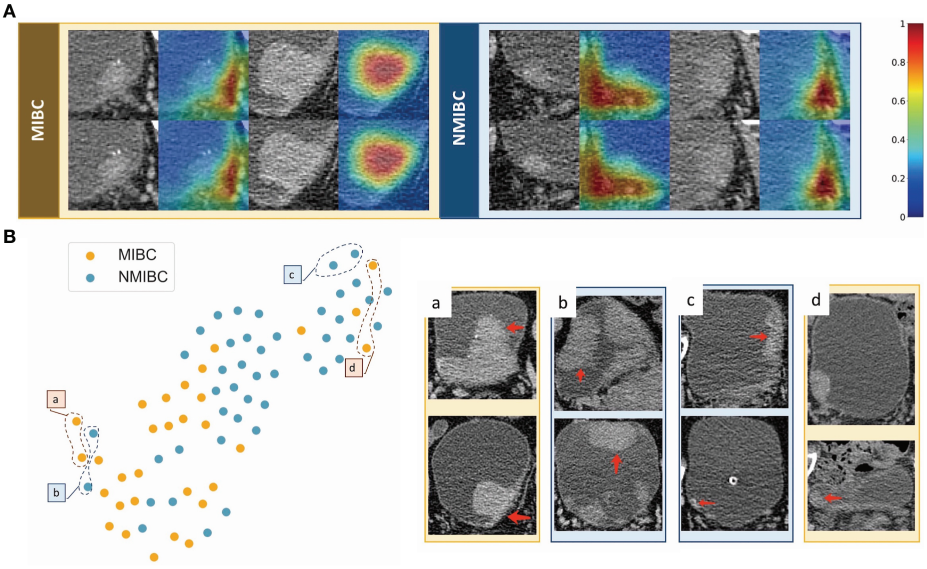


FIGURE 5 | Examples of feature maps from validation cohorts and visualization of the effectiveness of the learned features. **(A)** Two cases from MIBC and two cases from NMIBC are shown. The active regions were mainly overlaid on the areas with visual characteristics that were helpful for discriminating between MIBC and NMIBC, including the internal region of the tumor, corresponding bladder wall, and the surrounding outside pelvic fat. **(B)** Colored points represent the NIMBC (blue) and MIBC (orange). Effective features were learned by the model, and the two categories of nodules were well clustered. The eight examples show images corresponding to circled points. Nodules in sets a and c were highly discriminated by the model, whereas nodules in sets b and d were less discriminated because they shared similar features with the opposite tumor.

accuracy for BCa remains unclear. In this study, we applied a semiautomatic approach to segment each tumor. This method is a combination of automated segmentation by the platform and small modification by the radiologist. According to our experience, this semiautomatic method not only greatly accelerates the study process but also ensures the accuracy of ROI delineation. The interclass dice coefficient of 0.706 for ROI segmentation was slightly low. We analyzed those significantly different segmentations between the two radiologists and found that the radiologist with less experience mistakenly identified BCa lesions in patients with an irregular bladder shape or with prostate hyperplasia. This radiologist also failed to correctly segment some BCa lesions that presented as abnormal enhancement of the focal bladder wall. This result reminded us of the importance of the experience and the training of radiologists for ROI segmentation to reach a solid and reliable result.

Although the results were not very satisfactory, our study still has several strengths. First, this study explored the capacity of a DL model based on CT images to determine the status of muscle invasiveness of BCa, which provided a basis for subsequent studies to apply this technique to tackle relevant clinical problems. Second, unlike some other studies that used cross validation or single-center validation, this study used an external validation cohort enrolled from a different hospital, which allowed us to investigate the generalizability of the DL model. In addition, the study population of this study was larger than many other studies focusing on the application of machine learning in BCa. Third, CT-related studies of discriminating MIBC from NMIBC are limited. There is no doubt that CT has its limitations due to its low resolution of soft tissue. However, our study indicated that with the help of novel techniques, such as DL, we can also obtain valuable information from routine CT images to guide patient therapy. Thus, it is still worth performing CT-based studies to solve clinical problems in BCa management.

Our study has some limitations. First, this is a two-center study, but the number of patients in Center 2 is relatively small. Multicenter studies with larger population or prospective clinical trials should be conducted to validate the results in the future. Second, the proposed DL model exhibited its potential but the performance was less than satisfactory and has yet to be improved. Constant efforts should be made to optimize the model before it could be applied in real clinical practice. Third, the model was based on visible tumors on enhanced CT given that we excluded tumors detected by cystoscopy but invisible on CT images. Although these tumors constitute a small proportion of BCa, they may still limit the scope of the model's application to some extent. Fourth, in this study, we did not incorporate other clinical information which may be helpful for determining the invasiveness of BCa, such as urine DNA or RNA. We aimed to investigate the potential of deep learning to facilitate CT evaluation of BCa, thus we focused on CT images only. It's possible that integrating those useful clinical information into the model might further improve the prediction accuracy. Fifth, we only chose the largest one among multiple lesions for segmentation and there is a chance that the largest one didn't have the highest T stage. But usually larger lesions are supposed to have higher T stage, and it's very difficult to make one-to-one correspondence between the lesion on CT images and the lesion pathologist evaluated, we think choosing the largest one for analysis is acceptable.

In conclusion, we developed a DL model based on enhanced CT images to predict muscle invasiveness of BCa. This model should favorable performance. It could provide more useful information for individual preoperative evaluation, may facilitate clinical decision making and improve patient care.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of Peking Union Medical College Hospital and Fushun Central Hospital of Liaoning Province. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

GZ: conception and design, acquisition of data, analysis and interpretation of data, manuscript drafting and revision, and statistical analysis. ZW: acquisition of data, manuscript drafting and revision, and administrative and material support. LX: acquisition of data. XZ: analysis and interpretation of data. DZ: analysis and interpretation of data. YX: acquisition of data. LM: analysis and interpretation of data, and statistical analysis. XL: conception and design, manuscript drafting and revision, and administrative and material support. JG: acquisition of data. ZhiJ: administrative and material support. HS: conception and design, manuscript drafting and revision, administrative and material support, and supervision. ZheJ: administrative and material support, and supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China [81901742, 91859119], the Natural Science Foundation of Beijing Municipality [7192176], the Clinical and Translational Research Project of Chinese Academy of Medical Sciences [XK320028], and the National Public Welfare Basic Scientific Research Project of Chinese Academy of Medical Sciences [2018PT32003, 2019PT320008].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.654685/full#supplementary-material

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2020. *CA Cancer J Clin* (2020) 70:7–30. doi: 10.3322/caac.21590
2. Svatek RS, Hollenbeck BK, Holmang S, Lee R, Kim SP, Stenzl A, et al. The Economics of Bladder Cancer: Costs and Considerations of Caring for This Disease. *Eur Urol* (2014) 66:253–62. doi: 10.1016/j.eururo.2014.01.006
3. Spiess PE, Agarwal N, Bangs R, Boorjian SA, Buyyounouski MK, Clark PE, et al. Bladder Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Cancer Network* (2017) 15:1240–67. doi: 10.6004/jnccn.2017.0156
4. Kamat AM, Hahn NM, Efstathiou JA, Lerner SP, Malmstrom PU, Choi W, et al. Bladder Cancer. *Lancet* (2016) 388:2796–810. doi: 10.1016/S0140-6736(16)30512-8
5. Shariat SF, Palapattu GS, Karakiewicz PI, Rogers CG, Vazina A, Bastian PJ, et al. Discrepancy Between Clinical and Pathologic Stage: Impact on Prognosis After Radical Cystectomy. *Eur Urol* (2007) 51:137–49. doi: 10.1016/j.eururo.2006.05.021
6. Mariappan P, Zachou A, Grigor KM, Edinburgh Uro-Oncology G. Detrusor Muscle in the First, Apparently Complete Transurethral Resection of Bladder Tumour Specimen Is a Surrogate Marker of Resection Quality, Predicts Risk of Early Recurrence, and Is Dependent on Operator Experience. *Eur Urol* (2010) 57:843–9. doi: 10.1016/j.eururo.2009.05.047
7. Bellmunt J, Orsola A, Leow JJ, Weigel T, Santis MD, Horwich A, et al. Bladder Cancer: ESMO Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann Oncol* (2014) 25(Suppl 3):iii40–8. doi: 10.1093/annonc/mdu223
8. Zhang XY, Wang L, Zhu HT, Li ZW, Ye M, Li XT, et al. Predicting Rectal Cancer Response to Neoadjuvant Chemoradiotherapy Using Deep Learning of Diffusion Kurtosis Mri. *Radiology* (2020) 296:56–64. doi: 10.1148/radiol.2020190936
9. Liu KL, Wu T, Chen PT, Tsai YM, Wang W. Deep Learning to Distinguish Pancreatic Cancer Tissue From Non-Cancerous Pancreatic Tissue: A Retrospective Study With Cross-Racial External Validation. *Lancet Digital Health* (2020) 2:e303–13. doi: 10.1016/S2589-7500(20)30078-9
10. Dong D, Fang MJ, Tang L, Shan XH, Gao JB, Giganti F, et al. Deep Learning Radiomic Nomogram Can Predict the Number of Lymph Node Metastasis in Locally Advanced Gastric Cancer: An International Multicenter Study. *Ann Oncol* (2020) 31:912–20. doi: 10.1016/j.annonc.2020.04.003
11. Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, et al. Deep Learning Radiomics of Shear Wave Elastography Significantly Improved Diagnostic Performance for Assessing Liver Fibrosis in Chronic Hepatitis B: A Prospective Multicentre Study. *Gut* (2019) 68:729–41. doi: 10.1136/gutjnl-2018-316204
12. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-End Lung Cancer Screening With Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nat Med* (2019) 25:954–61. doi: 10.1038/s41591-019-0447-x
13. Cha KH, Hadjiiski L, Chan H-P, Weizer AZ, Alva A, Cohan RH, et al. Bladder Cancer Treatment Response Assessment in CT Using Radiomics With Deep-Learning. *Sci Rep* (2017) 7:8738. doi: 10.1038/s41598-017-09315-w
14. Huang C, Lv W, Zhou C, Mao L, Xu Q, Li X, et al. Discrimination Between Transient and Persistent Subsolid Pulmonary Nodules on Baseline CT Using Deep Transfer Learning. *Eur Radiol* (2020) 30:6913–23. doi: 10.1007/s00330-020-07071-6
15. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. *IEEE Trans Pattern Anal Mach Intell* (2020) 42:318–27. doi: 10.1109/TPAMI.2018.2858826

16. Detmer FJ, Chung BJ, Mut F, Slawski M, Hamzei-Sichani F, Putman C, et al. Development and Internal Validation of an Aneurysm Rupture Probability Model Based on Patient Characteristics and Aneurysm Location, Morphology, and Hemodynamics. *Int J Comput Assist Radiol Surg* (2018) 13:1767–79. doi: 10.1007/s11548-018-1837-0
17. DeLong ER DD, Clarke-Pearson DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* (1988) 44:837–45. doi: 10.2307/2531595
18. Power NE, Izawa J. Comparison of Guidelines on Non-Muscle Invasive Bladder Cancer (Eau, CUA, Aua, NCCN, Nice). *Bladder Cancer* (2016) 2:27–36. doi: 10.3233/BLC-150034
19. Woldu SL, Bagrodia A, Lotan Y. Guideline of Guidelines: Non-Muscle-Invasive Bladder Cancer. *BJU Int* (2017) 119:371–80. doi: 10.1111/bju.13760
20. Gurram S, Muthigi A, Egan J, Stamatakis L. Imaging in Localized Bladder Cancer: Can Current Diagnostic Modalities Provide Accurate Local Tumor Staging? *Curr Urol Rep* (2019) 20:82. doi: 10.1007/s11934-019-0948-7
21. Mirmomen SM, Shinagare AB, Williams KE, Silverman SG, Malayeri AA. Preoperative Imaging for Locoregional Staging of Bladder Cancer. *Abdom Radiol (NY)* (2019) 44:3843–57. doi: 10.1007/s00261-019-02168-z
22. Ueno Y, Takeuchi M, Tamada T, Sofue K, Takahashi S, Kamishima Y, et al. Diagnostic Accuracy and Interobserver Agreement for the Vesical Imaging-Reporting and Data System for Muscle-Invasive Bladder Cancer: A Multireader Validation Study. *Eur Urol* (2019) 76:54–6. doi: 10.1016/j.eururo.2019.03.012
23. Garapati SS, Hadjiiski L, Cha KH, Chan HP, Caoili EM, Cohan RH, et al. Urinary Bladder Cancer Staging in CT Urography Using Machine Learning. *Med Phys* (2017) 44:5814–23. doi: 10.1002/mp.12510
24. Zheng J, Kong J, Wu S, Li Y, Cai J, Yu H, et al. Development of a Noninvasive Tool to Preoperatively Evaluate the Muscular Invasiveness of Bladder Cancer Using a Radiomics Approach. *Cancer* (2019) 125:4388–98. doi: 10.1002/cncr.32490
25. Wang H, Xu X, Zhang X, Liu Y, Ouyang L, Du P, et al. Elaboration of a Multisequence MRI-Based Radiomics Signature for the Preoperative Prediction of the Muscle-Invasive Status of Bladder Cancer: A Double-Center Study. *Eur Radiol* (2020) 30:4816–27. doi: 10.1007/s00330-020-06796-8
26. Xu S, Yao Q, Liu G, Jin D, Chen H, Xu J, et al. Combining DWI Radiomics Features With Transurethral Resection Promotes the Differentiation Between Muscle-Invasive Bladder Cancer and Non-Muscle-Invasive Bladder Cancer. *Eur Radiol* (2020) 30:1804–12. doi: 10.1007/s00330-019-06484-2
27. Xu X, Zhang X, Tian Q, Wang H, Cui LB, Li S, et al. Quantitative Identification of Nonmuscle-Invasive and Muscle-Invasive Bladder Carcinomas: A Multiparametric MRI Radiomics Analysis. *J Magn Reson Imaging* (2019) 49:1489–98. doi: 10.1002/jmri.26327

Conflict of Interest: LM and XL are employees of Deepwise AI Lab, Deepwise Inc., which contributed to the development of radiomics models described in the study.

The remaining 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.

frontiers in Oncology

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fonc.2021.700158

Predicting Malignancy and Invasiveness of Pulmonary Subsolid Nodules on CT Images Using Deep Learning

Tianle Shen[1†], *Runping Hou*[1,2†], *Xiaodan Ye*[3], *Xiaoyang Li*[1], *Junfeng Xiong*[2], *Qin Zhang*[1], *Chenchen Zhang*[1], *Xuwei Cai*[1], *Wen Yu*[1], *Jun Zhao*[2*‡] and *Xiaolong Fu*[1*‡]

[1] Department of Radiation Oncology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China, [2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, [3] Department of Radiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China

Background: To develop and validate a deep learning–based model on CT images for the malignancy and invasiveness prediction of pulmonary subsolid nodules (SSNs).

Materials and Methods: This study retrospectively collected patients with pulmonary SSNs treated by surgery in our hospital from 2012 to 2018. Postoperative pathology was used as the diagnostic reference standard. Three-dimensional convolutional neural network (3D CNN) models were constructed using preoperative CT images to predict the malignancy and invasiveness of SSNs. Then, an observer reader study conducted by two thoracic radiologists was used to compare with the CNN model. The diagnostic power of the models was evaluated with receiver operating characteristic curve (ROC) analysis.

Results: A total of 2,614 patients were finally included and randomly divided for training (60.9%), validation (19.1%), and testing (20%). For the benign and malignant classification, the best 3D CNN model achieved a satisfactory AUC of 0.913 (95% CI: 0.885–0.940), sensitivity of 86.1%, and specificity of 83.8% at the optimal decision point, which outperformed all observer readers' performance (AUC: 0.846±0.031). For pre-invasive and invasive classification of malignant SSNs, the 3D CNN also achieved satisfactory AUC of 0.908 (95% CI: 0.877–0.939), sensitivity of 87.4%, and specificity of 80.8%.

Conclusion: The deep-learning model showed its potential to accurately identify the malignancy and invasiveness of SSNs and thus can help surgeons make treatment decisions.

Keywords: pulmonary subsolid nodules, computed tomography, diagnosis, computer-aided diagnosis (CAD), deep learning

OPEN ACCESS

Edited by:
Abhishek Mahajan,
Tata Memorial Hospital, India

Reviewed by:
Shouliang Qi,
Northeastern University, China
Wei Qian,
The University of Texas at El Paso,
United States

***Correspondence:**
Jun Zhao
junzhao@sjtu.edu.cn
Xiaolong Fu
xlfu1964@hotmail.com

†These authors have contributed equally to this work and share first authorship

‡These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *25 April 2021*
Accepted: *08 July 2021*
Published: *26 July 2021*

Citation:
Shen T, Hou R, Ye X, Li X, Xiong J, Zhang Q, Zhang C, Cai X, Yu W, Zhao J and Fu X (2021) Predicting Malignancy and Invasiveness of Pulmonary Subsolid Nodules on CT Images Using Deep Learning. Front. Oncol. 11:700158. doi: 10.3389/fonc.2021.700158

INTRODUCTION

Lung cancer is one of the most lethal malignancies worldwide (1). Early detection and accurate diagnosis of pulmonary nodules can decrease the mortality of lung cancer (2). According to the content of solid component, pulmonary nodules can be divided into solid nodules and subsolid nodules (SSNs). They have great difference in clinical management due to their different biological characteristics (3).

SSNs are defined as nodular areas of homogeneous or heterogeneous attenuation that did not completely cover the whole lung parenchyma within them, including pure ground-glass nodules (PGGNs) and part-solid nodules (PSNs) (4) (**Supplementary Figure S1**). According to the pathology, SSNs can be further divided into benign and malignant lesions, of which malignant SSNs include pre-invasive (atypical adenomatous hyperplasia, AAH; adenocarcinoma *in situ*, AIS; minimally invasive adenocarcinoma, MIA) and invasive lesions (invasive pulmonary adenocarcinoma, IA) (5). The three categories of SSNs have different biological characteristics and need different clinical management. Benign SSNs include hemorrhage, inflammation, fibrosis, pulmonary alveolar proteinosis, etc. (6), which need almost no intervention but only follow-up. In contrast, malignant SSNs include subtypes of adenocarcinoma, and those malignant pathological types need careful intervention, such as surgical resection and stereotactic body radiation therapy (SBRT) (7). To be specific, receiving systematic lymph node dissection has no statistical significance on improving the prognosis of patients with pre-invasive SSNs (8, 9). The pre-invasive malignant SSNs may just need to be treated with conservative approach (sub-lobectomy or wedge resection) with long-term CT follow-up, while more aggressive surgical treatment (standard lobectomy with extended lymph node dissection) is necessary for patients with invasive (IA) SSNs. Also, the prognosis of different pathological subtypes varies greatly after the corresponding treatment (10, 11). Therefore, accurate classification of SSNs has a great importance for clinical decision-making and prognosis evaluating, especially for thoracic surgeons as it determines the candidates of surgery and the type of lung resection.

Nowadays, the prevalence application of high-resolution CT scanning makes more SSNs be detected at an early stage. However, for those detected SSNs, there exist many difficulties for accurate diagnosis during clinical practice. For example, the synchronous or asynchronous appearance of multiple primary SSNs, the inappropriate location of the SSNs, and the poor physical condition of the patients make it impossible to access each SSN by biopsy. Therefore, CT imaging has become the most important method to help clinicians make the diagnostic decisions of SSNs. As reported, clinicians often make decisions according to some CT morphological features (12, 13). Nevertheless, these morphological features are subjective and qualitative, which often lead to low inter-observer agreement and unsatisfied accuracy (14–16). The inaccurate diagnosis caused by the above limitations have led to undertreatment or overtreatment for patients with SSNs in clinical practice. Therefore, a more objective and quantitative method to accurately distinguish the malignancy and invasiveness of SSNs is urgently needed.

Recently, deep learning has been widely used to analyze medical images on various image modalities (17–20). Previous studies have shown the efficiency of deep learning in pulmonary nodule detection and classification areas (21–23). However, most of these studies are based on solid nodules, and few concentrate on SSNs. Therefore, this study aims to develop and validate a deep learning–based malignancy and invasiveness prediction model in patients with SSNs from the realistic clinical cohort.

MATERIALS AND METHODS

Patients

With approval from the institutional review board, we retrospectively collected patients with pulmonary nodules in Shanghai Chest Hospital from January 1, 2012, to December 31, 2018. The inclusion criteria include the following: (1) Patients received surgical resection of pulmonary nodules in our hospital. (2) Patients received pre-surgery chest CT scanning (thickness ≤5 mm) in our hospital. (3) Subsolid nodules were confirmed in the chest CT. Patients were excluded if (1) post-surgery pathological results were not available; (2) distant metastasis was found in preoperative examinations; (3) other malignant radiological features were present including enlarged hilar nodes, pleural effusion, atelectasis, etc.

CT Image Acquisition and Nodule Segmentation

Chest CT scans were taken with a 64-detector CT row scanner (Brilliance 64; Philips, Eindhoven, Netherlands). Part of the patients conducted a target thin-section helical CT scan with layer thickness of 1 mm, while the others only had the whole lung scan with a layer thickness of 5 mm.

SSNs were manually segmented by one radiation oncologist (with 5 years of experience in CT interpretation) using the MIM software (version 5.5.1, shown with window level −400 and window width 1,600), then the region of interest (ROI) was confirmed by one radiologist (with over 10 years of experience in CT interpretation).

Image Preprocessing

The image preprocessing procedure are as follows: CT scans were converted into Hounsfield units (HU), then voxel intensity was clipped to [−1,024, 400] and [−160, 240] HU, respectively. Min-Max normalization was used to rescale the image to [0,1]. Linear interpolation was applied to get isotropic volumes with a resolution of 0.5 mm × 0.5 mm × 0.5 mm. Then, an image cube and the corresponding segmentation mask with 64 × 64 × 64 voxels were cropped from the interpolated CT image centered on the tumor. The cropped image cubes were used as the input of our 3D CNN classification model.

Pathological Information

According to the pathological report, each SSN was given a specific label (benign, AAH, AIS, MIA, IA). For the malignancy classification, patients who had at least one pathologically confirmed malignant SSN (including AAH, AIS, MIA, IA) were regarded as positive samples with label 1, and those without malignant findings were negative samples with label 0. For the invasiveness classification, patients who were pathologically confirmed as AAH, AIS, or MIA were regarded as pre-invasive samples with label 0, while patients confirmed as IA were regarded as invasive samples with label 1.

Development of the Classification Model

We respectively established a binary classifier to distinguish benign and malignant SSNs and another one to recognize pre-invasive and invasive SSNs. The framework of our models is shown in **Figure 1**.

We totally constructed three models for the malignancy and invasiveness prediction of SSNs, respectively. First, a logistic regression model built with nodule size was used as the baseline clinical model. Second, a 3D CNN model based on modified adaptive DenseNet using the lung window image as input was constructed (AdaDense) (24). The adaptive dense connected structure can effectively reuse the shallow layers' features by allowing each layer access to feature maps from all of its preceding layers, which makes it easier to get a smooth decision function with better generalization performance. However, as most of the subsolid nodules' size are small, there exist lots of noisy information from the background in the cropped image patches. Therefore, we considered incorporating the segmentation mask as attention map to help the network focus on regions within the nodule. Moreover, studies have shown that solid portions of SSNs detected by mediastinal window can help distinguish pure ground-glass nodules and part-solid nodules (25, 26), and the proportion of solid components are considered to be related with the malignancy and invasiveness classification (8, 27). Therefore, to take the segmentation mask and solid component factors into account, we finally built another 3D CNN model using the lung window image [HU: (−1,024,400)] incorporated with mediastinal window image [HU: (−160,240)] and mask image as input (AdaDense_M). Then, given the CT image of SSNs, the CNN model output the predicted probability of the SSN being malignancy or invasiveness.

The architecture of the AdaDense_M model can be seen in **Figure 1**, which consists of two parts, data fusion and main structure. For the data fusion part, the CT image patch in different windows and the corresponding segmentation mask were separately convolved by a kernel of 3×3×3 to obtain channels 1, 2, and 3, respectively. Then the three channels were concatenated together and convolved by a 3×3×3 kernel with stride=2 as the input of the main structure. This operation reduced the original feature map of 64×64×64 to the size of 32×32×32. For the main structure part, there were three dense blocks connected by transition layers. Each of the dense block contained four bottleneck structures, and after each bottleneck layer, all feature maps in the previous layers were adaptively concentrated together to realize feature reuse. The bottleneck layer can reduce the number of input feature maps, thereby improving the computational efficiency. The transition layer further compressed parameters by reducing half of the feature maps after dense blocks.

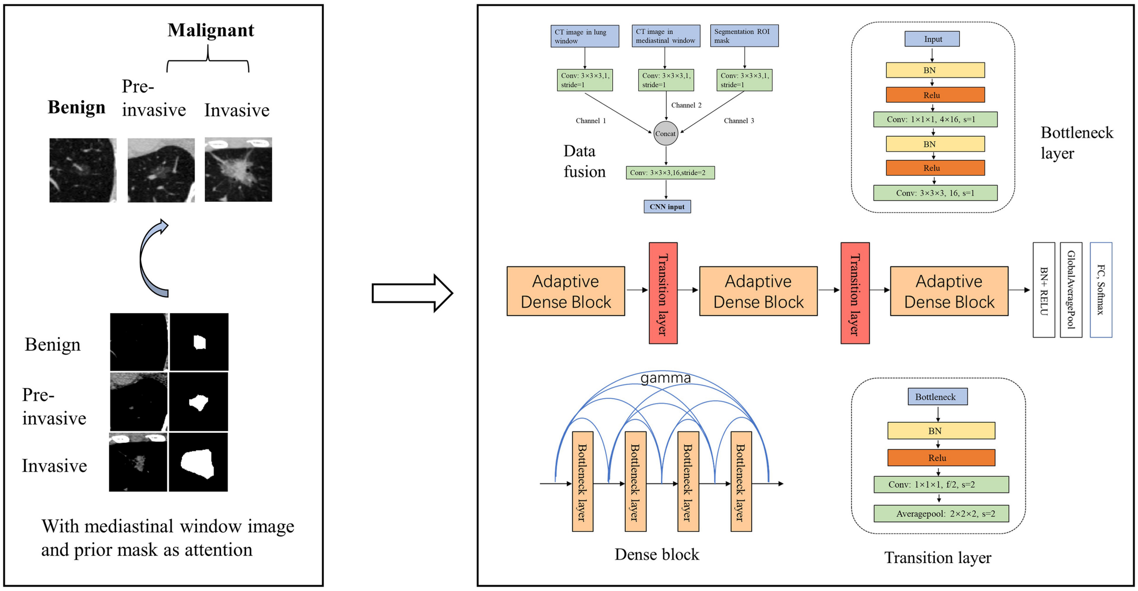


FIGURE 1 | Framework of our model. We developed a 3D CNN model for the malignancy and invasiveness recognition of subsolid pulmonary nodules. The 3D CNN model was based on modified 3D adaptive DenseNet and was improved by incorporating different window images and segmentation mask.

As the sample size was limited, we used data augmentation to avoid overfitting. We did online augmentation including rotations, reflection, and translation. For a given nodule patch and the corresponding mask, they were first translated by one to three voxels in three directions. Then the translated images were randomly rotated by 90, 180, 270, and 360° around the *x*-, *y*-, and *z*-axis. Finally, the rotated images were randomly flipped along the *x*-, *y*-, and *z*-axis.

For the network training, we used cross-entropy function as loss function and Adam optimizer to train the model. Xavier was used to initialize the network. The learning rate was set to 1e-4. Maximum iterative epoch was 1,000. We early stopped the training process when the validation dataset's performance had no improvement within five epochs. The batch size for each iteration was set to 24. The multiple test method was used to improve the stability of testing performance. Given a test example, the input image patch with different windows and the corresponding mask was randomly generated 10 times to obtain 10 different prediction probabilities, and the final prediction result was computed by averaging all prediction probabilities. The study was implemented with Tensorflow framework on a GeForce GTX 1080Ti GPU.

Observer Reader Study

To compare the performance of the CNN model with human experts for malignancy prediction, an observer reader study was conducted in the same testing dataset. Two radiologists (with over 10 years of clinical experience) were respectively asked to grade the SSNs based on preoperative CT images. The scores ranged from 0 to 10, and the higher the score was, the more likely they thought the SSN was malignant. The detailed scoring criteria can be found in **Supplementary Figure S2**. The radiologists made their own decisions independently. Also, the radiologists were given access to patients' demographics and clinical history as auxiliary information.

Model Evaluation and Statistical Analysis

To evaluate different models' performance, the receiver operating characteristic curve (ROC) was plotted, and the area under the ROC curve (AUC), sensitivity, and specificity were calculated to evaluate these models' discrimination ability. Delong test was used to pairwise compare different ROCs. Calibration curve was utilized to assess the calibration ability of the model. Brier score was calculated to quantify the calibration of those models, of which lower values (closer to 0) indicate better calibration. Decision curve analysis was used to determine the clinical usefulness of different models by calculating the net benefit of the constructed models at different threshold probabilities.

Mann-Whitney test was used to compare differences of the mean value of patient's age and max diameter in different groups. Pearson's $\chi 2$ test was used to compare differences of patients' gender and location proportion in different groups.

The statistical analysis was conducted with R software (Rproject.org) and python (version 3.7). P-value less than 0.05 was considered as statistically significant difference.

RESULTS

Patient Characteristics

From the total of 2,614 patients, 1,791 were malignant and 823 were benign nodules. The number of patients with 1 mm layer thickness was 1,735 (accounting for 66.4%), while the other 879 (33.6%) patients were with scans of 5 mm thickness. The median nodule diameter was 1 cm. All patients' characteristic statistical information are shown in **Table 1**. Detailed distribution of nodule sizes is shown in **Supplementary Figure S3**. Generally, female patients with larger diameter and location of right upper and left upper lobe were more likely to be malignant. The patients were randomly divided into training (60.9%), validation (19.1%), and testing datasets (20%) for the following analysis. The distribution of different subtypes of SSNs on each dataset is shown in **Table 2**. No significant difference was found among the datasets (**Supplementary Table S1**).

Performance of the Observer Reader Study

The observer readers' classification ROC, AUC, sensitivity, and specificity are shown in **Table 3** and **Supplementary Figure S4**. As we can see, one radiologist achieved the best performance with an AUC of 0.877 (95% CI: 0.843–0.911), sensitivity of 95.4%, and specificity of 66.7%, which was significantly better than another radiologist reader with an AUC of 0.815 (95% CI: 0.774–0.856). The difference also indicated the low inter-observer agreement of the malignancy recognition in clinical practice.

Performance of the 3D CNN Model for Malignancy Prediction

The ROC curves of the 3D CNN models for malignancy classification in the testing dataset are shown in **Figure 2**. As we can see, the best CNN model based on CT images was 3D CNN incorporated with different window images and the segmentation mask (AdaDense_M). The AUC of the best CNN model was 0.913 (95% CI: 0.885–0.940), which was significantly better than the 3D CNN only with the lung window image input (AdaDense) with an AUC of 0.848 (95% CI: 0.810–0.886). Also, the CNN model performed significantly better than clinical features-based model (AUC: 0.618), and adding clinical features to the CNN model yielded no significant improvement (AUC: 0.914, p = 0.489). The sensitivity and specificity of the AdaDense_M model at the optimal decision point were 86.1 and 83.8%. With a sensitivity of 100, 98, and 95%, the percentages of benign nodules that could be correctly identified was 32.5, 47.4, and 63.0%. Also, the Adadense_M model performed better than all the observer readers (AUC: 0.846±0.031).

The calibration curve and decision curve of the CNN model (AdaDense_M) were plotted in **Figure 3**. The Brier score was 0.101, showing satisfactory consistency between the predicted malignant probability and actual observation (**Figure 3A**). Also, the model can bring apparent benefits for the malignancy

TABLE 1 | Clinical characteristic of total patients.

Clinical Characteristics	Total Patients (n=2,614)	Malignant Nodules (n=1,791, 68.5%)	Benign Nodules (n=823, 31.5%)	Statistical Significance (Test Used)
Gender				
Male	924 (35.3%)	577 (32.2%)	347 (42.2%)	P<0.0001
Female	1,690 (64.7%)	1,214 (67.8%)	476 (57.8%)	(Pearson χ^2)
Age				
Median (Range)	57 (15–84)	58 (15–84)	57 (19–81)	P=0.055 (Mann-Whitney)
Max Diameter (cm)				
Median (Range)	1.0 (0.2–4.5)	1.1 (0.2–4.5)	0.9 (0.2–4.4)	p<0.0001 (Mann-Whitney)
Solid Ingredients				
PGGN[a]	1,768 (67.6%)	1,199 (66.9%)	569 (69.1%)	P=0.286 (Pearson χ^2)
PSN[b]	846 (32.4%)	592 (33.1%)	254 (30.9%)	
Location				
Right Upper Lobe	949 (36.3%)	671 (37.5%)	278 (33.8%)	p<0.0001 (Pearson χ^2)
Right Middle Lobe	198 (7.6%)	117 (6.5%)	81 (9.8%)	
Right Lower Lobe	469 (17.9%)	289 (16.1%)	180 (21.9%)	
Left Upper Lobe	670 (25.6%)	505 (28.2%)	165 (20.0%)	
Left Lower Lobe	328 (12.5%)	209 (11.7%)	119 (14.5%)	

[a] *PGGN, Pure ground-glass nodules.*
[b] *PSN, Part solid nodules.*

TABLE 2 | Distribution of SSN subtypes on each dataset.

	Training	Validation	Testing	Total
Benign	516	154	154	824
AAH/AIS	180	53	64	297
MIA	371	118	129	618
IA	525	175	175	875

TABLE 3 | Performance of the observer reader study.

	AUC	Sensitivity	Specificity
Radiologist1	0.815	80.8%	76.5%
Radiologist2	0.877	95.4%	66.7%

classification when the threshold was set to 0.01–0.99 compared with the treat-all strategies (perform surgeries in all patients) (**Figure 3B**).

Performance of the 3D CNN Model for Invasiveness Prediction

The ROC curves of the 3D CNN models for invasiveness classification in the testing dataset are shown in **Figure 4A**. The CNN model (AdaDense_M) achieved satisfactory AUC of 0.908 (95% CI: 0.877–0.939), sensitivity of 87.4%, and specificity of 80.8% at the optimal decision point. The confusion matrix is shown in **Table 4**. Calibration curve showed satisfactory consistency between the predicted invasiveness probability and the actual observation with a Brier score of 0.124 (**Figure 4B**).

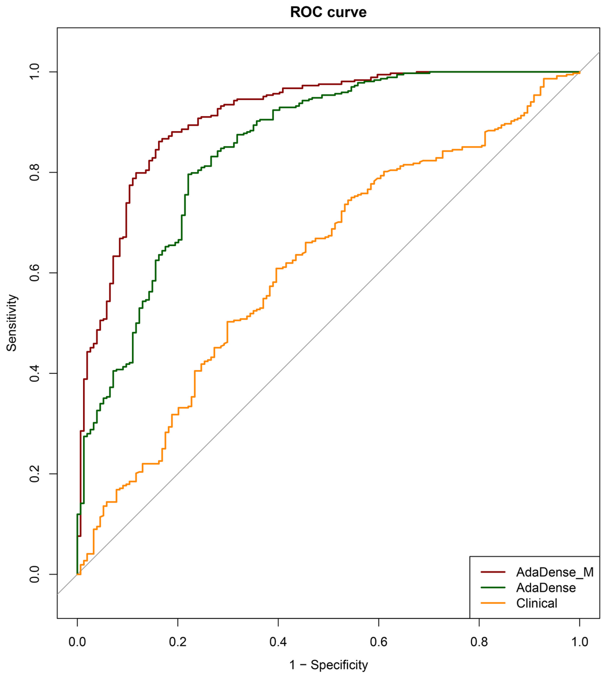


FIGURE 2 | The ROC curves of the CNN models for malignancy prediction. The ROC curves of the AdaDense_M (CNN incorporated with different window images and segmentation mask), AdaDense (CNN only with the lung window image as input), and baseline clinical model (diameter) for malignancy prediction in the testing dataset. The three models' corresponding AUCs were 0.913, 0.848, and 0.618, respectively. DeLong tests showed that the AdaDense_M performs significantly better than the AdaDense model and the clinical model (p<0.001).

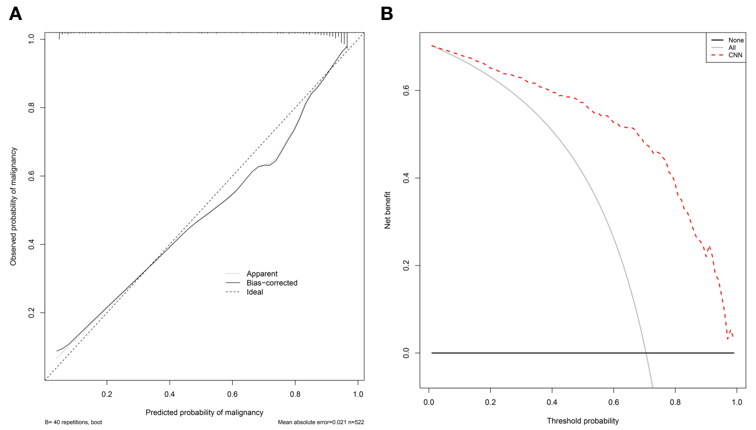


FIGURE 3 | The calibration curve and decision curve of the CNN model for malignancy prediction. **(A)** The calibration curve of the CNN model (AdaDense_M) for malignancy prediction. The diagonal dotted line represents a perfect prediction by an ideal model. **(B)** The decision curve of the CNN model (AdaDense_M) for malignancy prediction. The gray solid line represents the assumption that all patients had malignant nodules. The black solid line represents the assumption that no patients had malignant nodules. The net benefit was calculated by subtracting the proportion of all patients who are false positive from the proportion who are true positive, weighting by the relative harm of a false-positive and a false-negative result.

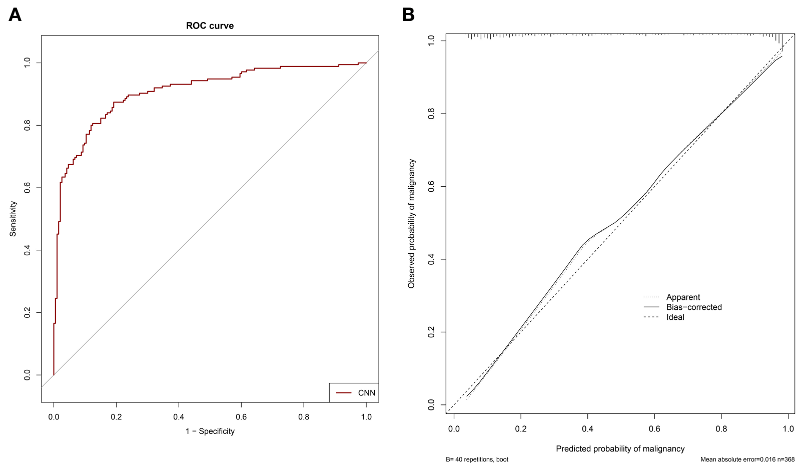


FIGURE 4 | The ROC curve and calibration curve of the CNN model for invasiveness prediction. **(A)** The ROC curve of the CNN model (AdaDense_M) for invasiveness prediction with an AUC of 0.908 in the testing dataset. **(B)** The calibration curve of the CNN model (AdaDense_M) for invasiveness prediction in the testing dataset. The diagonal dotted line represents a perfect prediction by an ideal model.

TABLE 4 | Confusion matrix of the CNN model for invasiveness prediction.

	CNN prediction		
Ground Truth	**Pre-invasive**	**Invasive**	**Total**
AAH/AIS	59	5	64
MIA	97	32	129
IA	22	153	175

DISCUSSION

Accurate diagnosis of malignancy and invasiveness of SSNs plays an important role in clinical decision-making, especially for thoracic surgeons. In this study, we developed and validated a novel deep-learning model based on preoperative CT images for accurate classification of SSNs. Moreover, the deep-learning model outperformed radiologists for malignancy prediction.

According to the Fleischner recommendations (3), follow-up CTs are recommended when subsolid nodules are initially detected to differentiate them between transient and persistent. Then, if the nodules are persistent, the management would be determined based on the patient's age, performance status, nodule size, and solid portion size. However, as there exist no national strategy for early-stage lung cancer screening in China, patients with pulmonary nodules may come to the hospital for a variety of reasons. Thus, for Chinese patients in clinical routine,

TABLE 5 | Other studies for the classification of pulmonary SSNs.

Author	Sample Size	Method	Task	AUC
Gon et al. (28)	123 malignant and 59 benign	Radiomics	Benign/Malignant	0.75
Digumarthy et al. (29)	77 malignant and 31 benign	Radiomics	Benign/Malignant	0.75–0.83
Yang et al. (30)	920 malignant and 94 benign	Qualitative feature synthesis	Benign/Malignant	0.89
Gong et al. (31)	828 malignant	CNN	AIS+MIA/IA	0.92
Zhao et al. (32)	651 malignant	CNN	Pre-invasive/Invasive	0.88

the lesions are usually larger at the first visit, resulting in the risk of diagnosis by dynamic follow-up. Therefore, it is necessary to diagnosis SSNs based on preoperative CT images at a single point. Furthermore, this diagnostic result greatly determines the subsequent treatment strategies in clinical practice. For SSNs that are basically diagnosed as benign, almost no intervention but only follow-up is needed. While for SSNs highly suspicious of malignancy, surgery or SBRT is usually adopted according to the individual condition of patients. More specifically, sublobectomy is more appropriate for pre-invasive SSNs, while lobectomy with extended lymph node dissection is more suitable for invasive SSNs. Currently, the inaccurate diagnosis based on radiologists' subjective judgment may cause overtreatment or undertreatment, which is harmful for the long-term survival of patients. Here, we established a quantitative deep-learning model that can accurately identify the malignancy and invasiveness of SSNs before the operation. This will play an important guiding role in the decision-making of the final surgical resection range, which can avoid unnecessary surgical trauma, reduce the complications of patients, and preserve the lung function to the greatest extent, and at the same time, patients can get radical treatment opportunities.

Considering that CNN has great advantage in automatically extracting deep representative image features, we decided to establish a CNN model for malignancy and invasiveness recognition of SSNs. Our established CNN model incorporated with different window images and segmentation mask (Adadense_M) finally achieved satisfying classification performance. Besides that, we tried to developed a fusion model by combining the CNN model's prediction result and the best radiologist's score with logistic regression. The fusion model finally achieved an AUC of 0.956 (95% CI: 0.938–0.975) for malignancy prediction, which was significantly better than the CNN model or radiologist alone. This result means that the CNN model has great potential to help the radiologist make better diagnosis of malignancy of SSNs.

Small sample size was the bottleneck to develop a high-efficacy prediction model for previous studies to distinguish pulmonary SSNs (28–32) (**Table 5**). Our study utilized the largest sample size to date with detailed CT images and pathologic information of SSNs. Compared with models built with qualitative features and radiomics (28–30), our CNN model can automatically learn deep representative features, which have stronger predictive ability than the hand-crafted features. Thus, our CNN model performs significantly better than other radiomics models for malignancy prediction of SSNs. Furthermore, in comparison with models developed with CNN (31, 32), our AdaDense_M model creatively uses the prior segmentation mask and tumor cube in mediastinal window as attention map, which can make the network focus on information within the tumor and its solid components. Results show that the CNN model we built achieved a high AUC value for invasiveness prediction of SSNs among the existing studies.

This study also has some limitations. First, we only included patients with pathologically confirmed SSNs who had undergone surgical resection, which results in a selection bias of more malignant patients. If more benign samples can be included, our model would be further improved. Second, there are 33.5% patients who only conducted regular CT scans with the layer thickness of 5 mm. Due to the small size and unique morphology of SSNs, the regular CT scans of SSNs are too blurred to excavate deep features for CNN. More thin-section CT scan data will be collected in the future, and the model performance may be further improved. Moreover, external dataset and prospective cohort are also required to validate the generalization ability of our model.

CONCLUSION

We constructed a deep learning–based model to identify the malignancy and invasiveness of pulmonary SSNs based on CT images. The model achieved a satisfactory performance and was proven with potential to guide the selection of surgery candidates and type of lung resection methods.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the datasets are privately owned by Shanghai Chest Hospital and are not made public. Requests to access the datasets should be directed to XF, xlfu1964@hotmail.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Shanghai Chest Hospital, Shanghai Jiaotong University. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version. XF, JZ, TS, and RH contributed to the study concept and design. TS, RH and XL contributed to acquisition of data. RH, TS, XY, XL, JX, QZ, CZ, XC, and WY contributed to analysis and interpretation of data. TS and RH contributed to drafting of the manuscript.

FUNDING

This work was supported in part by the Major Research Plan of the National Natural Science Foundation of China (Grant No. 92059206), Shanghai Jiao Tong University Medical Engineering Cross Research Funds (No. YG2017ZD10 and YG2014ZD05), National Key Research and Development Program (No. 2016YFC0905502 and 2016YFC0104608), and National Natural Science Foundation of China (No. 81371634).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.700158/full#supplementary-material

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2019. *CA Cancer J Clin* (2019) 69(1):7–34. doi: 10.3322/caac.21551
2. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced Lung-Cancer Mortality With Low-Dose Computed Tomographic Screening. *N Engl J Med* (2011) 365(5):395–409. doi: 10.1056/NEJMoa1102873
3. Macmahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society. *Radiology* (2017) 284(1):228–43. doi: 10.1148/radiol.2017161659
4. Henschke C. CT Screening for Lung Cancer : Frequency and Significance of Part-Solid and Nonsolid Nodules. *Ajr Am J Roentgenol* (2002) 178(5):1053–7. doi: 10.2214/ajr.178.5.1781053
5. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol* (2015) 10(9):1243–60. doi: 10.1097/jto.0000000000000630
6. Godoy MCB, Naidich DP. Subsolid Pulmonary Nodules and the Spectrum of Peripheral Adenocarcinomas of the Lung: Recommended Interim Guidelines for Assessment and Management. *Article Radiol* (2009) 253(3):606–22. doi: 10.1148/radiol.2533090179
7. Hammer MM, Hatabu H. Subsolid Pulmonary Nodules: Controversy and Perspective. *Eur J Radiol Open* (2020) 7:100267. doi: 10.1016/j.ejro.2020.100267
8. Ye T, Deng L, Wang S, Xiang J, Zhang Y, Hu H, et al. Lung Adenocarcinomas Manifesting as Radiological Part-Solid Nodules Define a Special Clinical Subtype. *J Thorac Oncol* (2019) 14(4):617–27. doi: 10.1016/j.jtho.2018.12.030
9. Ye T, Deng L, Xiang J, Zhang Y, Hu H, Sun Y, et al. Predictors of Pathologic Tumor Invasion and Prognosis for Ground Glass Opacity Featured Lung Adenocarcinoma. *Ann Thorac Surg* (2018) 106(6):1682–90. doi: 10.1016/j.athoracsur.2018.06.058
10. Tsutani Y, Miyata Y, Nakayama H, Okumura S, Adachi S, Yoshimura M, et al. Appropriate Sublobar Resection Choice for Ground Glass Opacity-Dominant Clinical Stage IA Lung Adenocarcinoma: Wedge Resection or Segmentectomy. *Chest* (2014) 145(1):66–71. doi: 10.1378/chest.13-1094
11. Zhang J, Wu J, Tan Q, Zhu L, Gao W. Why do Pathological Stage IA Lung Adenocarcinomas Vary From Prognosis?: A Clinicopathologic Study of 176 Patients With Pathological Stage IA Lung Adenocarcinoma Based on the IASLC/ATS/ERS Classification. *J Thorac Oncol* (2013) 8(9):1196–202. doi: 10.1097/JTO.0b013e31829f09a7
12. Yang J, Wang H, Geng C, Dai Y, Ji J. Advances in Intelligent Diagnosis Methods for Pulmonary Ground-Glass Opacity Nodules. *Biomed Eng Online* (2018) 17(1):1–18. doi: 10.1186/s12938-018-0435-2
13. Kim H, Park CM, Koh JM, Lee SM, Goo JM. Pulmonary Subsolid Nodules: What Radiologists Need to Know About the Imaging Features and Management Strategy. *Diagn Interv Radiol* (2014) 20(1):47–57. doi: 10.5152/dir.2013.13223
14. Jin X, Zhao S-H, Gao J, Wang D-J, Wu J, Wu C-C. CT Characteristics and Pathological Implications of Early Stage (T1N0M0) Lung Adenocarcinoma With Pure Ground-Glass Opacity. *Eur Radiol* (2015) 25(9):2532–40. doi: 10.1007/s00330-015-3637-z
15. Dai C, Ren Y, Xie H, Jiang S, Fei K, Jiang G, et al. Clinical and Radiological Features of Synchronous Pure Ground-Glass Nodules Observed Along With Operable Non-Small Cell Lung Cancer. *J Surg Oncol* (2016) 113(7):738–44. doi: 10.1002/jso.24235
16. Gierada DS, Pilgram TK, Ford M, Fagerstrom RM, Church TR, Nath H, et al. Lung Cancer: Interobserver Agreement on Interpretation of Pulmonary Findings at Low-Dose CT Screening. *Radiology* (2008) 246(1):265–72. doi: 10.1148/radiol.2461062097
17. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-To-End Lung Cancer Screening With Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nat Med* (2019) 25(6):954–61. doi: 10.1038/s41591-019-0447-x
18. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks. *Nature* (2017) 542(7639):115–8. doi: 10.1038/nature21056
19. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep Learning for Automated Skeletal Bone Age Assessment in X-Ray Images. *Med Image Anal* (2017) 36:41–51. doi: 10.1016/j.media.2016.10.010
20. van der Burgh HK, Schmidt R, Westeneng HJ, de Reus MA, van den Berg LH, van den Heuvel MP. Deep Learning Predictions of Survival Based on MRI in Amyotrophic Lateral Sclerosis. *NeuroImage Clin* (2017) 13:361–9. doi: 10.1016/j.nicl.2016.10.008
21. Jiang H, Ma H, Qian W, Gao M, Li Y. An Automatic Detection System of Lung Nodule Based on Multi-Group Patch-Based Deep Learning Network. *IEEE J Biomed Health Informatics* (2017) PP(99):1. doi: 10.1109/JBHI.2017.2725903
22. Xie Y, Xia Y, Zhang J, Song Y, Feng D, Fulham M, et al. Knowledge-Based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT. *IEEE Trans Med Imaging* (2019) 38(4):991–1004. doi: 10.1109/tmi.2018.2876510
23. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-Aided Classification of Lung Nodules on Computed Tomography Images via Deep Learning Technique. *Oncotargets Ther* (2015) 8:2015–22. doi: 10.2147/OTT.S80733
24. Huang G, Liu Z, Weinberger KQ. Densely Connected Convolutional Networks. *2017 IEEE Conf Computer Vision Pattern Recog (CVPR)* (2017) 2261–9. doi: 10.1109/CVPR.2017.243
25. Kamiya S, Iwano S, Umakoshi H, Ito R, Shimamoto H, Nakamura S, et al. Computer-Aided Volumetry of Part-Solid Lung Cancers by Using CT: Solid Component Size Predicts Prognosis. *Radiology* (2018) 287(3):1030–40. doi: 10.1148/radiol.2018172319
26. Revel MP, Mannes I, Benzakoun J, Guinet C, Leger T, Grenier P, et al. Subsolid Lung Nodule Classification: A CT Criterion for Improving Interobserver Agreement. *Radiology* (2018) 286(1):316–25. doi: 10.1148/radiol.2017170044
27. Ge X, Gao F, Li M, Chen Y, Hua Y. Diagnostic Value of Solid Component for Lung Adenocarcinoma Shown as Ground-Glass Nodule on Computed Tomography. *Zhonghua Yi Xue Za Zhi* (2014) 94(13):1010–3. doi: 10.3760/cma.j.issn.0376-2491.2014.13.014
28. Gong J, Liu J, Hao W, Nie S, Wang S, Peng W. Computer-Aided Diagnosis of Ground-Glass Opacity Pulmonary Nodules Using Radiomic Features

Analysis. *Phys Med Biol* (2019) 64(13):135015. doi: 10.1088/1361-6560/ab2757

29. Digumarthy SR, Padole AM, Rastogi S, Price M, Mooradian MJ, Sequist LV, et al. Predicting Malignant Potential of Subsolid Nodules: Can Radiomics Preempt Longitudinal Follow Up CT? *Cancer Imaging* (2019) 19(1):36. doi: 10.1186/s40644-019-0223-7
30. Yang W, Sun Y, Fang W, Qian F, Ye J, Chen Q, et al. High-Resolution Computed Tomography Features Distinguishing Benign and Malignant Lesions Manifesting as Persistent Solitary Subsolid Nodules. *Clin Lung Cancer* (2018) 19(1):e75–83. doi: 10.1016/j.cllc.2017.05.023
31. Gong J, Liu J, Hao W, Nie S, Zheng B, Wang S, et al. A Deep Residual Learning Network for Predicting Lung Adenocarcinoma Manifesting as Ground-Glass Nodule on CT Images. *Eur Radiol* (2019) 30(4):1847–55. doi: 10.1007/s00330-019-06533-w
32. Zhao W, Yang J, Sun Y, Li C, Wu W, Jin L, et al. 3d Deep Learning From CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas. *Cancer Res* (2018) 78(24):6881–9. doi: 10.1158/0008-5472.CAN-18-0696

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Oncology

ORIGINAL RESEARCH
published: 29 July 2021
doi: 10.3389/fonc.2021.696706

Comparable Performance of Deep Learning–Based to Manual-Based Tumor Segmentation in KRAS/NRAS/BRAF Mutation Prediction With MR-Based Radiomics in Rectal Cancer

Guangwen Zhang[1†], Lei Chen[2†], Aie Liu[2], Xianpan Pan[2], Jun Shu[1], Ye Han[1], Yi Huan[1*] and Jinsong Zhang[1*]

[1] Department of Radiology, Xijing Hospital, Fourth Military Medical University, Xi'an, China, [2] Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

OPEN ACCESS

Edited by:
Bo Gao,
Affiliated Hospital of Guizhou Medical University, China

Reviewed by:
Guolin Ma,
China-Japan Friendship Hospital, China
Sarthak Pati,
University of Pennsylvania, United States

***Correspondence:**
Yi Huan
huanyi3000@163.com
Jinsong Zhang
stspine@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: 20 April 2021
Accepted: 15 July 2021
Published: 29 July 2021

Citation:
Zhang G, Chen L, Liu A, Pan X, Shu J, Han Y, Huan Y and Zhang J (2021) Comparable Performance of Deep Learning–Based to Manual-Based Tumor Segmentation in KRAS/NRAS/BRAF Mutation Prediction With MR-Based Radiomics in Rectal Cancer. Front. Oncol. 11:696706. doi: 10.3389/fonc.2021.696706

Radiomic features extracted from segmented tumor regions have shown great power in gene mutation prediction, while deep learning–based (DL-based) segmentation helps to address the inherent limitations of manual segmentation. We therefore investigated whether deep learning–based segmentation is feasible in predicting KRAS/NRAS/BRAF mutations of rectal cancer using MR-based radiomics. In this study, we proposed DL-based segmentation models with 3D V-net architecture. One hundred and eight patients' images (T2WI and DWI) were collected for training, and another 94 patients' images were collected for validation. We evaluated the DL-based segmentation manner and compared it with the manual-based segmentation manner through comparing the gene prediction performance of six radiomics-based models on the test set. The performance of the DL-based segmentation was evaluated by Dice coefficients, which are 0.878 ± 0.214 and 0.955 ± 0.055 for T2WI and DWI, respectively. The performance of the radiomics-based model in gene prediction based on DL-segmented VOI was evaluated by AUCs (0.714 for T2WI, 0.816 for DWI, and 0.887 for T2WI+DWI), which were comparable to that of corresponding manual-based VOI (0.637 for T2WI, P=0.188; 0.872 for DWI, P=0.181; and 0.906 for T2WI+DWI, P=0.676). The results showed that 3D V-Net architecture could conduct reliable rectal cancer segmentation on T2WI and DWI images. All-relevant radiomics-based models presented similar performances in KRAS/NRAS/BRAF prediction between the two segmentation manners.

Keywords: rectal cancer, deep learning, radiomics, magnetic resonance imaging, gene mutation

INTRODUCTION

It is clear that (1) Epidermal Growth Factor Receptor (EGFR) inhibitors could provide a beneficial clinical outcome for metastatic Colorectal Cancer (mCRC) patients with wild-type rat sarcoma viral oncogene homolog (RAS) genes rather than mutant types. However, some patients with wild-type RAS still exhibit no response to anti-EGFR therapies. To address this confusion, the downstream factors of the RAS pathway was explored, and a specific mutation in the BRAF gene (V600E) (2) was confirmed to be responsible for less response from EGFR inhibitors and a worse prognosis. Therefore, the National Comprehensive Cancer Network (NCCN) guideline (3) recommends that the genotype of KRAS/NRAS/BRAF should be determined in patients with mCRC and further claims that patients with these mutations should not be provided with medication such as cetuximab or panitumumab, either alone or in combination with other anticancer drugs, since there is little chance of them having any benefit and the toxicity and expense suffered will not be reasonable.

Up to now, it is still a state-of-the-art routine practice to detect gene mutation status by pathologically analyzing biopsy samples or resected tissues. However, there is a growing recognition (4) that tissue-based genetic tests have some limitations such as intratumoral heterogeneity, clonal evolution, and poor DNA quality, especially in biopsy samples, which can lead to a suboptimal profile of tumor genetic characteristics and be of limited value in routine practice. In recent years, liquid biopsy has emerged to be an alternative method to determine gene status. However, this newly raised technology is still limited for clinical practice due to the availability of samples for testing, the non-standardized method, and the low sensitivity in low-stage tumor (5). Therefore, efficient identification of RAS and BRAF status in rectal cancers using a non-invasive method, which could feasibly reveal the whole tumor gene features in real-time, would be of meaningful assistance in providing individual tailored therapy.

There have been a certain number of researches based on PET/CT (6), CT (7), or MRI (8) focusing on detecting RAS gene mutations in rectal cancer, while these studies all delineated tumors manually. It is worth noting that the inherent limitations of manual segmentation, such as long time-consumption and inter- and intra-observer variability, have significant impact on medical image quantitative analysis (9) and the efficacy and safety of the radiotherapy plan (10). Fortunately, state-of-the-art auto segmentation based on deep-learning architecture has been developed and shown to be able to address these problems. Successful application included making differential diagnosis in brain (11) and contouring gross tumor volumes in rectal cancer radiotherapy (12). For 3D medical image segmentation, 3D V-Net, a special fully convolutional neural network (CNN), has been shown to be able to produce satisfactory segmentation results (13). The network first detects the boundary from a “coarse” resolution, then provides accurate spatial localization through a “fine” resolution.

Radiomics, with its high-throughput quantitative image features, has shown exciting power in assessing treatment response (14), genetic profile (8), predicting lymph node (15), and distant metastasis (16) in respect to rectal cancers. Furthermore, combinations of DL-based automatic segmentation and radiomics have been demonstrated with great potential in glioma grading (17), treatment response assessment (18), and the isocitrate dehydrogenase-1 (IDH1) mutation prediction (19) of glioblastoma. However, the combination of DL-based auto segmentation with MR-based radiomics in predicting gene mutation for rectal cancer has not been investigated. Thus, we attempt to segment rectal cancer *via* 3D V-Net on T2WI and DWI and then compare the performance of radiomics in predicting the KRAS/NRAS/BRAF status between DL-based auto segmentation and manual-based segmentation.

MATERIALS AND METHODS

Dataset

This retrospective study was approved by the institutional review board in our hospital, and informed patient consent was waived. A total of 202 participants (mean age 59.88 ± 11.82 years, 139 males and 63 females) with rectal adenocarcinoma confirmed by colonoscopy biopsy were recruited from 333 patients who had underwent pelvic MR imaging on a 3.0T scanner (November 2016 to May 2019) after screening according to the following exclusive criteria: (a) treated with any strategy before MR imaging or surgery (n=75); (b) the interval between MR imaging and postoperative pathology was more than 4 weeks (n=8); (c) gross artifacts or severe distortion of MR images (n=18); (d) absence of visible lesion or the volume of lesion was less than 1 cm^3 on MR image (n=7); (e) other pathological types of tumor (mucinous adenocarcinoma, neuroendocrine carcinoma, and malignant melanoma) (n=23). Among the 202 participants, 94 patients were subject to a KRAS/NRAS/BRAF mutation test, and the interval was less than 4 weeks between MR imaging and the gene test. Among the 94 patients who underwent the gene test, 53 patients harbored mutant KRAS/NRAS/BRAF, and 41 patients were wild type. The remaining 108 patients were not tested for mutations and could not be used to assess mutation prediction, but they were suitable for modeling segmentation. Therefore, we used the 108 patients without the gene test as the training set for the auto segmentation model and the 94 patients with the gene test as the test dataset, each including both the T2WI and DWI images. The radiomics-based model for gene mutation prediction was constructed based on 94 patients’ MR images *via* 5-fold cross validation. Considering the different imaging modalities and tumor segmentation manners, we constructed six radiomics-based models, which were T2WI+manual-based VOI, T2WI+DL-based VOI, DWI+manual-based VOI, DWI+DL-based VOI, T2WI+DWI+manual-based VOI, and T2WI+DWI+DL-based VOI. The detailed experiment flow chart is shown in **Figure 1**, and patients’ baseline clinical characteristics for genotype prediction is summarized in **Table 1**.

MR Image Acquisition

All MR scanning was performed on a 3.0T MR scanner (Discovery MR750, GE Medical Systems) with an eight-channel phased-array coil. Bowel preparation was implemented by drinking folium sennae soup (a kind of laxative) after dinner the night before the examination. Antispasmodic and other intestinal contrast agents were not used. Rectal MRI protocols included axial T1WI (TR/TE = 487/8 ms), coronal and sagittal T2WI (TR/TE = 7,355/136 ms), oblique axial small FOV FRFSE

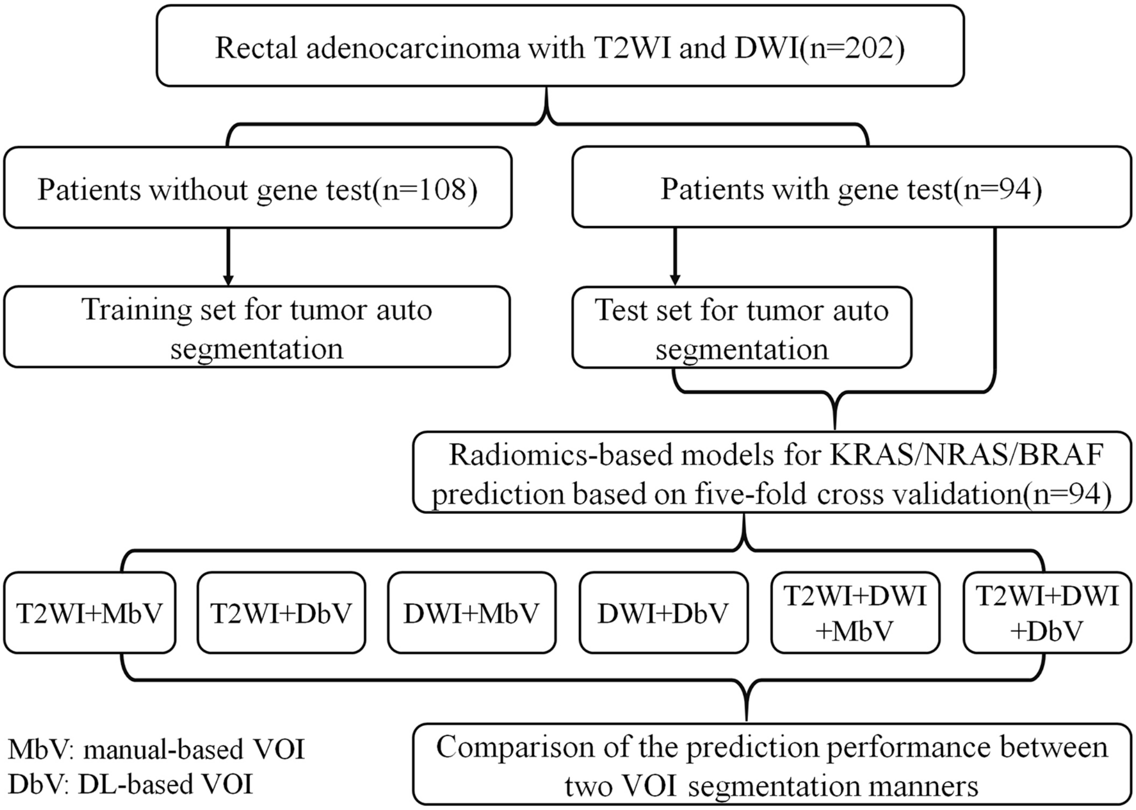

Rectal adenocarcinoma with T2WI and DWI(n=202)
Patients without gene test(n=108)
Patients with gene test(n=94)
Training set for tumor auto segmentation
Test set for tumor auto segmentation
Radiomics-based models for KRAS/NRAS/BRAF prediction based on five-fold cross validation(n=94)
T2WI+MbV
T2WI+DbV
DWI+MbV
DWI+DbV
T2WI+DWI +MbV
T2WI+DWI +DbV
MbV: manual-based VOI
DbV: DL-based VOI
Comparison of the prediction performance between two VOI segmentation manners

FIGURE 1 | Experiment flow chart. VOI, volumes of interest.

TABLE 1 | Patient baseline characteristics for genotype (KRAS/NRAS/BRAF) prediction.

Characteristics	Wild type (n = 41)	Mutant type (n = 53)	*P*
Age, years (Mean ± SD)	60.44 ± 12.93	61.57 ± 10.30	0.639
Gender, n (%)			0.297
Male	29 (70.7%)	32 (60.4%)	
Female	12 (29.3%)	21 (39.6%)	
Histologic grade, n (%)			0.206
Well	5 (12.2%)	9 (17.0%)	
Moderate	35 (85.4%)	38 (71.7%)	
Poor	1 (2.4%)	6 (11.3%)	
pT stage, n (%)			**0.021**
T1/2	22 (53.7%)	16 (30.2%)	
T3/4	19 (46.3%)	37 (69.8%)	
pN stage, n (%)			0.183
N0	25 (61.0%)	25 (47.2%)	
N1	16 (39.0%)	28 (52.8%)	
CEA, n (%)			0.543
≤5 ng/ml (normal)	27 (65.9%)	38 (71.7%)	
>5 ng/ml (abnormal)	14 (34.1%)	15 (28.3%)	
CA-199, n (%)			0.588
≤27 u/ml (normal)	35 (85.4%)	43 (81.1%)	
>27 u/ml (abnormal)	6 (14.6%)	10 (18.9%)	

Chi-squared or Fisher's exact tests, as appropriate, were used to compare the differences in categorical variables, while independent samples t test was used to compare the differences in age. Bold value: Rectal cancer with more advanced T stage is prone to evolve mutant KRAS/NRAS/BRAF (P=0.021). p, pathological.

T2WI (TR/TE = 6,055/130 ms, Slice Thickness = 3 mm, Gap = 0.3 mm, FOV=200 × 200 mm, Matrix = 352×256), and axial single-shot EPI DWI (TR/TE = 4,734/80 ms, Slice Thickness = 4 mm, Spacing = 0.5 mm, FOV=340 × 340 mm, Matrix = 128×140, NEX = 8, b = 0, 1,000 s/mm^2). An oblique axial T2WI high-resolution sequence was planned perpendicularly to the bowel with the tumor, while the axial DWI sequence was performed parallelly to the horizontal line.

Imaging Pre-Processing

As the reliability of manual VOI delineation had been reported in our previous study (20), the whole-tumor volume was manually delineated as the ground truth annotation on T2WI and DWI (b=1,000 s/mm^2) images by one radiologist with 8 years of experience in abdominal MRI and scrutinized by another senior abdominal MRI radiologist with 20 years of experience. The regions of contiguous normal rectal wall and lumen against tumor were manually labeled on T2WI images, and the magnetic susceptibility artifacts were labeled on DWI images, which were used for the training and validation of the automated tumor segmentation algorithm. All manual delineations were performed using ITK-SNAP (version 3.8) (21). Because of the peristalsis of rectum and different imaging parameters such as matrix, FOV (Field of View), slice thickness, and scan position line, the processing of the registration and image fusion between T2 and DWI images was not performed.

All MR images were normalized to accelerate the convergence of neural network training. First, the MR images were resampled to the same spatial resolution: 0.4×0.4×3.3 (mm), and then the gray values were linearly normalized into the range [0, 1]. Considering the GPU memory, the input 3D patch size was set to 96×96×32. Due to the limited amount of training images, image augmentation was performed, which included shift, rotation, scale, and flip slightly.

Network Architecture of 3D V-Net

We applied cascade learning in this work based on 3D V-Net for the tumoral tissue segmentation of the rectum on T2WI and DWI sequences. The code of 3D V-Net was improved from the V-Net (13). The architecture of the conventional V-Net has two pathways: the left part of the network consists of a compression path, while the right part decompresses the signal until its original size is reached. The detailed network architecture is shown in **Figure 2**. The proposed cascade neural network includes one coarse model and one fine model. The coarse-to-fine segmentation method detects the boundary from coarse resolution to the highest fine resolution to provide accurate spatial localization. The input of the 3D V-Net is a single

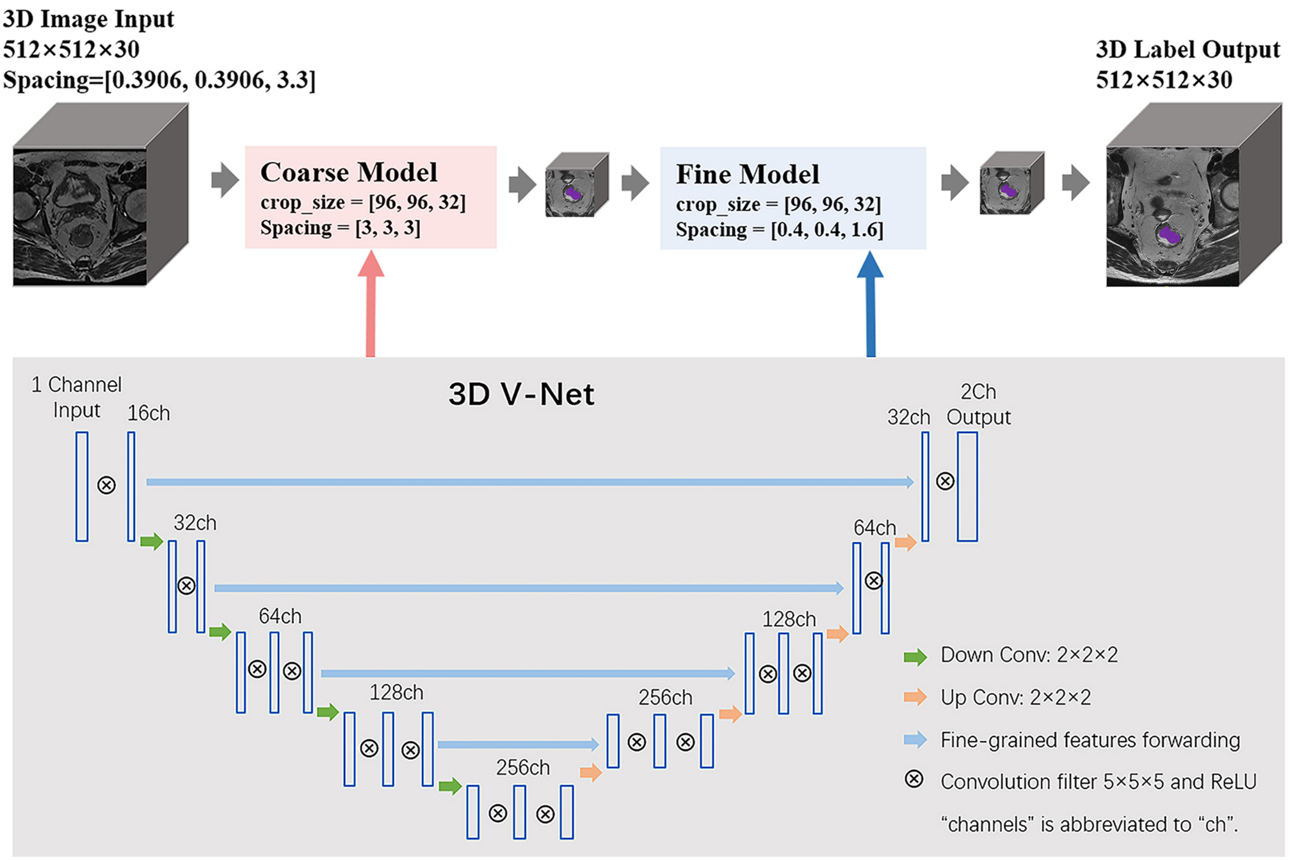


FIGURE 2 | The schematic network architecture of cascade V-Net.

sequence of a patient such as T2WI, while the output is a map of classification probability, which determines whether voxels of image belong to tumor or background. The loss function based on the Dice coefficient (range [0, 1]), which we sought to maximize, was performed in the training process. It is defined as

$$D = \frac{2\sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2}$$

Where N is the number of voxels of the image, p_i is the prediction probability of the i-th voxel which belongs to the target region, and g_i denotes whether the i-th voxel belongs to ground truth annotation or not (1 means yes, 0 means no). The volume size of the input and output image is 512×512×30, and the parameters of spacing for the coarse model and fine model are [3,3,3] and [0.4,0.4,1.6], respectively. Similar to other CNN, the training process was iterated with min-batch and stochastic gradient descent to ensure quick convergence. Tumor volume was segmented using forward propagation in the test process.

Genes (KRAS/NRAS/BRAF) Mutational Status Analysis

The tissue blocks were acquired from resected tumors, and pathologists selected the samples for gene mutational analysis. Genomic DNA was extracted from 5 mm formalin-fixed, paraffin-embedded (FFPE) tumor tissue sections, using a DNA FFPE Tissue Kit (AmoyDx, China). KRAS (exons 2, 3, and 4), NRAS (exons 2, 3), and BRAF (exons 15, V600E) mutations were detected by using polymerase chain reaction (PCR) and amplification-refractory mutation system (ARMS). Among the 53 patients with mutant genes, 48 patients were KRAS mutation, four patients were NRAS mutation, and one was BRAF mutation.

Radiomics Features Extraction, Selection, and Classifier Modeling for Gene Mutation Prediction

Radiomics analysis was performed by a clinical research platform (uAI Research Portal, United Imaging Intelligence Co., Ltd, China). The code for radiomics analysis was developed based on pyradiomics (https://pyradiomics.readthedocs.io/en). First, a total of 2,600 features were extracted from the labeled tumor volume of each MR sequence. These features were computed by the combination of 104 original image features with 25 image filters. The original image features include First-order, Shape, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighborhood Gray-Tone Difference Matrix (NGTDM). The image filters consist of Gaussian noise, curvature flow, Laplacian of Gaussian, Discrete Gaussian, Speckle noise, Recursive Gaussian, shot noise, and Wavelets. Second, feature selection was performed on the extracted features (2,600 dimensions) by least absolute shrinkage and selection operator (Lasso) method to work out an optimal feature subset (around 10 dimensions, for example). We set two parameters for LASSO, the feature scaler and shrinkage penalty, as min-max scaler and 0.02, respectively. The selected features for each radiomics-based model are presented in the supplementary material. Then, a radiomics-based model was built by support vector machine (SVM) classifier with the selected features. The parameters of SVM consist of penalty factor C (3.0), Gamma (0.03), and kernel (radial basis function). The predict models were verified by five-fold cross-validation and thus derived an average performance.

Statistics

Differences of patient baseline characteristics between the wild-type and mutant groups were tested using independent samples t test and chi-squared or Fisher's exact tests, as appropriate. Performance of the V-Net with respect to tumor segmentation was evaluated in the test dataset using the Dice coefficient. The AUC (area under the curve), accuracy, sensitivity, and specificity were calculated to evaluate the performance of the radiomics-based model in differentiating gene status. DeLong's test was used to compare two AUCs of the manual based model and deep learning–based model of identical imaging modality. The statistical analyses were conducted with SPSS (version 26.0), Medcalc (version 20.0), and PyCharm (version 2018, Python version 3.0). A two-sided p value < 0.05 was statistically considered significant difference.

RESULTS

Performance of 3D V-Net Segmentation Algorithm

The ground truth annotation includes 202 rectal cancers on T2WI and DWI sequences. To evaluate the performance of the 3D V-Net, the Dice Similarity Coefficient (DSC) was used to compare segmentations between AI and a radiologist. The volumetric segmentations generated from the deep learning model are probability maps. The mean and standard deviation of the Dice is 0.878 ± 0.214 and 0.955 ± 0.055 for T2WI and DWI separately in the test dataset. A paradigm of tumor segmentation results are shown in **Figure 3**.

Clinical and Pathological Characteristics

Among the 202 participants, 94 patients underwent a KRAS/NRAS/BRAF mutation test. There were 53 patients who harbored mutant genes, and 41patients were wild type. A statistical difference in terms of age, gender, histologic grade, pN stage, CEA, and CA-199 levels was not found between wild-type and mutant groups, except at the pT stage ($p = 0.021$). It seems that a tumor with more advanced T stage is prone to evolve mutant gene (**Table 1**).

Testing of Gene Mutation Prediction With Radiomics Signature

We built radiomics-based models with extracted features from two MR sequences of T2WI and DWI. Each sequence was processed separately to compute features from DL-based and manual-based VOI, respectively. Furthermore, we combined all features computed from T2WI and DWI sequences, and then applied the feature selection method LASSO to obtain an optimal feature subset. Thus, in total we collected six feature subsets and built six radiomics-based models for gene prediction. The mean performance of each model based on five-fold cross-validation is

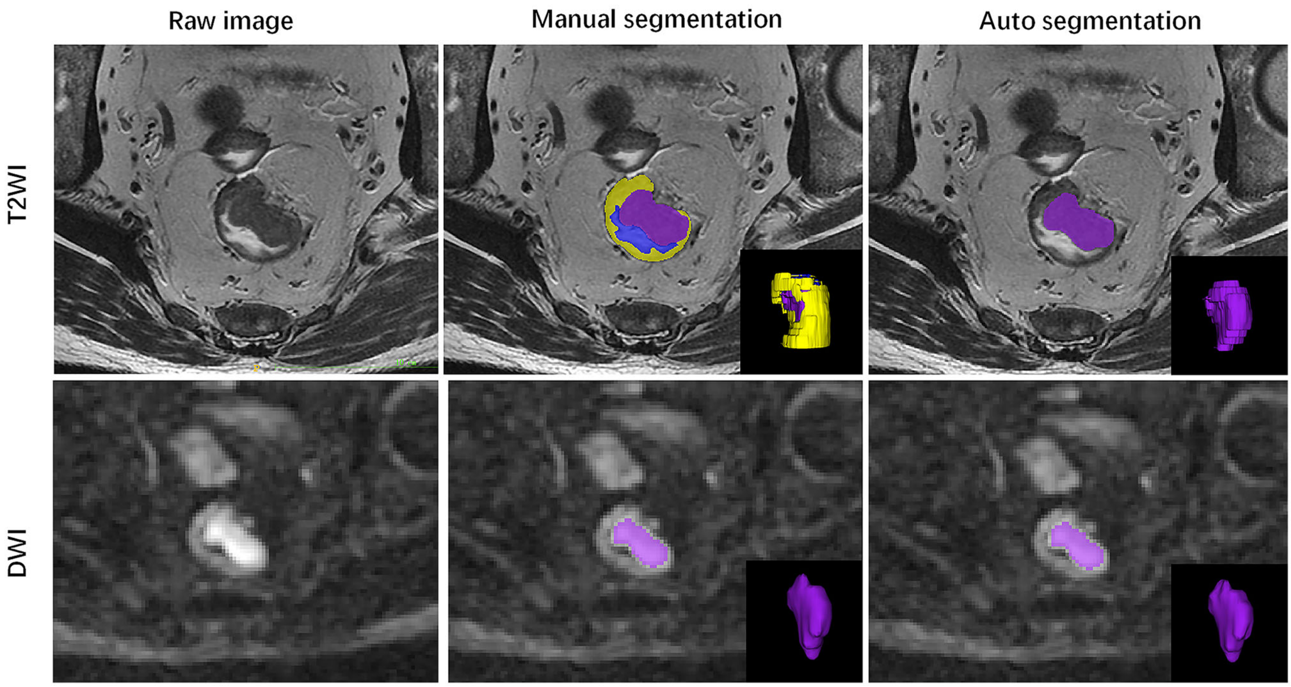


FIGURE 3 | Illustration of automated segmentation using 3D V-Net *versus* ground truth on rectal MR images of a 51-year-old male. Purple indicates tumor, yellow indicates normal rectal wall, and blue indicates lumen. The Dice was 0.980 on T2WI and 0.981 on DWI for this patient.

listed in **Table 2** and includes accuracy, specificity, sensitivity, and AUC. For each imaging modality, the prediction performance of gene mutation did not show any statistical difference between DL-based segmentation and manual-based segmentation (**Table 2** and **Figure 4**).

DISCUSSION

In this study, we segmented rectal cancer *via* 3D V-Net on T2WI and DWI and then compared the radiomics performance in predicting KRAS/NRAS/BRAF status between DL-based auto segmentation and manual-based segmentation. By virtue of volumetric convolution and coarse-to-fine segmentation models, higher tumor segmentation performance (Disc=0.878 and 0.955 for T2WI and DWI) was achieved by V-Net in our study compared with Trebeschi's (22) (Dice=0.70 for confusion image of T2WI and DWI) and Wang's (Dice=0.74 for T2WI) (12) work. This could be explained with low signal noisy ratio caused by 1.5T MR scanner in Trebeschi's work, volumetric information loss with 2D U-net architecture in Wang's work, and their relatively small sample size (n=140 and 93, respectively). It has been widely recognized that qualified standard input image data are crucial for training CNN architecture to obtain high performance (23). We recruited MR images from 202 rectal patients who underwent 3.0T MR scans, which ensured eligible input data with high signal noise ratio and spatial resolution. Furthermore, we manually labeled regions of contiguous normal rectal wall and lumen against tumor on T2WI images and the magnetic susceptibility artifacts on DWI image. This process is distinctive to previous work (12, 22) and helpful to confirm the boundary of VOI.

TABLE 2 | Performance of the radiomics-based models in predicting genotype (KRAS/NRAS/BRAF).

Imaging modality	VOI	Accuracy	Specificity	Sensitivity	AUC	*P*
T2WI	Manual	0.669	0.614	0.716	0.637	0.188
	DL	0.674	0.464	0.744	0.714	
DWI	Manual	0.776	0.731	0.809	0.872	0.181
	DL	0.711	0.678	0.736	0.816	
T2WI+DWI	Manual	0.829	0.803	0.847	0.906	0.676
	DL	0.783	0.661	0.882	0.887	

For each model, the mean performance from five-fold cross-validation is presented in this table. DeLong's test was used to compare the two AUCs of the manual-based model and the deep learning–based model for identical imaging modality. DL, deep learning; VOI, volumes of interest; AUC, area under the curve.

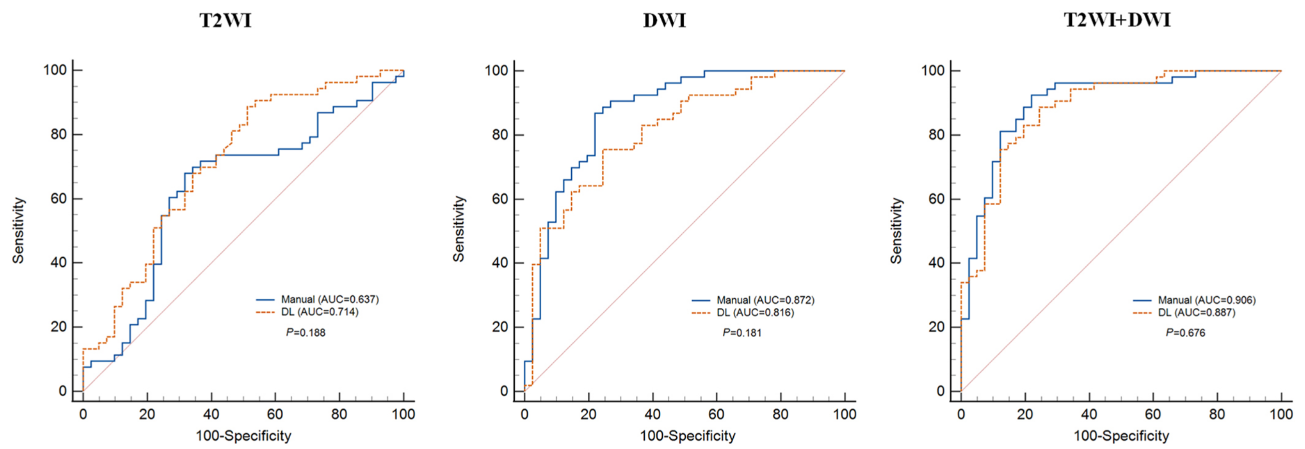


FIGURE 4 | Mean receiver operating characteristics (ROC) curve of five-fold cross validation for each radiomics-based model. DL, deep learning; AUC, area under the curve.

Though the T2WI had higher spatial resolution, the higher Dice was achieved on DWI. The noise, intensity non-uniformity, partial volume averaging, and tumor background contrast were key elements to influence the accuracy of segmentation (24). Compared to T2WI, the tumor background contrast and intensity uniformity on DWI were greater, which may facilitate the computer to identify and recognize the tumor region. As there was greater non-uniformity of intensity and low tumor background contrast on T2WI, especially in respect to muscle, bladder, and normal rectal wall, which may present similar signal intensity and texture to tumorous tissue, we found that after intensity histogram match there were still two samples that totally failed to give the correct segmentation (Disc=0). One of them put the segmentation label on the right piriformis, and the other put the segmentation label on the uterus (**Supplementary Figure 1**). Except for intensity non-uniformity and low tumor background contrast on T2WI, limited training samples may be another reason that contributes to failed segmentations.

The ultimate goal of auto segmentation is to facilitate clinical or experimental application. Since the genotype (KRAS/NRAS/BRAF) is strongly correlated with response to anti-EGFR therapies (25), we evaluated the reliability and usefulness of auto segmentation with radiomics analysis on these genotype predictions. No matter whether referring to single imaging modality or combined imaging modality, we found that the performance of genotype prediction is similar between manual-based and DL-based segmentation (**Table 2**). For example, the AUC is 0.906 for manual-based and 0.887 for DL-based VOI in combination of T2WI and DWI features on the test dataset (*P*=0.676). When referring to radiomics analysis, the genotype prediction performance of DWI is superior to that of T2WI, and combination modality surpasses any single imaging modality no matter whether it is manual-based VOI or DL-based VOI (**Table 2**). The KRAS/NRAS/BRAF are the downstream effectors of the EGFR signal pathway involved in tumor cell proliferation, differentiation, and invasion (26). Tumors with mutant genes more likely exhibit greater aggressiveness and angiogenesis, which will result in faster progress, worse survival, and lower apparent diffusion coefficient (ADC) value (27). The DWI can indicate the functional information of tissue by evaluation of water molecular mobility, which is estimated with ADC value, while T2WI are prone to indicate anatomic information, which might explain the higher genotype prediction performance of DWI compared to that of T2WI. Cui and his colleagues (8) developed a radiomics signature to predict KRAS mutations with moderate performance on T2WI (AUC=0.682 for internal validation and 0.714 for external validation), which is concordant to our genotype prediction performance with T2WI (AUC=0.714, DL-based VOI). We noted that for T2WI modality, the AUC of the radiomics-based model with DL-based VOI is higher than that of manual-based VOI (0.714 vs 0.637). In theory, the manual segmentation is the ground truth for radiomics analysis. So, the performance of the DL-based model should not be superior to manual-based VOI. To assess the difference of gene prediction performance between these two models, a Delong's test was used, and the result showed no statistical significance (*P*=0.181). We speculate that limited sample size may be one reason. On the other hand, DL-based VOI may contain some peritumoral region, which could exhibit an inflammatory response and tumor microinvasion. The inflammatory response and tumor microinvasion may provide additional information that is related to gene mutation. Meng et al. (28) investigated a radiomics-based model in predicting the KRAS-2 genotype based on multiparametric MRI (T1WI, T2WI, DWI, and DCE) with 0.651 of AUC in the validation cohort, which is slightly inferior to our combination model (T2WI+DWI, AUC=0.878 for manual VOI) and may be attributed to low signal noise ratio and spatial resolution of their 1.5T MR scanner. Several studies have demonstrated the value of CT radiomics (7) (AUC = 0.829) or texture analysis (AUC = 0.82) (29) or PET/CT (AUC = 0.684 ~ 0.75) (30) on genotype prediction of KRAS/NRAS/BRAF or KRAS alone. Compared with CT or

PET/CT, MRI can be of benefit with no concern about radiation exposure and contrast agent injection and simultaneously provide a wonderful detailed tissue contrast.

Though 202 patients were involved in our analysis, it is still necessary to validate this framework and compare it with different architectures, such as the recently developed Generally Nuanced Deep Learning Framework (31), in larger and diverse datasets. Currently, all segmentation acquired with deep learning architecture should be carefully reviewed before being submitted for further application, especially for making a radiotherapy plan. The requirement of high-quality annotated data is a great challenge for auto segmentation, which needs a standard imaging protocol, strict quality control, and accurate annotation. For rectum DWI, magnetic susceptibility artifact is the main obstacle that affects the accuracy of auto segmentation. Therefore, we labeled the artifact on DWI of the training dataset. If possible, labeling all anatomic structures and artifacts on the training dataset will definitely improve the performance of deep learning architecture, but that will be a huge workload. Considering the great performance of combined imaging modality on predicting genotype, further investigation of combining CT and MRI is needed.

CONCLUSIONS

In this study, 3D V-Net architecture provided reliable rectal cancer segmentation on T2WI and DWI compared with expert-based segmentation, and auto segmentation was subjected to radiomics analysis in the prediction of KRAS/NRAS/BRAF mutation status and may produce a good prediction result.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by institutional review board of Xijing hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

GZ and LC produced the manuscript. GZ and LC conceived and designed framework of this article. GZ, LC, AL, XP, JS, and YeH collected and analysed the data. JZ and YiH supervised this study and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by Key Research and Development Projects in Shaanxi, grant number 2018ZDXM-SF-059.

ACKNOWLEDGMENTS

We thank Huan Wang, PhD, for his constructive comments in revision. We thank Dr. Liz White for helping to proofread the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.696706/full#supplementary-material

REFERENCES

1. Sorich MJ, Wiese MD, Rowland A, Kichenadasse G, McKinnon RA, Karapetis CS. Extended RAS Mutations and Anti-EGFR Monoclonal Antibody Survival Benefit in Metastatic Colorectal Cancer: A Meta-Analysis of Randomized, Controlled Trials. *Ann Oncol* (2015) 26(1):13–21. doi: 10.1093/annonc/mdu378
2. Pietrantonio F, Petrelli F, Coinu A, Di Bartolomeo M, Borgonovo K, Maggi C, et al. Predictive Role of BRAF Mutations in Patients With Advanced Colorectal Cancer Receiving Cetuximab and Panitumumab: A Meta-Analysis. *Eur J Cancer* (2015) 51(5):587–94. doi: 10.1016/j.ejca.2015.01.054
3. NCCN Clinical Practice Guidelines in Oncology. *Colon Cancer, Version 4* (2020). Available at: https://www.nccn.org/professionals/physician_gls/default.aspx.
4. Sclafani F, Chau I, Cunningham D, Hahne JC, Vlachogiannis G, Eltahir Z, et al. KRAS and BRAF Mutations in Circulating Tumour DNA From Locally Advanced Rectal Cancer. *Sci Rep* (2018) 8(1):1445. doi: 10.1038/s41598-018-19212-5
5. Vymetalkova V, Cervena K, Bartu L, Vodicka P. Circulating Cell-Free DNA and Colorectal Cancer: A Systematic Review. *Int J Mol Sci* (2018) 19(11):3356. doi: 10.3390/ijms19113356
6. Kim SJ, Pak K, Kim K. Diagnostic Performance of F-18 FDG PET/CT for Prediction of KRAS Mutation in Colorectal Cancer Patients: A Systematic Review and Meta-Analysis. *Abdom Radiol* (2019) 44(5):1703–11. doi: 10.1007/s00261-018-01891-3
7. Yang L, Dong D, Fang M, Zhu Y, Zang Y, Liu Z, et al. Can CT-Based Radiomics Signature Predict KRAS/NRAS/BRAF Mutations in Colorectal Cancer? *Eur Radiol* (2018) 28(5):2058–67. doi: 10.1007/s00330-017-5146-8
8. Cui Y, Liu H, Ren J, Du X, Xin L, Li D, et al. Development and Validation of a MRI-Based Radiomics Signature for Prediction of KRAS Mutation in Rectal Cancer. *Eur Radiol* (2020) 30(4):1948–58. doi: 10.1007/s00330-019-06572-3
9. Owens CA, Peterson CB, Tang C, Koay EJ, Yu W, Mackin DS, et al. Lung Tumor Segmentation Methods: Impact on the Uncertainty of Radiomics Features for Non-Small Cell Lung Cancer. *PloS One* (2018) 13(10):e0205003. doi: 10.1371/journal.pone.0205003
10. Chen AM, Chin R, Beron P, Yoshizaki T, Mikaeilian AG, Cao M. Inadequate Target Volume Delineation and Local-Regional Recurrence After Intensity-Modulated Radiotherapy for Human Papillomavirus-Positive Oropharynx Cancer. *Radiother Oncol* (2017) 123(3):412–8. doi: 10.1016/j.radonc.2017.04.015
11. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Artificial Intelligence System Approaching Neuroradiologist-Level

Differential Diagnosis Accuracy at Brain MRI. *Radiology* (2020) 295(3):626–37. doi: 10.1148/radiol.2020190283

12. Wang J, Lu J, Qin G, Shen L, Sun Y, Ying H, et al. Technical Note: A Deep Learning-Based Autosegmentation of Rectal Tumors in MR Images. *Med Phys* (2018) 45(6):2560–4. doi: 10.1002/mp.12918
13. Milletari F, Navab N, Ahmadi S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, 25-28 Oct. 2016. IEEE (2016). p. 565–71. doi: 10.1109/3DV.2016.79
14. Jeon SH, Song C, Chie EK, Kim B, Kim YH, Chang W, et al. Delta-Radiomics Signature Predicts Treatment Outcomes After Preoperative Chemoradiotherapy and Surgery in Rectal Cancer. *Radiat Oncol* (2019) 14(1):43. doi: 10.1186/s13014-019-1246-8
15. Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. *J Clin Oncol* (2016) 34(18):2157–64. doi: 10.1200/JCO.2015.65.9128
16. Liu H, Zhang C, Wang L, Luo R, Li J, Zheng H, et al. MRI Radiomics Analysis for Predicting Preoperative Synchronous Distant Metastasis in Patients With Rectal Cancer. *Eur Radiol* (2019) 29(8):4418–26. doi: 10.1007/s00330-018-5802-7
17. Chen W, Liu B, Peng S, Sun J, Qiao X. Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. *Int J Biomed Imaging* (2018) 2018:2512037. doi: 10.1155/2018/2512037
18. Park JE, Ham S, Kim HS, Park SY, Yun J, Lee H, et al. Diffusion and Perfusion MRI Radiomics Obtained From Deep Learning Segmentation Provides Reproducible and Comparable Diagnostic Model to Human in Post-Treatment Glioblastoma. *Eur Radiol* (2020) 31:3127–37. doi: 10.1007/s00330-020-07414-3
19. Choi Y, Nam Y, Lee YS, Kim J, Ahn KJ, Jang J, et al. IDH1 Mutation Prediction Using MR-Based Radiomics in Glioblastoma: Comparison Between Manual and Fully Automated Deep Learning-Based Approach of Tumor Segmentation. *Eur J Radiol* (2020) 128:109031. doi: 10.1016/j.ejrad.2020.109031
20. Zhang G, Ma W, Dong H, Shu J, Hou W, Guo Y, et al. Based on Histogram Analysis: ADCaqp Derived From Ultra-High B-Value DWI Could be a Non-Invasive Specific Biomarker for Rectal Cancer Prognosis. *Sci Rep* (2020) 10 (1):10158. doi: 10.1038/s41598-020-67263-4
21. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *NeuroImage* (2006) 31 (3):1116–28. doi: 10.1016/j.neuroimage.2006.01.015
22. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, Lahaye MJ, Parmar C, Bakers FCH, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci Rep* (2017) 7 (1):5301. doi: 10.1038/s41598-017-05728-9
23. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial Intelligence in Radiology. *Nat Rev Cancer* (2018) 18(8):500–10. doi: 10.1038/s41568-018-0016-5
24. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol* (2019) 29(3):185–97. doi: 10.1016/j.semradonc.2019.02.001
25. De Roock W, Claes B, Bernasconi D, De Schutter J, Biesmans B, Fountzilas G, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA Mutations on the Efficacy of Cetuximab Plus Chemotherapy in Chemotherapy-Refractory Metastatic Colorectal Cancer: A Retrospective Consortium Analysis. *Lancet Oncol* (2010) 11(8):753–62. doi: 10.1016/S1470-2045(10)70130-3
26. Khan K, Valeri N, Dearman C, Rao S, Watkins D, Starling N, et al. Targeting EGFR Pathway in Metastatic Colorectal Cancer-Tumour Heterogeniety and Convergent Evolution. *Crit Rev Oncol Hematol* (2019) 143:153–63. doi: 10.1016/j.critrevonc.2019.09.001
27. Beckers RCJ, Lambregts DMJ, Lahaye MJ, Rao SX, Kleinen K, Grootscholten C, et al. Advanced Imaging to Predict Response to Chemotherapy in Colorectal Liver Metastases - A Systematic Review. *HPB* (2018) 20(2):120–7. doi: 10.1016/j.hpb.2017.10.013
28. Meng X, Xia W, Xie P, Zhang R, Li W, Wang M, et al. Preoperative Radiomic Signature Based on Multiparametric Magnetic Resonance Imaging for Noninvasive Evaluation of Biological Characteristics in Rectal Cancer. *Eur Radiol* (2019) 29(6):3200–9. doi: 10.1007/s00330-018-5763-x
29. Taguchi N, Oda S, Yokota Y, Yamamura S, Imuta M, Tsuchigame T, et al. CT Texture Analysis for the Prediction of KRAS Mutation Status in Colorectal Cancer *Via* a Machine Learning Approach. *Eur J Radiol* (2019) 118:38–43. doi: 10.1016/j.ejrad.2019.06.028
30. Mao W, Zhou J, Zhang H, Qiu L, Tan H, Hu Y, et al. Relationship Between KRAS Mutations and Dual Time Point (18)F-FDG PET/CT Imaging in Colorectal Liver Metastases. *Abdom Radiol* (2019) 44(6):2059–66. doi: 10.1007/s00261-018-1740-8
31. Pati S, Thakur SP, Bhalerao M, Baid U, Grenko CM, Edwards B, et al. GaNDLF: A Generally Nuanced Deep Learning Framework for Scalable End-To-End Clinical Workflows in Medical Imaging. *ArXiv* (2021), abs/2103.01006.

Conflict of Interest: Authors LC, AL and XP were employed by Shanghai United Imaging Intelligence Co., Ltd.

The remaining 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fonc.2021.792062

Convolutional Neural Network-Based Diagnostic Model for a Solid, Indeterminate Solitary Pulmonary Nodule or Mass on Computed Tomography

Ke Sun[1,2†], *Shouyu Chen*[3†], *Jiabi Zhao*[2†], *Bin Wang*[2], *Yang Yang*[2], *Yin Wang*[3], *Chunyan Wu*[4*] and *Xiwen Sun*[2*]

[1] Department of Radiology, Huashan Hospital, Fudan University, Shanghai, China, [2] Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, [3] Department of Computer Science and Technology, College of Electronics and Information Engineering, Tongji University, Shanghai, China, [4] Department of Pathology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

Purpose: To establish a non-invasive diagnostic model based on convolutional neural networks (CNNs) to distinguish benign from malignant lesions manifesting as a solid, indeterminate solitary pulmonary nodule (SPN) or mass (SPM) on computed tomography (CT).

Method: A total of 459 patients with solid indeterminate SPNs/SPMs on CT were ultimately included in this retrospective study and assigned to the train (n=366), validation (n=46), and test (n=47) sets. Histopathologic analysis was available for each patient. An end-to-end CNN model was proposed to predict the natural history of solid indeterminate SPN/SPMs on CT. Receiver operating characteristic curves were plotted to evaluate the predictive performance of the proposed CNN model. The accuracy, sensitivity, and specificity of diagnoses by radiologists alone were compared with those of diagnoses by radiologists by using the CNN model to assess its clinical utility.

Results: For the CNN model, the AUC was 91% (95% confidence interval [CI]: 0.83–0.99) in the test set. The diagnostic accuracy of radiologists with the CNN model was significantly higher than that without the model (89 *vs.* 66%, P<0.01; 87 *vs.* 61%, P<0.01; 85 *vs.* 66%, P=0.03, in the train, validation, and test sets, respectively). In addition, while there was a slight increase in sensitivity, the specificity improved significantly by an average of 42% (the corresponding improvements in the three sets ranged from 43, 33, and 42% to 82, 78, and 84%, respectively; P<0.01 for all).

Conclusion: The CNN model could be a valuable tool in non-invasively differentiating benign from malignant lesions manifesting as solid, indeterminate SPNs/SPMs on CT.

Keywords: neural network model, computed tomography, differential diagnosis, solid, indeterminate solitary pulmonary nodule, lung adenocarcinoma

Edited by:
Haibin Shi,
Soochow University, China

Reviewed by:
Huaiqiang Sun,
Sichuan University, China
Amit Mehndiratta,
Indian Institute of Technology Delhi, India

****Correspondence:***
Chunyan Wu
wuchunyan581@sina.com
Xiwen Sun
sunxiwen5256@163.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *09 October 2021*
Accepted: *19 November 2021*
Published: *21 December 2021*

Citation:
Sun K, Chen S, Zhao J, Wang B, Yang Y, Wang Y, Wu C and Sun X (2021) Convolutional Neural Network-Based Diagnostic Model for a Solid, Indeterminate Solitary Pulmonary Nodule or Mass on Computed Tomography. Front. Oncol. 11:792062. doi: 10.3389/fonc.2021.792062

1 INTRODUCTION

With the use of thoracic low-dose computed tomography (CT) for lung cancer screening, an increasing number of solitary pulmonary nodules (SPNs) or masses (SPMs) are deliberately or incidentally discovered. Solid SPNs are extremely common, and malignancy account for approximately 60% (range: 55–66%) (1, 2). Data from the Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial indicated that SPMs were highly predictive of malignancy (odds ratio, 10.3; 95% confidence interval [CI], 2.46–43.38) (3). Solid malignant lesions are related to rapid cancer growth and high risks of recurrence and metastasis, despite their small size (4, 5). Therefore, the most crucial task for radiologists and clinicians is to accurately determine the natural history of the lesions. Surgery is the diagnostic gold standard and definitive treatment for malignant cases. However, 25–46% of patients with SPNs have benign disease despite a preoperative suspicion of cancer, and an incorrect diagnosis results in unnecessary invasive resection and monetary and time costs (6, 7).

High-resolution computed tomography (HRCT) can non-invasively provide specific information about pulmonary lesions (8). However, there are challenges associated with the visual assessment of CT images. First, a series of CT images consist of hundreds of slices; radiologists have to browse through these slices and carefully consider them, which is time-consuming, tedious, and subjective. Second, visual evaluations are inadequate to distinguish benign from malignant lesions manifesting as solid, indeterminate SPNs or SPMs because of the considerable overlap in the radiographic characteristics of these lesion types (**Figure 1**). For example, 21–58% of malignant lesions have smooth edges, and approximately 25% of benign nodules are irregularly shaped with spiculated or lobulated margins (9–12). In this study, a solid, indeterminate lesion was defined as a non-calcified lesion or a lesion without features strongly suggestive of a benign etiology, usually greater than 8 mm in size (13).

Recently, machine learning has shown outstanding capabilities as one of the most promising tools for the detection, diagnosis, and differentiation of lung lesions. Over the years, two computational strategies have been developed to predict the malignancy of lung lesions on CT images: radiomics based on quantitative radiological image features and deep learning methods such as those based on cascade convolutional neural networks (CNNs). The extraction of radiomics features relies heavily on accurate lesions boundary outline, and predictive models are built based on a prior knowledge of which features are significant. Whereas CNNs could automatically extract potential features beyond human perception from medical images to predict whether a lesion is benign or malignant by amplifying aspects of the input images that are important for discrimination and suppressing irrelevant variations (14). When successfully applied, it is expected to improve diagnostic accuracy and reduce unnecessary invasive procedures and costs and anxiety of patients. Several studies have revealed the predictive value of CNNs and the promising prospects they afford for lung lesion differentiation (15–18). However, (1) these models lack interpretability and are often referred to as "black boxes", which renders them difficult for the users to understand; (2) no specific emphasis has been given to distinguishing benign from malignant lesions manifesting as solid, indeterminate pulmonary lesions. Therefore, this study aimed to develop an interpretable CNN-based non-invasive diagnostic model for solid, indeterminate SPNs or SPMs on CT and to evaluate its clinical utility.

2 MATERIALS AND METHODS

The retrospective study was approved by the ethics committee of "Shanghai Pulmonary" Hospital. The informed consent requirement was waived.

2.1 Study Population

We retrospectively included 459 consecutive patients with solid, indeterminate SPNs or SPMs on CT between January 2018 and December 2018. Patients who met the following criteria were included: (1) presence of a primary intrapulmonary lesion; (2) the diameter of an existing lesion of >8 mm [because pure-solid nodules measuring <8 mm has the relatively low prevalence of malignancy, and the risks of surgical diagnosis usually outweigh the benefits (13); thus, the Fleischner Society guidelines recommend routine follow-up for management (19), and additionally, in our hospital, one of the criteria for surgical excision is a diameter greater than 8 mm (20)]; (3) histologically confirmed diagnosis after surgical resection; and (4) preoperative CT slice thickness of 1–1.25 mm. The exclusion criteria were as follows: (1) a clearly benign diagnosis based on the initial CT reports; (2) a history of malignancy; (3) lesions with calcification regardless of type; (4) obvious artefacts on CT images. Eligible patients were sorted randomly, and the benign and malignant groups were divided into the training (n=366), validation (n=46), and test (n=47) sets according to the 8:1:1 ratio for model learning, respectively, shown in **Figure 2**.

2.2 CT Parameter Acquisition and Image Annotation and Interpretation

All patients underwent Chest CT examinations before surgery in our institution, and the detailed scanning parameters are shown in **Supplement Table 1**. Two thoracic radiologists (with 3 and 7 years of work experience) detected the location of pulmonary lesions, marked their coordinates (X, Y, and Z axes), and measured their diameters on the section that displayed the longest diameter of the lesion. When annotations differed, radiologists discussed them until consensus was achieved.

Additionally, these lesions were evaluated for shape (regular or irregular), the presence of spiculation, lobulation, and pleural retraction. The mean CT density was calculated by measuring

Abbreviations: CT, computed tomography; SPNs, solitary pulmonary nodules; SPMs, solitary pulmonary masses; CI, confidence interval; HRCT, high-resolution computed tomography; CNN, cascade convolutional neural network; ROI, region of interest; HU, Hounsfield unit; 3D, three-dimensional; CAM, class activation map; SD, standard deviations; ROC, receiver operating characteristic; AUC, area under the ROC; PSPs, pulmonary sclerosing pneumocytomas; FOP, focal organizing pneumonia.

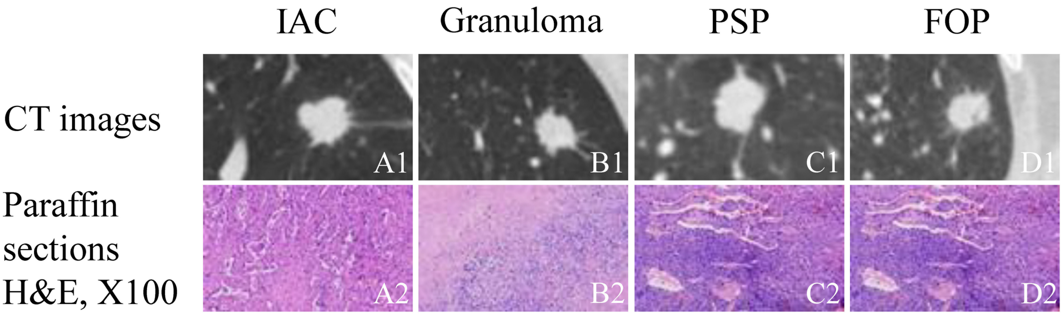


FIGURE 1 | Examples of solid, indeterminate SPN/SPMs without features strongly suggestive of a benign etiology. **(A1)** Invasive adenocarcinoma (IAC); **(B1)** granuloma; **(C1)** pulmonary sclerosing pneumocytoma (PSP); **(D1)** focal organizing pneumonia (FOP). **(A2–D2)** paraffin section (hematoxylin and eosin [H&E], 100 ×) of IAC, granuloma, PSP, and FOP, respectively.

the average CT value of the region of interest (ROI) that carefully placed in an area away from vessels, bronchi, and necrosis. The readings were interpreted using Radiant software (http://radiantviewer.com) with the lung window setting (window level, −450 Hounsfield unit [HU]; width, 1,500 HU) and mediastinal window setting (window level, 40 HU; width, 400 HU). Based on the experienced evaluation, the other two radiologists (with 4 and 9 years of experience in reading thoracic CT scans, respectively), uninformed of the pathological results, made a diagnosis. In case of a discrepancy between the two radiologists, a third radiologist with an experience of more than 29 years in thoracic CT made the final decision.

2.3 CNN Model Construction

2.3.1 Image and Data Preprocessing

The voxel spatial resolution of all patients' raw CT were standardized on all three axes, to 0.6 × 0.6 × 0.6 mm^3 each voxel. Then small three-dimensional (3D) tensor with size of 128×128×128 voxels centered at each nodule is extracted using corresponding coordinates annotation. The size ensured that each nodule was entirely covered. In the training phase, it was necessary to randomly rotate the tensor at arbitrary angle in the 3D space, as a data augmentation method. Then we selected three orthogonal slices passing through the center point and stacked them, resulting in a 3×128×128 tensor. Furthermore, we cropped a 3×104×104 sub-region that could completely cover all lesions, and resized it to the voxels of 3×224×224, as a data augmentation method also. Finally, the CT value interval was clipped to [−1,100 HU, 100 HU], and the result was further linearly mapped to the value interval [0, 1]. Each 3×224×224 tensor represented one patient in the network pipeline. The preprocessing was shown in **Figure 3**.

2.3.2 The Structure of the CNN Model

Figure 4 shows the pipeline of benign and malignant prediction for solid, indeterminate SPNs or SPMs on CT. ResNet was used as the basis of the deep learning model (21). Specifically, the selected network was ResNet-101. As a transfer learning method, ResNet's weights parameter pre-trained on the ImageNet image

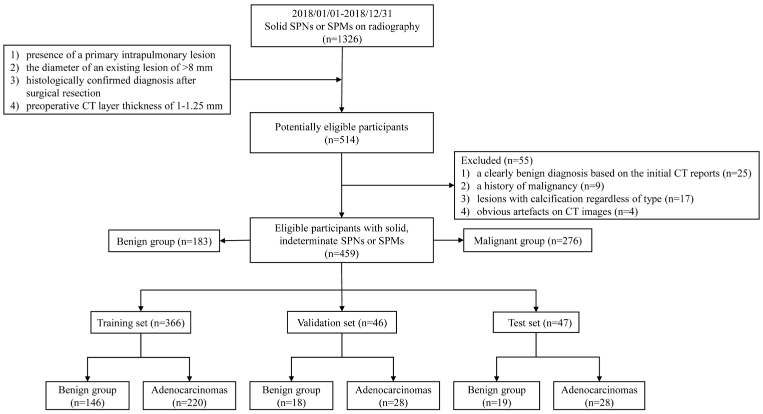

2018/01/01-2018/12/31
Solid SPNs or SPMs on radiography
(n=1326)
1) presence of a primary intrapulmonary lesion
2) the diameter of an existing lesion of >8 mm
3) histologically confirmed diagnosis after surgical resection
4) preoperative CT layer thickness of 1-1.25 mm
Potentially eligible participants
(n=514)
Excluded (n=55)
1) a clearly benign diagnosis based on the initial CT reports (n=25)
2) a history of malignancy (n=9)
3) lesions with calcification regardless of type (n=17)
4) obvious artefacts on CT images (n=4)
Benign group (n=183)
Eligible participants with solid, indeterminate SPNs or SPMs
(n=459)
Malignant group (n=276)
Training set (n=366)
Validation set (n=46)
Test set (n=47)
Benign group (n=146)
Adenocarcinomas (n=220)
Benign group (n=18)
Adenocarcinomas (n=28)
Benign group (n=19)
Adenocarcinomas (n=28)

FIGURE 2 | Flow chart of inclusion and exclusion criteria for eligible patients and specific allocations in the train, validation, and test sets.

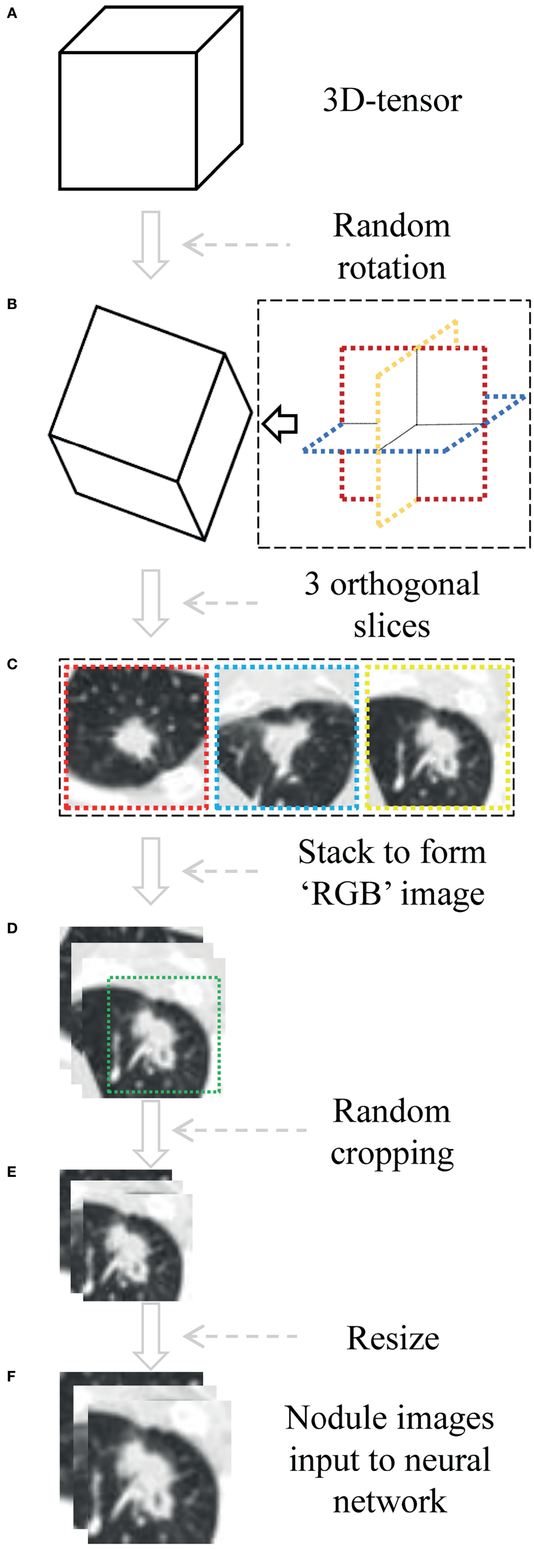


FIGURE 3 | The process of data preprocessing. **(A)** Three-dimensional (3D) tensor was obtained from the original CT sequence according to nodule coordinates labeled by radiologist. **(B)** The tensor is rotated at arbitrary angle around its center point. **(C, D)** Three orthogonal slices spanning center point were extracted and stacked to form a pseudo-RGB map (3×128×128 tensor). **(E)** Random cropping with 3×104×104 subregion. **(F)** Nodule images were resized to voxels of 3×224×224.

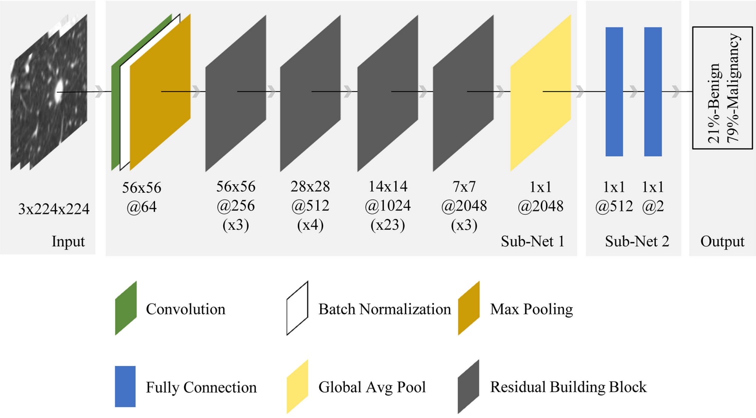


FIGURE 4 | End-to-end CNN model illustration. For the input nodule images (from the left side), the neural network made the prediction (right side) and outputted two values, representing benign and malignant probabilities (summed to 1). The final diagnosis of each nodule by the model depended on which class was predicted with a probability greater than 50%. The architecture was composed of convolution, batch normalization, max pooling, fully connection, global average pooling, and residual building block. Sub-net 1 was pretrained on ImageNet dataset with ~15 million neutral images, while sub-net 2 was trained from scratch. The 56x56@256 (x3) below the first residual building block meant the spatial size and number of channels of the output feature map in this block were 56x56 and 256 respectively, while x3 meant the block contained 3 residual units.

dataset was loaded to initialize sub-network 1 of our model (22). In sub-network 2, the last two layers in the original network were replaced by two fully connected layers with 512 and 2 output nodes, respectively, and their weight parameters were initialized randomly. We chose to use 2D-CNN rather 3D-CNN for the following reasons: (1) the number of CT cases is small and training CNNs from scratch on top of this would lead to overfitting, and there are currently no pre-trained 3D-CNN model weights available on large publicly available 3D CT datasets; (2) the data augmentation method used in the paper that rotated in 3D space could also help 2D-CNN capture the 3D features of nodules, and better prediction results could be achieved using 2D-CNN model with transfer learning. Furthermore, to increase the generalizability of the model and avoid overfitting, mix-up algorithm was adopted (23).

2.3.3 Experiment Parameter Setting

The train dataset was used to train the deep learning algorithm, a separate validation dataset to tune parameter, and the test dataset to assess the final model. During the training stage, only weight parameters in the last two fully connected layers and all batch normalization layers in the network were trained for 1 epoch, and others remained unchanged. This can be considered as a warm-up training. Then the entire model was trained for additional 60 epochs. This process simultaneously optimized all network layers, making the lower convolutional layer more suitable for edge and corner features in CT data, as well as for the specific data distribution resulting from our combination of orthogonal slices. The weights corresponding to the epoch with the lowest validation loss were chosen as the optimal model and saved. The model used Adam as weights optimizer and cross-entropy as loss function (24). The learning rate was 1e-2, and weight decay was 5e-5. One cycle strategy was used to adjust the learning rate during model training (25). The dropout probabilities of the last two fully connected layers in the model were set 0.25 and 0.5, respectively. A batch size of 64 was used. It took about 5 s to train the neural network on all 366 training samples (tensor size: 366×3×224×244) for one epoch. See the code for the detailed procedure. Code implementation was based on the fastai framework (26) and available online https://github.com/DrIsDr/TJU_Chen_SK.

2.3.4 Visualization of the CNN Model

The CNN models were often referred to "black-box" technology due to lack of interpretability, making it difficult for users to understand the inference procedure. We used the class activation map (CAM) to visualize the discriminative process of the neural network (27), and the results are shown in **Figure 5**. The CAM could generate the response heatmaps to reversely deduce the process of the model making diagnosis. Red areas had the highest activation value, which suggested that the model mainly extracted diagnostic characteristics from the region, whereas the blue areas had the lowest activation value, meaning that less discriminative features were found in this region.

As can be seen in the **Figure 5**, the CNN model produced high activation value (red areas) in the regions where the nodule was located and adjacent to the nodule only when the nodule was correctly classified. In other words, the model captured the internal and external features of nodules to make a diagnosis. More examples are shown in the **Supplementary Figure 1**.

2.5 Statistical Analysis

Baseline characteristics and image information of the participants were summarized as mean ± standard deviations (SD) values for continuous variables, and as frequency and percentage for categorical variables. Statistical significance was

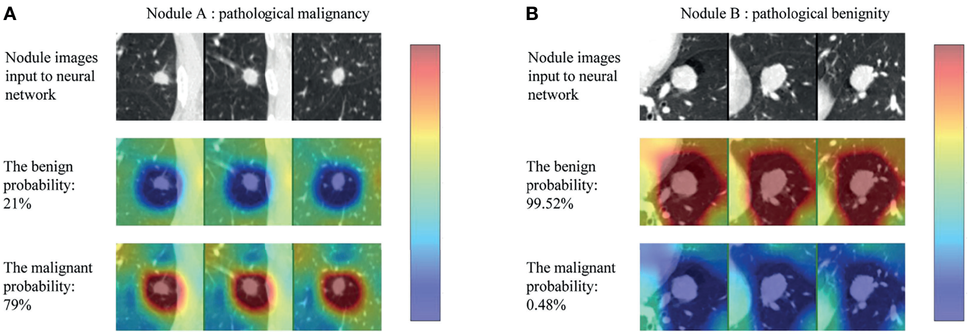

A
Nodule A : pathological malignancy
Nodule images input to neural network
The benign probability: 21%
The malignant probability: 79%
B
Nodule B : pathological benignity
Nodule images input to neural network
The benign probability: 99.52%
The malignant probability: 0.48%

FIGURE 5 | Class activation map (CAM) for two example nodules (nodule **A** and **B**) in the test set. For each nodule, the first row (nodule images input to neural network) represented the three views of each nodule, the second (the benign probability) and third row (the malignant probability) represented the corresponding response heatmaps when the model classifies the nodule as benign and malignant, respectively (red regions are of highest interest and blue lowest).

tested using Student's t-test, Welch's t-test, Mann-Whitney U-test, and Kruskal-Wallis for continuous variables as appropriate, and Chi-square test for categorical variables. A p-value < 0.05 was considered to be statistically significant. The predictive results of CNN model were compared with the pathological gold standard. The diagnostic performance of the CNN model was described using the area under the receiver operating characteristic (ROC) curve (AUC) with 95% CI. The mean value of accuracy, sensitivity, and specificity of diagnoses by radiologists alone were compared with those of diagnoses by radiologists with the assistance of the CNN model to evaluate its clinical utility. All statistical analyses are based on SPSS 20.0 software (SPSS Inc., Chicago, IL, USA) and R version 3.6.3 (R foundation for Statistical Computing).

3 RESULTS

3.1 Baseline Characteristics

A total of 459 patients with solid, indeterminate SPNs or SPMs were included. Of the 459 patients, 183 had benign disease (83 males and 100 females; mean age, 53.67 ± 12.33) and 276 had malignant disease (151 males and 125 females; mean age, 60.53 ± 9.30). Among the 183 benign cases, there were 124, 55, and 4 granulomas, pulmonary sclerosing pneumocytomas (PSPs), and focal organizing pneumonia (FOP), respectively. The subtype of malignancy only included lung adenocarcinomas.

The clinical baseline characteristics and image features are listed in **Table 1**. Between the benign and malignant groups, clinical variables, such as age (P<0.01) and gender (P=0.05), demonstrated statistical difference. However, no statistically significant association was observed in terms of radiological features.

3.2 Performance of the CNN Model

The model demonstrated superior performance in the train set (AUC: 0.94, 95% CI: 0.92–0.96); the results in the validation and test sets showed slightly lower but still satisfactory differentiation performance (validation set: AUC 0.88, 95% CI: 0.78–0.99; test set: AUC 0.91, 95% CI: 0.83–0.99) (**Figure 6**). The total concordance rates between the CNN model and final pathological assessments generated by the final paraffin section in the train, validation, test cohorts were 87% (318/366), 83% (38/46), and 83% (39/47), respectively (**Supplement Table 1**). The sensitivity and specificity were 89% (95% CI: 0.85–0.92) and 84% (95% CI:0.80–0.87) in the train set, 86% (95% CI: 0.73–0.93) and 78% (95% CI: 0.64–0.88) in the validation set, and 86% (95% CI: 0.73–0.93) and 79% (95% CI: 0.65–0.88) in the test set (**Table 2**).

3.3 Clinical Utility of the CNN Model

Three radiologists blinded to the pathological results twice assessed the benignity or malignancy of each patient and made a final decision in consensus. The average time required for diagnosing each patient was 3 min. The diagnostic accuracy of radiologists alone was lower than that of the CNN model in all patients (train set: 66 *vs.* 87%, P<0.01; validation set: 61 *vs.* 83%, P=0.02; test set: 66 *vs.* 83%, P=0.06). When radiologists used the CNN model, their diagnostic accuracy was higher than that achieved by radiologists alone (train set: 89 *vs.* 66%, P<0.01; validation set: 87 *vs.* 61%, P<0.01; test set: 85 *vs.* 66%, P=0.03) (**Supplement Table 2**). Additionally, specificities increased significantly, by an average of 42% (train set: from 43 to 82%; validation set: from 33 to 78%; test set: from 42 to 84%; all P-values < 0.01); and sensitivities improved slightly (train set: 81 *vs.* 95%, P<0.01; validation set: 79 *vs.* 93%, P=0.04; test set: 82 *vs.* 89%, P=0.37) (**Table 3**). Thus, the CNN model could help radiologists to enhance the capability of distinguishing benign from malignant lesions with radiographic solid, indeterminate SPN or SPM characteristics at all three levels of CT expertise, effectively preventing misdiagnosis.

4 DISCUSSION

Deep CNN is a type of deep learning approach in which computers are not explicitly programmed but can perform tasks by analyzing relationships of existing data. In this retrospective study, our CNN model achieved better accuracy than three radiologists in differentiation between benignity and

TABLE 1 | The baseline characteristics and imaging information of patients included in the study.

Variables	Total (n=459)	Benign (n=183)	Malignant (n=276)	P-value
Age, mean ± SD, y	57.80 ± 11.12	53.67 ± 12.33	60.53 ± 9.30	<0.01
Gender, n (%)				0.05
Male	234 (51)	83 (45)	151 (55)	
Female	225 (49)	100 (55)	125 (45)	
Image information				
Diameter, n (%)				0.14
≤30 mm	379 (83)	157 (86)	222 (80)	
>30 mm	80 (17)	26 (14)	54 (20)	
Tumors location, n (%)				0.07
RUL	123 (27)	40 (22)	83 (30)	
RML	45 (10)	19 (10)	26 (9)	
RLL	97 (21)	50 (27)	47 (17)	
LUL	111 (24)	41 (22)	70 (25)	
LLL	83 (18)	33 (18)	50 (18)	
Shape, n (%)				0.50
Regular	103 (22)	44 (24)	59 (21)	
Irregular	356 (78)	139 (76)	217 (79)	
Lobulation, n (%)				0.59
Presence	290 (63)	117 (64)	173 (63)	
Absence	168 (37)	66 (36)	102 (37)	
Spiculation, n (%)				
Presence	244 (53)	94 (51)	150 (54)	0.53
Absence	215 (47)	89 (49)	126 (46)	
Pleural retraction, n (%)				0.92
Presence	232 (51)	93 (51)	139 (50)	
Absence	227 (49)	90 (49)	137 (50)	
CT value, mean ± SD, HU	31.24 ± 24.86	33.07 ± 27.22	30.03 ± 23.14	0.20

The data are expressed as mean ± standard deviations for continuous variables, and the frequency and percentage for categorical variables.
A p-value < 0.05 was supported to be statistically significant.

malignancy for solid, indeterminate SPNs or SPMs. When radiologists used the CNN model, the mean accuracy was 87%, and the specificity improved by 42%, which would have facilitated timely diagnosis and treatment for lung cancer and avoided unnecessary excision for benign cases by a non-invasive, highly efficient, and reproducible method. Furthermore, to enhance interpretability, we used visualization techniques to analyze the process of the CNN model classification. To the best of our knowledge, this is the first attempt to differentiate radiographically solid, indeterminate lesions using interpretable CNN technology based on thin-section CT scans.

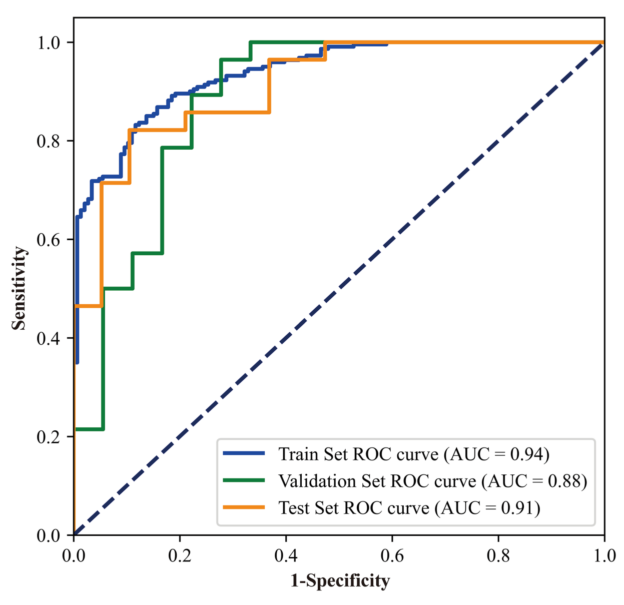


FIGURE 6 | The receiver operating characteristic curves of the CNN model used in this study.

The solid SPN is an extremely common type of tumor, and approximately 60% of solid SPNs are malignant (1, 2). Published studies have reported that nodal metastasis and intrapulmonary and extrapulmonary diffusion could be found in malignant solid lesions, even in subcentimeter small nodules (4). Thus, differentiation of benign and malignant lesions is the most critical step for patient management.

Chest CT examinations can provide specific information about morphological and density characteristics and are helpful to estimate the probability of malignancy for pulmonary solid lesions. Multiple studies have revealed that spiculation, lobulation, irregular shape, and pleural retraction are associated with malignancy, whereas lesions with a regular shape and the smooth margin are more likely to be benign (28, 29). However, in our dataset, no radiologically available features were observed (**Table 1**), which means that trained radiologists have difficulty distinguishing the nature history of solid, indeterminate solitary pulmonary by visual assessment alone. The overlap of radiographic characteristics does not seem too unusual. As noted previously, for pulmonary lesions with smooth edges, the risk of malignancy was

TABLE 2 | Predictive performance of the CNN model.

	Train set	Validation set	Test set
AUC	94 (0.92–0.96)	88 (0.78–0.99)	91 (0.83–0.99)
ACC	87 (0.83–0.90)	83 (0.70–0.91)	83 (0.70–0.91)
SE	89 (0.85–0.92)	86 (0.73–0.93)	86 (0.73–0.93)
SP	84 (0.80–0.87)	78 (0.64–0.88)	79 (0.65–0.88)

All values shown as % (95% confidence interval).
CNN, convolutional neural network; AUC: area under curve; ACC, accuracy; SE, sensitivity; SP, specificity.

approximately 35% (range: 21–58%) (9–12). Chu et al. reported that 95% (214/225) of solid cancerous nodule had a regular shape (30). Also, Zerhouni et al. recorded that 25% of benign nodules showed irregular margins with lobulation or spiculation, and only 18% of these lesions were correctly assessed on CT (9). In addition, in the study by Xu et at., lung cancer risk was absent in solid indeterminate nodules attached to the pleural or a fissure during 1 year of follow-up (31).

Thus, it is very pivotal to differentiate benign from malignant solid, indeterminate SPNs or SPMs by using a new approach to overcome the naked limitation. However, some studies revealed and exploited the massive potential of image features that may be visually imperceptible to even very experienced thoracic radiologists and can be extracted from CT scans by using (1) radiomics methods or (2) deep learning approaches based on CNNs (32–34). Both methods have been widely used to classify and identify the natural history of sub-solid nodules including the part-solid and pure ground glass nodules, scoring tremendous achievements (35–39). Nevertheless, few studies have focused on the differentiation of solid pulmonary lesions. Shen et al. established a multiclassifier fusion based on radiomic features, including geometric features, textures features, gray-level features, and wavelet features to predict benign and malignant primary solid nodules, achieving an AUC of 0.915 in the test set (40). However, not all benign cases in this study were pathologically confirmed, a stable 2-year follow-up period does not guarantee its benign nature. In addition, radiomics methods that extract quantitative biological features are limited by prior knowledge of significant characteristics, which may be unbefitting for pulmonary lesions with considerable overlapping features. The CNN method could simplify the redundancies and learn discriminating features directly from CT images, facilitating greater reproducibility. In this study, our CNN model in the test set had an AUC of 0.91, comparable with the previously reported value, which indicated good performance. The specificity of our model was significantly higher than that of the three radiologists. In fact, most benign lesions in our study mimicked the morphological characteristics of lung cancer. Radiologists are prone to classifying these lesions as malignant in clinical practice, yielding high sensitivity with low specificity. However, when radiologists used the CNN model, their specificity improved significantly by 42% while maintaining the high sensitivity.

Additionally, the CNN model is highly efficient in distinguishing benignity from malignancy for solid, indeterminate lesions. Radiologists spent an average of 3 min to read and interpret a set of CT images of one patient, while the CNN model could process the 366 patient images in just 5 s. Moreover, in routine clinical practice, radiologists usually need to review and compare prior CT images to make a diagnosis, which would require more time despite yielding higher accuracy. In our model, on the basis of coordinate information, we adopted a supervised learning method, guided the neural network model to extract features layer-by-layer from CT images of interest, constantly enhanced the intensity of feature abstraction, and finally output the result of the prediction. Thus, we used an end-to-end computational method that could greatly simplify the traditional workflow.

A limitation of our model is the overfitting issue caused by the single-institute small data size. To compensate for this limitation, we utilized the pretrained network on ImageNet that included millions of natural images, in a process termed transfer learning. Although there is no intuitive approach for using a pretrained model with non-medical images for differentiation of medical images, some features including the edges, corners, orientations, and textures are generic. We compared the performance of the CNN model with or without pretrained procedure using the same experimental parameters, and the results showed the pretrained CNN model performed much better than the untrained one (**Supplementary Figure 2**). In addition, data augmentation was also used to resolve this problem. Randomly rotating nodules/masses in 3D space,

TABLE 3 | Comparison of the diagnostic performance of radiologists without and with the CNN model.

	Training set		Validation set		Test set	
	Radiologists alone	Radiologists with CNN	Radiologists alone	Radiologists with CNN	Radiologists alone	Radiologists with CNN
ACC	66 (0.61–0.71)	89 (0.85–0.92)	61 (0.47–0.74)	87 (0.74–0.94)	66 (0.52–0.78)	85 (0.72–0.93)
SE	81 (0.77–0.85)	95 (0.92–0.97)	79 (0.65–0.88)	93 (0.82–0.98)	82 (0.69–0.90)	89 (0.77–0.95)
SP	43 (0.38–0.48)	82 (0.78–0.86)	33 (0.21–0.47)	78 (0.64–0.88)	42 (0.29–0.56)	84 (0.71–0.92)
FPV	57 (0.52–0.62)	19 (0.15–0.23)	67 (0.53–0.79)	22 (0.12–0.36)	58 (0.44–0.71)	16 (0.08–0.29)
FNV	19 (0.15–0.23)	6 (0.04–0.09)	21 (0.12–0.35)	7 (0.02–0.18)	18 (0.10–0.31)	11 (0.05–0.23)

All values shown as % (95% confidence interval).
CNN, convolutional neural network; ACC, accuracy; SE, sensitivity; SP, specificity; FPV, false positive value; FNV, false negative value.

extracting three orthogonal slices to form pseudo-RGB map, cropping, and resizing the input images for neural network could help 2D-CNN capture rich 3D features. In the future, a larger multicenter study should be used to validate this model and improve the performance of this algorithm.

Other limitations should be mentioned. Lesions were not automatically detected, but based on radiologist annotations, which would lead to interobserver variability and error propagation to CNN model because of this process. However, to reduce this bias, all lesions were marked in consensus by two experienced radiologists. Furthermore, we used the Grad-CAM method to visualize the intermediate variables generated by the trained model for the prediction process of the images. Given an image patch, the model does focus on the nodule, demonstrating that the region of interest used by the model for feature recognition is correct and that such interpretable analysis is appropriate for the form of our annotation (which includes nodule location and class) currently provided. At present, the design of CNN algorithms and the abundance of clinical data are mutually reinforcing. In the future, as more data become available and finer-level annotation information becomes more widespread, CNNs can be more useful for clinical applications. Considering actual clinical limitations, the design of our cohort was restricted to allow differentiation between adenocarcinomas and benign diseases including granulomas, PSP, and FOP. Actually, it makes sense to use CNN model to further predict the results of benign lesions for subclassification. However, in our study, we did not perform this task. Understandably, the multi-classification tasks for benign lesions are difficult due to the disparity in sample distribution of benign subtypes (124 granulomas; 55 PSPs; 4 FOPs), as a well-performing model requires a large number of sample data of each type of disease. We plan to conduct a more in-depth evaluation of the application of CNN model to multi-classification tasks based on large samples in the upcoming studies.

In conclusion, we established a CNN model based on CT images that can serve as a valuable tool for radiologists to differentiate radiographic solid, indeterminate SPNs or SPMs. Moreover, a visualization procedure was presented to enhance interpretability of CNN model.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/DrIsDr/TJU_Chen_SK.

AUTHOR CONTRIBUTIONS

KS: Conception and design, data analysis and interpretation, manuscript writing, final approval of manuscript. SC: Data analysis and interpretation, manuscript writing, final approval of manuscript. JZ: Collection and assembly of data, manuscript writing, final approval of manuscript. BW: Manuscript writing, final approval of manuscript. YY: Manuscript writing, final approval of manuscript. YW: Data analysis and interpretation, manuscript writing, final approval of manuscript. CW: Provision of study materials or patients, manuscript writing, final approval of manuscript. XS: Conception and design, administrative support, manuscript writing, final approval of manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of Shanghai (grant Number 21Y11910400). The funders had no role in the study design, data collection, data analysis, and writing the manuscripts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.792062/full#supplementary-material

Supplementary Figure 1 | The more represented results for using CAM to visualize the discriminative process of the neural network. (nodules **A–D**): examples of CNN model incorrectly predicting the benignity and malignancy of nodules. (nodules **E–J**) examples of CNN model for accurate prediction of nodal benignity and malignancy. As shown in nodule H, the ROI heat values used to determine malignancy is low, but the ROI heat values used to determine non-benign is high, which can be explained by the fact the CNN determines this case as malignant by referring more to the surrounding area than to the nodal region. What is more, we can learn from nodule I that the CT images of this case obviously have texture noise that is not present in other cases, but the CNN still correctly detects the ROI and makes a judgment, which reflects the robustness of the CNN model. CAM, class activation map; CNN, convolutional neural network; ROI, region of interest; CT, computed tomography.

REFERENCES

1. She Y, Zhao L, Dai C, Ren Y, Jiang G, Xie H, et al. Development and Validation of a Nomogram to Estimate the Pretest Probability of Cancer in Chinese Patients With Solid Solitary Pulmonary Nodules: A Multi-Institutional Study. *J Surg Oncol* (2017) 116:756–62. doi: 10.1002/jso.24704
2. Harders SW, Madsen HH, Rasmussen TR, Hager H, Rasmussen F. High Resolution Spiral CT for Determining the Malignant Potential of Solitary Pulmonary Nodules: Refining and Testing the Test. *Acta Radiol* (2011) 52:401–9. doi: 10.1258/ar.2011.100377
3. Tammemagi MC, Freedman MT, Pinsky PF, Oken MM, Hu P, Riley TL, et al. Prediction of True Positive Lung Cancers in Individuals With Abnormal Suspicious Chest Radiographs—A Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Study. *J Thorac Oncol* (2009) 4:710–21. doi: 10.1097/JTO.0b013e31819e77ce
4. Hattori A, Matsunaga T, Hayashi T, Takamochi K, Oh S, Suzuki K. Prognostic Impact of the Findings on Thin-Section Computed

Tomography in Patients With Subcentimeter Non-Small Cell Lung Cancer. *J Thorac Oncol* (2017) 12:954–62. doi: 10.1016/j.jtho.2017.02.015

5. Hattori A, Matsunaga T, Takamochi K, Oh S, Suzuki K. Locoregional Recurrence After Segmentectomy for Clinical-T1aN0M0 Radiologically Solid Non-Small-Cell Lung Carcinoma. *Eur J Cardiothorac Surg* (2017) 51:518–25. doi: 10.1093/ejcts/ezw336
6. Yonemori K, Tateishi U, Uno H, Yonemori Y, Tsuta K, Takeuchi M, et al. Development and Validation of Diagnostic Prediction Model for Solitary Pulmonary Nodules. *Respirology* (2007) 12:856–62. doi: 10.1111/j.1440-1843.2007.01158.x
7. Li Y, Chen KZ, Wang J. Development and Validation of a Clinical Prediction Model to Estimate the Probability of Malignancy in Solitary Pulmonary Nodules in Chinese People. *Clin Lung Cancer* (2011) 12:313–9. doi: 10.1016/j.cllc.2011.06.005
8. Zhou Z, Zhan P, Jin J, Liu Y, Li Q, Ma C, et al. The Imaging of Small Pulmonary Nodules. *Transl Lung Cancer Res* (2017) 6:62–7. doi: 10.21037/tlcr.2017.02.02
9. Zerhouni EA, Stitik FP, Siegelman SS, Naidich DP, Sagel SS, Proto AV, et al. CT of the Pulmonary Nodule: A Cooperative Study. *Radiology* (1986) 160:319–27. doi: 10.1148/radiology.160.2.3726107
10. Tozaki M, Ichiba N, Kunihiko Fukuda M. Dynamic Magnetic Resonance Imaging of Solitary Pulmonary Nodules: Utility of Kinetic Patterns in Differential Diagnosis. *J Comput Assist Tomogr* (2005) 29:13–9. doi: 10.1097/01.rct.0000153287.79730.9b
11. Siegelman SS, Khouri NF, Leo FP, Fishman EK, Braverman RM, Zerhouni EA. Solitary Pulmonary Nodules: CT Assessment. *Radiology* (1986) 160:307–12. doi: 10.1148/radiology.160.2.3726105
12. Swensen SJ, Morin RL, Schueler BA, Brown LR, Cortese DA, Pairolero PC, et al. Solitary Pulmonary Nodule: CT Evaluation of Enhancement With Iodinated Contrast Material—A Preliminary Report. *Radiology* (1992) 182:343–7. doi: 10.1148/radiology.182.2.1732947
13. Gould MK, Donington J, Lynch WR, Mazzone PJ, Midthun DE, Naidich DP, et al. Evaluation of Individuals With Pulmonary Nodules: When Is it Lung Cancer? Diagnosis and Management of Lung Cancer, 3rd Ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* (2013) 143:e93S–e120S. doi: 10.1378/chest.12-2351
14. LeCun Y, Bengio Y, Hinton G. Deep Learning. *Nature* (2015) 521:436–44. doi: 10.1038/nature14539
15. Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Politte DG, et al. Highly Accurate Model for Prediction of Lung Nodule Malignancy With CT Scans. *Sci Rep* (2018) 8:1–12. doi: 10.1038/s41598-018-27569-w
16. Nibali A, He Z, Wollersheim D. Pulmonary Nodule Classification With Deep Residual Networks. *Int J Comput Assist Radiol Surg* (2017) 12:1799–808. doi: 10.1007/s11548-017-1605-6
17. Ohno Y, Aoyagi K, Yaguchi A, Seki S, Ueno Y, Kishida Y, et al. Differentiation of Benign From Malignant Pulmonary Nodules by Using a Convolutional Neural Network to Determine Volume Change at Chest CT. *Radiology* (2020) 296:432–43. doi: 10.1148/radiol.2020191740
18. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification. *Expert Syst Appl* (2019) 128:84–95. doi: 10.1016/j.eswa.2019.01.048
19. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* (2017) 284:228–43. doi: 10.1148/radiol.2017161659
20. Jiang G, Chen C, Zhu Y, Xie D, Dai J, Jin K, et al. Shanghai Pulmonary Hospital Experts Consensus on the Management of Ground-Glass Nodules Suspected as Lung Adenocarcinoma (Version 1). *Chin J Lung Cancer* (2018) 21(3):147–59.
21. He K, Zhang X, Ren S, Sun J, Research M. Deep Residual Learning for Image Recognition. *IEEE Xplore* (2015), 770–8. doi: 10.1109/CVPR.2016.90
22. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Access* (2009), 248–55. doi: 10.1109/CVPR.2009.5206848
23. Zhang H, Cisse M, Dauphin YN, David LP. Mixup: Beyond Empirical Risk Minimization. *arXiv preprint arXiv* (2017) 171009412:1–13.
24. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv* (2017) 14126980:1–15.
25. Smith LN. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv preprint arXiv* (2018) 180309820:5510–026.
26. Howard J, Gugger S. Fastai: A Layered Api for Deep Learning. *Information* (2020) 11(2):108. doi: 10.3390/info11020108
27. Zhou B, Khosla A, Lapedriza A, Oliva A. Learning Deep Features for Discriminative Localization. *arXiv preprint arXiv* (2015) 1512.04150v1:1–10.
28. Li F, Shusuke S, Abe H, Macmahon H, Doi K. Malignant Versus Benign Nodules at CT Screening for Lung Cancer: Comparison of Thin-Section CT Findings. *Radiology* (2004) 233:793–8. doi: 10.1148/radiol.2333031018
29. Xu DM, van Klaveren RJ, de Bock GH, Leusveld A, Zhao Y, Wang Y, et al. Limited Value of Shape, Margin and CT Density in the Discrimination Between Benign and Malignant Screen Detected Solid Pulmonary Nodules of the NELSON Trial. *Eur J Radiol* (2008) 68:347–52. doi: 10.1016/j.ejrad.2007.08.027
30. Chu ZG, Zhang Y, Li WJ, Li Q, Zheng YN, Lv FJ. Primary Solid Lung Cancerous Nodules With Different Sizes: Computed Tomography Features and Their Variations. *BMC Cancer* (2019) 19:1060–8. doi: 10.1186/s12885-019-6274-0
31. Xu DM, Zaag-Loonen HJVD, Oudkerk M, Wang Y, Vliegenthart R, Scholten ET, et al. Smooth or Attached Solid Indeterminate Nodules Detected at Baseline Ct Screening in the Nelson Study: Cancer Risk During 1 Year of Follow-Up. *Radiology* (2009) 250:264–72. doi: 10.1148/radiol.2493070847
32. Ather S, Kadir T, Gleeson F. Artificial Intelligence and Radiomics in Pulmonary Nodule Management: Current Status and Future Applications. *Clin Radiol* (2020) 75:13–9. doi: 10.1016/j.crad.2019.04.017
33. Khawaja A, Bartholmai BJ, Rajagopalan S, Karwoski RA, Varghese C, Maldonado F, et al. Do We Need to See to Believe? - Radiomics for Lung Nodule Classification and Lung Cancer Risk Stratification. *J Thorac Dis* (2020) 12(6):3303–16. doi: 10.21037/jtd.2020.03.105
34. Ding H, Xia W, Zhang L, Mao Q, Cao B, Zhao Y, et al. CT-Based Deep Learning Model for Invasiveness Classification and Micropapillary Pattern Prediction Within Lung Adenocarcinoma. *Front Oncol* (2020) 10:1186. doi: 10.3389/fonc.2020.01186
35. Qi LL, Wu BT, Tang W, Zhou LN, Huang Y, Zhao SJ, et al. Long-Term Follow-Up of Persistent Pulmonary Pure Ground-Glass Nodules With Deep Learning-Assisted Nodule Segmentation. *Eur Radiol* (2020) 30:744–55. doi: 10.1007/s00330-019-06344-z
36. Chae HD, Park CM, Park SJ, Lee SM, Kim KG, Goo JM. Computerized Texture Analysis of Persistent Part-Solid Ground Glass Nodules: Differentiation of Preinvasive Lesions From Invasive Pulmonary Adenocarcinomas. *Radiology* (2014) 273:285–93. doi: 10.1148/radiol.14132187
37. Weng Q, Zhou L, Wang H, Hui J, Chen M, Pang P, et al. A Radiomics Model for Determining the Invasiveness of Solitary Pulmonary Nodules That Manifest as Part-Solid Nodules. *Clin Radiol* (2019) 74:933–43. doi: 10.1016/j.crad.2019.07.026
38. Gao C, Xiang P, Ye J, Pang P, Wang S, Xu M. Can Texture Features Improve the Differentiation of Infiltrative Lung Adenocarcinoma Appearing as Ground Glass Nodules in Contrast-Enhanced CT? *Eur J Radiol* (2019) 117:126–31. doi: 10.1016/j.ejrad.2019.06.010
39. Xia X, Gong J, Hao W, Yang T, Lin Y, Wang S, et al. Comparison and Fusion of Deep Learning and Radiomics Features of Ground-Glass Nodules to Predict the Invasiveness Risk of Stage-I Lung Adenocarcinomas in CT Scan. *Front Oncol* (2020) 10:418. doi: 10.3389/fonc.2020.00418
40. Shen Y, Xu F, Zhu W, Hu H, Chen T, Li Q, et al. Multiclassifier Fusion Based on Radiomics Features for the Prediction of Benign and Malignant Primary Pulmonary Solid Nodules. *Ann Transl Med* (2020) 8(5):171. doi: 10.21037/atm.2020.01.135

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Oncology

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fonc.2021.793417

Improved Diagnostic Accuracy of Ameloblastoma and Odontogenic Keratocyst on Cone-Beam CT by Artificial Intelligence

Zi-Kang Chai[1†], Liang Mao[1,2†], Hua Chen[3], Ting-Guan Sun[1], Xue-Meng Shen[1], Juan Liu[3] and Zhi-Jun Sun[1,2*]*

[1] The State Key Laboratory Breeding Base of Basic Science of Stomatology (Hubei-MOST) & Key Laboratory of Oral Biomedicine, Ministry of Education, School and Hospital of Stomatology, Wuhan University, Wuhan, China, [2] Department of Oral Maxillofacial-Head Neck Oncology, School and Hospital of Stomatology, Wuhan University, Wuhan, China, [3] Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, China

Objective: The purpose of this study was to utilize a convolutional neural network (CNN) to make preoperative differential diagnoses between ameloblastoma (AME) and odontogenic keratocyst (OKC) on cone-beam CT (CBCT).

Methods: The CBCT images of 178 AMEs and 172 OKCs were retrospectively retrieved from the Hospital of Stomatology, Wuhan University. The datasets were randomly split into a training dataset of 272 cases and a testing dataset of 78 cases. Slices comprising lesions were retained and then cropped to suitable patches for training. The Inception v3 deep learning algorithm was utilized, and its diagnostic performance was compared with that of oral and maxillofacial surgeons.

Results: The sensitivity, specificity, accuracy, and F1 score were 87.2%, 82.1%, 84.6%, and 85.0%, respectively. Furthermore, the average scores of the same indexes for 7 senior oral and maxillofacial surgeons were 60.0%, 71.4%, 65.7%, and 63.6%, respectively, and those of 30 junior oral and maxillofacial surgeons were 63.9%, 53.2%, 58.5%, and 60.7%, respectively.

Conclusion: The deep learning model was able to differentiate these two lesions with better diagnostic accuracy than clinical surgeons. The results indicate that the CNN may provide assistance for clinical diagnosis, especially for inexperienced surgeons.

Keywords: deep learning, convolutional neural network, Inception v3, ameloblastoma, odontogenic keratocyst, cone-beam CT

OPEN ACCESS

Edited by:
*Abhishek Mahajan,
Tata Memorial Hospital, India*

Reviewed by:
*Xiaowei Han,
Nanjing Drum Tower Hospital, China
Mohammed Amjed Alsaegh,
University of Sharjah, United Arab Emirates*

***Correspondence:**
*Zhi-Jun Sun
sunzj@whu.edu.cn
Juan Liu
liujuan@whu.edu.cn*

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

***Received:** 12 October 2021*
***Accepted:** 30 December 2021*
***Published:** 27 January 2022*

Citation:
Chai Z-K, Mao L, Chen H, Sun T-G, Shen X-M, Liu J and Sun Z-J (2022) Improved Diagnostic Accuracy of Ameloblastoma and Odontogenic Keratocyst on Cone-Beam CT by Artificial Intelligence. Front. Oncol. 11:793417. doi: 10.3389/fonc.2021.793417

INTRODUCTION

Ameloblastoma (AME) and odontogenic keratocyst (OKC) are common radiolucent lesions of the jaws in oral and maxillofacial surgery (1, 2). Radiographic examinations are vital for patients with odontogenic lesions, notwithstanding that histopathological findings are the gold-standard diagnostic criteria (3, 4). However, because of the overlap of morphological characteristics in radiography, it is usually difficult to accurately distinguish these two diseases. Current treatment modalities for AME are wide local excision and immediate reconstruction, but OKC is generally treated with more conservative surgical methods, such as marsupialization and/or enucleation. Given that they have different treatment strategies, it is imperative to differentiate these conditions before surgery (5–8).

Clinically, the differentiation between AME and OKC in radiography is mainly based on some features, such as buccolingual expansion, the number of locules, internal density, and the root resorption of the adjacent teeth (**Figure 1**). Nevertheless, only relying on these features is insufficient to obtain a strong differential diagnosis. Previous studies have sought more instrumental radiographic findings, such as the width-to-length ratio, volumetric measurement, and assessment of the Hounsfield unit, to distinguish these two lesions (9–11). However, these studies have the same limitation in that they only focused on low-level and limited features. Therefore, it can be contended that the current knowledge of radiography is still at tip of the iceberg, and more undetected information waits to be mined.

Recently, deep learning, which has been shown to outperform humans in object recognition and visual tasks, has achieved tremendous progress (12, 13). Deep learning algorithms have already been successfully used in medical practice, such as for the detection of incidental esophageal cancers, dermatologist-level classification of skin cancer, prediction of tyrosine kinase inhibitor treatment response, and diagnosis of COVID-19 pneumonia (14–17). In oral and maxillofacial oncology, some researchers have used deep learning methods to distinguish AME and OKC in panoramic radiographs and benefited greatly from the methods (18–20). However, panoramic radiography is not as good as cone-beam CT (CBCT) in demonstrating lesions. As the optimal examination for jaw lesions, CBCT has a high resolution, enabling it to comprehensively and clearly display lesions without distortion, superimposition, and misrepresentation of structures (21, 22). Lee et al. have demonstrated that their deep learning model trained with CBCT images performed better than that trained with panoramic images in diagnosing odontogenic cystic lesions (23). Consequently, we aimed to use a convolutional neural network (CNN) to automatically classify AME and OKC in CBCT data. Furthermore, we compared the diagnostic accuracy of the proposed model with that of senior and junior oral and maxillofacial surgeons.

MATERIALS AND METHODS

Patient and Data Collection

The 350 patients in this study were obtained from the Hospital of Stomatology, Wuhan University, and all of them underwent

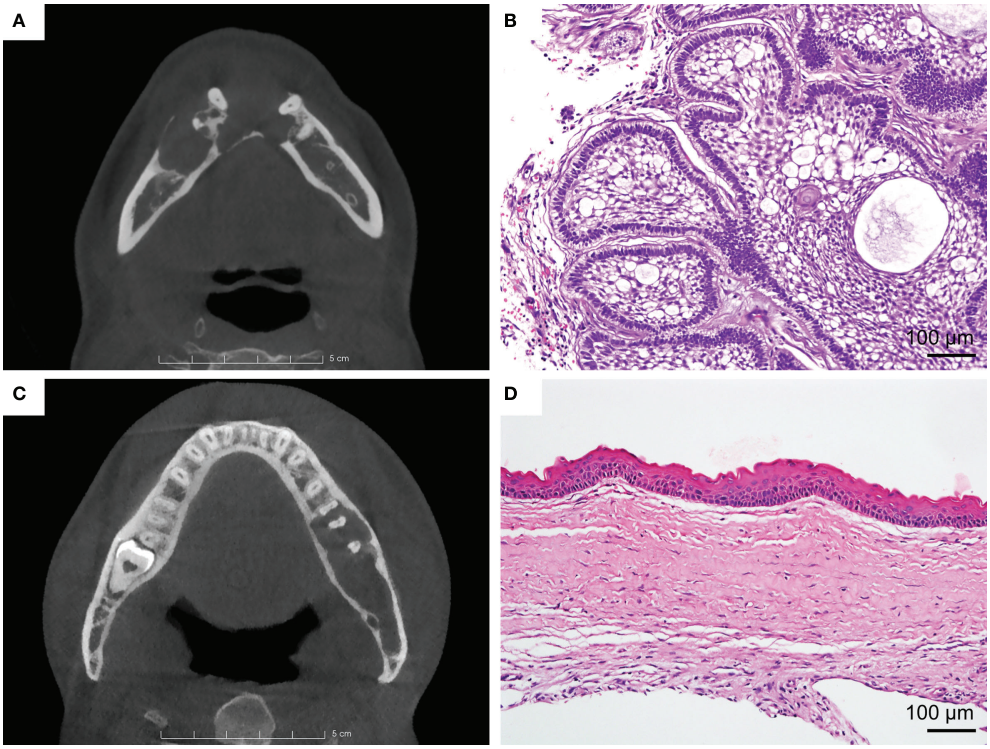


FIGURE 1 | **(A)** Ameloblastoma (AME). Axial view of CBCT shows the lesion with buccal expansion, obvious cortical bone resorption, and a multilocular pattern. **(B)** Typical H&E staining of AME (×200). **(C)** Odontogenic keratocyst (OKC). Axial view of CBCT shows that the lesion grows along the bone, with unapparent disruption of the cortical bone and the unilocular pattern. **(D)** Typical H&E staining of OKC (×200).

surgical treatment with a diagnosis of jaw cystic disease from 2012 to 2020. The pathological diagnosis was made by one pathologist and reviewed by one pathologist from the Department of Oral Pathology, Wuhan University, based on criteria according to the World Health Organization Classification of Head and Neck Tumors (4th, 2017) (24). Their imaging data were retrieved from the picture archiving and communication system (PACS) and saved in DICOM format. All CBCT scans of patients were performed with the same CBCT device (NewTom VG, Italy). The tube voltage was set to 110 kV, and the tube current and the exposure time were regulated by the automatic exposure control system. The images were reconstructed with an isotropic voxel size of 0.3 mm and a 0.3-mm axial pitch.

The inclusion criteria were as follows: 1) complete clinical records, 2) definitive histopathological confirmation of the lesion as AME or OKC, and 3) availability of preoperative CBCT. The exclusion criteria included the following: 1) multiple OKCs or nevoid basal cell carcinoma syndrome and 2) images with apparent artifacts involving the regions of interest (ROIs).

Finally, an equalized dataset consisting of 178 AMEs (130 solid/multicystic ameloblastomas and 48 unicystic ameloblastomas) and 172 OKCs was included in this study. The data were randomly partitioned into two parts: 272 patients in the training set and 78 patients in the testing set, at a ratio of approximately 7:3 (**Table S1**).

Image Processing

The CBCT data were loaded in the open source software 3D Slicer (version 4.11; www.slicer.org) and were demonstrated in three dimensions. The ROI of each slice was manually delineated by a junior surgeon using the semiautomatic segmentation method and then examined and modified by two professional surgeons. The labeled masks were saved in the axial sequence for the subsequent training process. To manifest the lesions more clearly, the open source software mDicom (MicroDicom) was utilized to adjust the raw DICOM images into the bone window (WW/WL, 1,000/300 HU), and then all axial sequences were exported as 512 * 512 pictures in PNG format.

The original pictures were cut into smaller rectangular patches that comprised only the lesions according to labeled masks. The rectangles should be reshaped to squares by padding the black-filled region to fit the CNN architectures and resized to 150 * 150 due to the inconsistent sizes of cropped images. In the training process, in order to reduce redundancy and avoid overfitting of the model, we selected one out of every three consecutive images in a series of each patient. Hence, only one-third of the images of each patient were retained. In the test phase, all images of each patient were tested, and the final classification result was up to the category with larger numbers. If the numbers of the two categories are equal, it means that the model made an incorrect diagnosis of the patient. The experimental procedure is illustrated in **Figure 2**, and some processed pictures of one patient are presented in **Figure 3**. After each case was processed identically, we obtained 272 patients in the training dataset and 78 in the testing dataset, consisting of 11,820 and 11,455 slices, respectively.

Model Interpretation and Training Process

We selected the Inception v3 network as the classifier in our study because it performed better than the other three models (**Table S2**). Inception v3 clustered similar sparse nodes into a dense structure to increase both the depth and width of the network and reduce the computation process efficiently (25). The network consisted of five convolutional layers, two max-pooling layers, 11 inception modules, one average pooling layer, and one

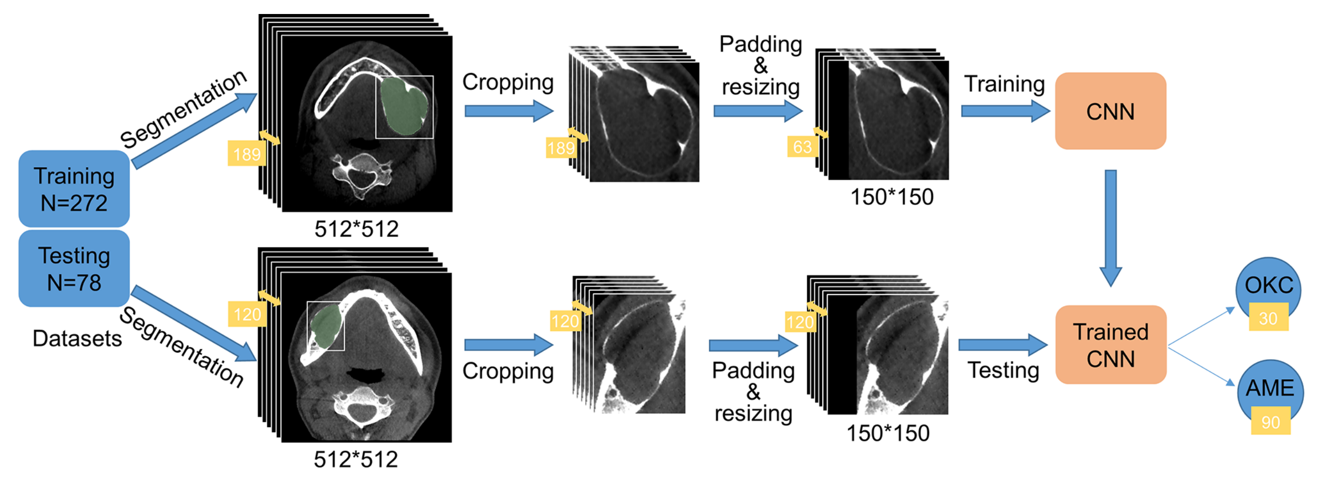


FIGURE 2 | Flow diagram of the study. The training and testing datasets contained 272 and 78 patients, respectively. A total of 189 images comprising the region of interest (ROI) of one patient in the training dataset were cropped into smaller rectangles, padded into square images using black, and resized to 150 * 150 to gear the CNN. To reduce redundancy, we selected one image out of every three images. The series in the testing dataset underwent the same process except for changing the number of images. As shown in the picture, 120 slices of AME patient were tested by the trained CNN. Ninety slices were predicted to have AME and 30 slices were predicted to have OKC, so AME was ultimately considered.

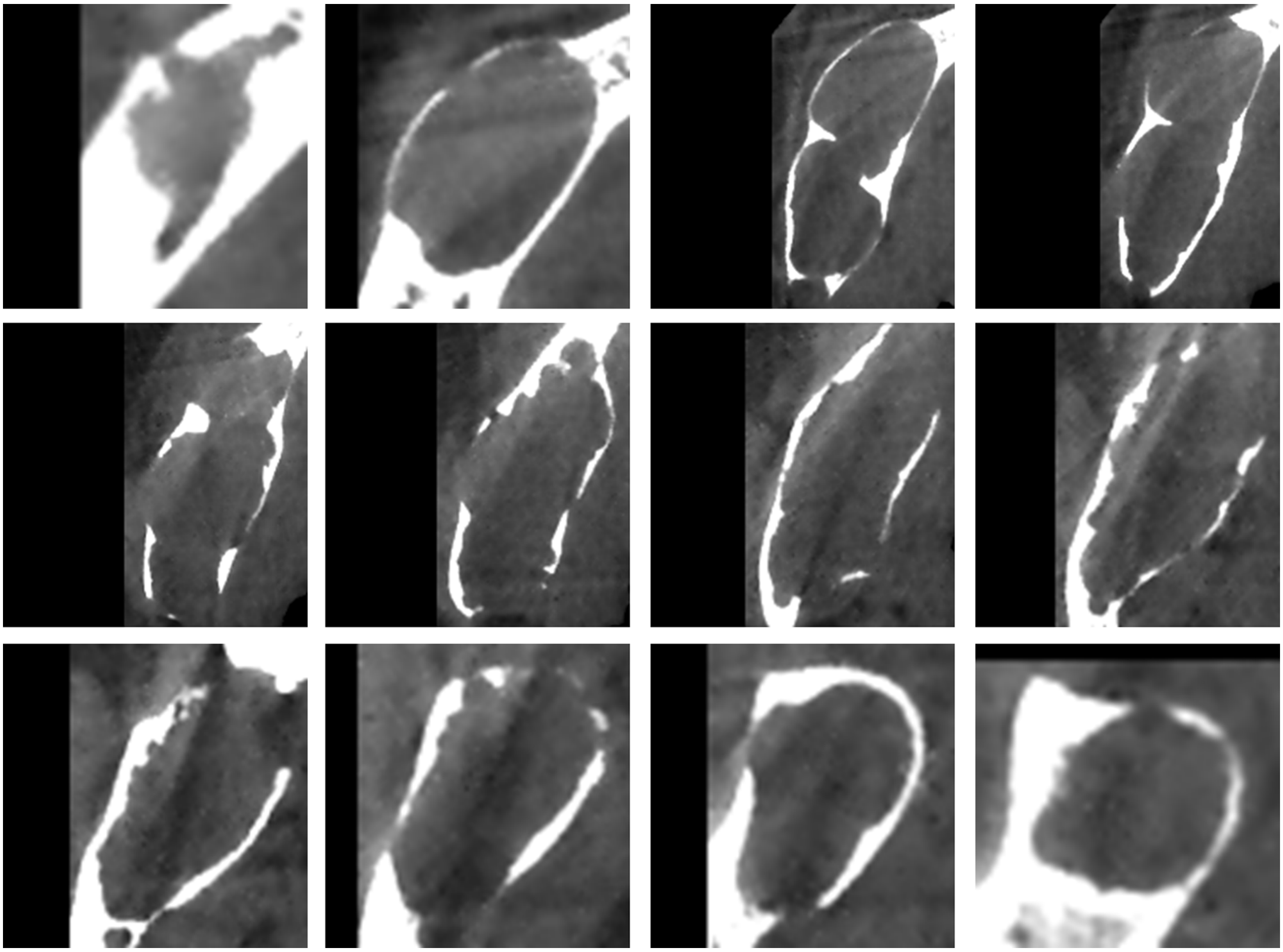


FIGURE 3 | The images used for training were obtained after a series of processing steps, and these 12 images came from one patient.

fully-connected layer (**Figure 4**). The convolution layers were used to extract the features in the jaw images. The pooling layers, including the max-pooling layers and average pooling layer, were utilized to reduce the dimension of features and reduce the amount of calculation. The inception module applied different sized convolution kernels to realize multiscale feature fusion. The fully connected layer integrated the output features of the convolution layer or pooling layer and output the probability

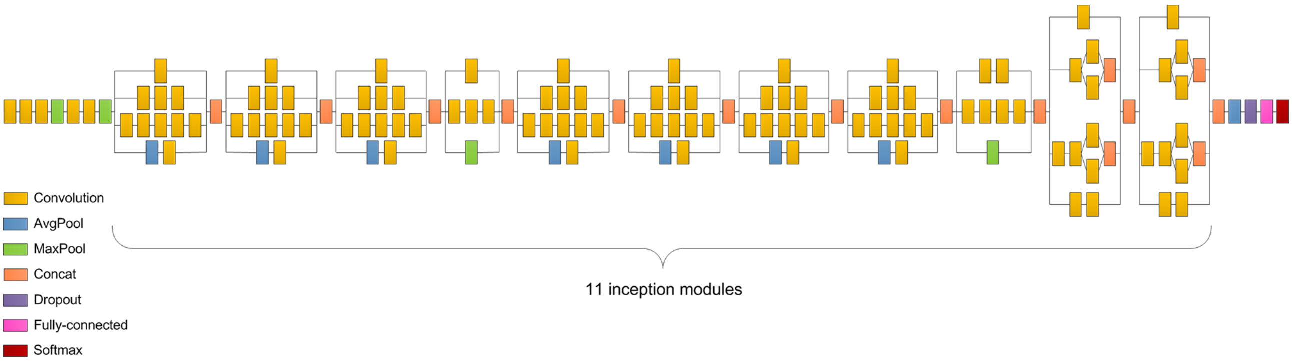


FIGURE 4 | Inception v3 consists of five convolutional layers, two max-pooling layers, 11 inception modules, one average pooling layer, and one fully connected layer.

value of each category after the Softmax activation function. To tackle the problem of limited dataset in medicine, transfer learning was applied in most situations. As done before (26), the CNN was trained on a large ImageNet dataset to learn the hierarchical features. Then, we applied the pretrained CNN with properly adjusted weights in our task.

In this work, our model was performed using a PC with the 64-bit Ubuntu 16.04 operating system, CUDA 9.0, an Intel E5-2650 v4 CPU, 256 GB RAM, a TITAN Xp GPU, and Python 3.5. In the model training process, the datasets were split into training and validation sets at a ratio of 4:1. We utilized the RAdam optimizer to train the layers in batches with a step size of 12 images and a learning rate of 0.0001. After 100 epochs, the training was stopped since both the accuracy and cross-entropy loss were not further improved. The learning history of the model is shown in **Figure 5**.

Testing Surgeons

Clinical surgeons were tested using the identical testing dataset to obtain an objective assessment of the model. Seven senior surgeons and 30 junior surgeons participated in this study, and their results were classified into two groups: senior surgeons and junior surgeons. For each patient, two screenshots comprising three CBCT views, instead of the complete CBCT series, were offered for testing. Only the pictures of patients were summarized into a questionnaire with no more clinical information provided (**Figure S1**).

Statistical Analysis

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ score = \frac{2 * TP}{2 * TP + FP + FN}$$

(TP: true positive, FP: false positive, TN: true negative, FN: false negative)

In this study, the accuracy, specificity, sensitivity (recall), and F1 score were used to assess the performance of Inception v3 and surgeons. For statistical analysis, we regarded the AME as positive and the OKC as negative. The sensitivity was derived by dividing the total number of patients correctly classified as having AME by the total number of AME cases. The specificity was derived by dividing the total number of patients correctly classified as having OKC by the total number of OKC cases. The accuracy was calculated by dividing the number of correctly classified patients by the total number of test patients. The F1 score is the harmonic average of the precision and recall and is considered to comprehensively measure classification performance.

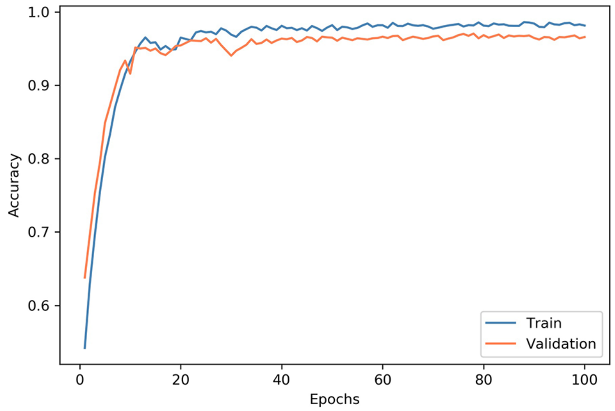


FIGURE 5 | Convergence of the network training. At each epoch, the model was trained using all images in the training dataset, and the accuracy was evaluated. At the end of each epoch, we measured the accuracy of the model on the validation dataset. After 100 epochs, the training was stopped since both accuracy and cross-entropy loss would not be further improved.

RESULTS

Patient Characteristics

The demographic and clinical data of the subjects in this study are presented in **Table 1**. The ages for AME cases range from 9 to 81 years, which is wider than the ages for OKC cases that range from 10 to 70 years. The average age of patients with AME and OKC are 40.3 ± 16.5 years (mean ± standard deviation) and 41.5 ± 17.6 years (mean ± standard deviation), respectively. Both AME and OKC have a predilection for the mandible.

Comparison Results Between Model and Surgeons

Inception v3 obtained the highest scores among the participants, with a sensitivity of 87.4%, a specificity of 82.1%, an accuracy of 84.6%, and an F1 score of 85.0% (**Table 2**). For Inception v3, the diagnostic accuracy of AME (87.4%) was slightly higher than that of OKC (82.1%). Compared with lesions in the maxilla, the model had better diagnostic performance for the mandible, and the accuracies are shown in **Table 3**. The average prediction time for an image was 3.13 ms using the model, and the total time for diagnosing the 78 patients was 35.87 s.

TABLE 1 | Demographic data of the study subjects.

Characteristics	OKC (*N* = 172)	AME (*N* = 178)
Age (mean ± SD)	41.5 ± 17.6	40.3 ± 16.5
Location		
Maxilla	63 (36.6%)	24 (13.5%)
Mandible	109 (63.4%)	154 (86.5%)
Gender		
Male	91 (52.9%)	108 (60.7%)
Female	81 (47.1%)	70 (39.3%)

SD, standard deviation; OKC, odontogenic keratocyst; AME, ameloblastoma.

TABLE 2 | Comparison results of Inception v3 and surgeons.

	Sensitive (%)	Specificity (%)	Accuracy (%)	F1 score (%)
Inception v3	87.2	82.1	84.6	85.0
Senior surgeons	60.0	71.4	65.7	63.6
Junior surgeons	63.9	53.2	58.5	60.7

TABLE 3 | Diagnostic accuracy in the maxilla and mandible.

	Testing number	Sensitive (%)	Specificity (%)	Accuracy (%)	F1 score (%)
Maxilla	15	50.0	81.8	73.3	50.0
Mandible	63	91.4	82.1	87.3	88.9

The average sensitivity, specificity, accuracy, and F1 score for the classification of the group of 7 senior surgeons were 60.0%, 71.4%, 65.7%, and 63.6%, respectively, and those of the group of 30 junior surgeons were 63.9%, 53.2%, 58.5%, and 60.7%, respectively. The diagnostic outcomes of the CNN model and 5 surgeons were presented by confusion matrices (**Figure 6**). The average time to make diagnoses for 78 patients by 7 senior surgeons was 1,471 s. For the 30 junior surgeons, the average time was 1,113 s.

DISCUSSION

AME is the most common benign odontogenic tumor, accounting for approximately 10% of all odontogenic tumors (27). AME can arise from any odontogenic epithelium, so it can manifest widely varied radiographic findings. As the third most common odontogenic cyst, OKC represents nearly 12% of all odontogenic cysts, also arising from odontogenic epithelium (28). According to the literature reports, OKC is inclined to grow along the bone without the same buccolingual expansion of AME that usually results in bone resorption. However, these results could also be observed when OKC reached a large size. These confusing radiographic manifestations contributed to the difficulty of differential diagnosis. For AME, the main treatment modality is wide local excision and immediate reconstruction (6). Nevertheless, OKC is generally treated with more conservative surgical methods, such as marsupialization and/or enucleation, followed by adjunctive treatments, including cryotherapy with liquid nitrogen or the

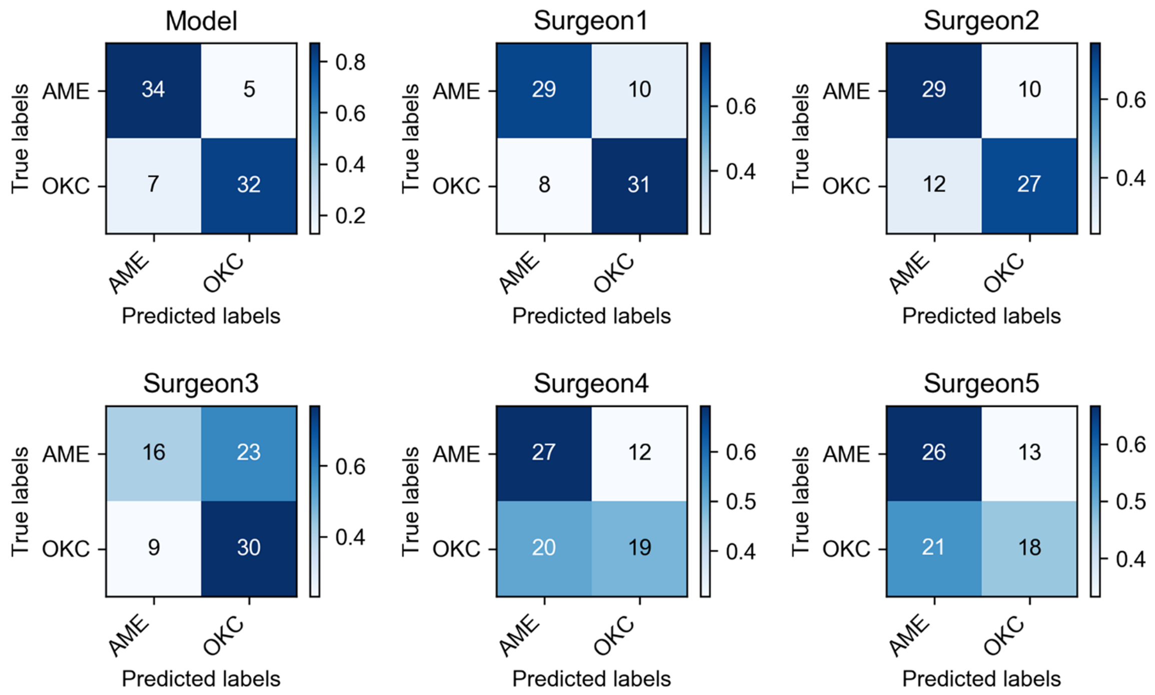


FIGURE 6 | Confusion matrices of Inception v3 and five oral and maxillofacial surgeons showed the specific diagnostic performance. The color shade of the grid represented the proportion of each class.

application of fixative Carnoy's solution, to reduce recurrence (5). As a consequence, precise preoperative diagnosis is necessary for determining appropriate treatment strategies.

This study can be regarded as a successful application of deep learning in the field of odontogenic diseases with CBCT data. The results showed that our CNN model exhibited superior performance in differentiating AME and OKC compared with the oral and maxillofacial surgeons. Its diagnostic capability considerably outperformed senior and junior surgeons. Notably, though the sensitivity of junior surgeons (63.9%) was higher than that of the senior surgeons (60.0%), it did not mean that the junior surgeons had better diagnostic capabilities. This was because the junior surgeons in this study were inclined to choose the AME. As shown in the results, the specificity of junior surgeons (53.2%) was significantly lower than that of senior surgeons (71.4%). Furthermore, the CNN model spent extremely less time in diagnosis than the senior and junior surgeons. The average diagnosis time for the group of senior surgeons was longer than that for the group of junior surgeons. A possible explanation for this might be that senior surgeons would consider more details when they made a diagnosis. There are also some studies that developed deep learning models for differentiating AME and OKC in panoramic radiographs and achieved a high classification accuracy for lesions of the mandible (18–20). However, these models cannot perform well for lesions of the maxilla due to the inherent limitations of the panoramic radiograph, including the distortion, superimposition, and misrepresentation of structures. In contrast, CBCT has a higher resolution, enabling it to comprehensively and clearly display lesions in the maxillofacial region, which has many complex anatomic structures (21, 22). As a result, in our work, it was not necessary to deliberately select the location of onset. Our CNN model could substantially distinguish OKC and AME regardless of whether the lesion was in the maxilla or mandible. Bispo et al. used deep learning methods to differentiate them in multidetector CT images. However, their work was based only on extremely limited data from 40 patients, which would weaken the credibility of their results (29). In contrast, a larger dataset consisting of 350 patients was used in our study. Consequently, the convincing results indicated that our model could provide assistance for clinical diagnosis, especially for inexperienced surgeons.

In our study, we found that the diagnostic accuracy in the maxilla was lower than that in the mandible, and the possible explanations might be as follows. First, the low incidence in the maxilla results in less available data. Second, there may be more similar manifestations in the maxilla. There are few bone absorptions when lesions are small because of the intrinsic sinus cavities in the maxilla. However, the flimsy maxillary cortex is more susceptible to extensive destruction which often involves the nasal cavity and ethmoidal and sphenoidal sinuses, by both AME and OKC (30).

In the present study, two special and effective methods were used to improve the performance of the model. First, we cropped the original images using a tailored processing method. The original images contained many irrelevant anatomical structures, such as teeth, craniofacial bones, and muscles, and such loud noise might interfere with the model accurately extracting the features from the ROIs. Shin et al. proved that slice-level classification is more challenging than patch-level classification (31). Monkam and his colleagues compared the performances of several models based on different sized patches (32). Given that the sizes of ROIs in our database covered a large variation, it was irrational to establish a one-size-fits-all patch size. We tailored the optimal patch size for each slice by automatically measuring the mask to determine a suitable width and length of the rectangle. This process proved to be conducive to reducing the memory footprint and increasing the accuracy. Second, we noticed that the adjacent slices in the CBCT scans of one patient were extremely similar, which could lead to redundancy. As a solution, we selected one image out of every three images. This processing not only improved the training speed in every epoch but also effectively avoided overfitting and improved the model performance.

Keep in mind that our study still has some limitations. First, the diagnostic accuracy of the surgeons might be underestimated. Neither the model nor the surgeons were allowed to utilize the clinical information of patients, which is indispensable in clinical practice. In addition, we tested surgeons using only partial images of the CBCT series. Second, we did not perform external data validation; therefore, the generalizability of the model should be considered. The difficulty of obtaining sufficient images restricts the application of deep learning in the field of medical research. It is no exception that we used a relatively small amount of data, and all data were from the same medical center. Third, the CNN model was only based on 2D ROI patches of axial images, which might result in ignoring contextual information. Apparently, we suboptimally used the CBCT data, which are amenable to providing 3D manifestations. Ciompi et al. effectively classified pulmonary perifissural nodules by combining several 2D views (33), and Xu et al. designed a 3D CNN for automatic bladder segmentation to fully exploit 3D CT images (34). These studies of predecessors are bound to guide subsequent works, which are worthy of undertaking in the future. For example, we can attempt to multistream architectures based on three dimensions of CBCT or utilize a 3D-CNN to improve the classification accuracy. We can also search for the most suitable window setting to fully manifest lesions and pay more attention to overcoming the conundrum in differentiating lesions in the maxilla. Furthermore, external validations are indispensable to strengthen the generalization and credibility of the model. Additionally, we expect that deep learning will make greater advances and yield greater benefits for medical systems.

In conclusion, the CNN model achieved a fulfilling accuracy in diagnosing AME and OKC through CBCT, and the model significantly outperformed senior and junior surgeons of oral and maxillofacial. While these results require further validation, our work suggests that the CNN model can provide substantial assistance with non-invasive diagnosis and therapy guidance for patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because we want to protect the interest of the patients. Requests to access the datasets should be directed to sunzj@whu.edu.cn.

ETHICS STATEMENT

This study followed the Declaration of Helsinki guidelines and its protocol was approved by the Institutional Medical Ethics Committee of School and Hospital of Stomatology, Wuhan University (2018LUNSHENZIA28). Written informed consent to participate in this study was provided by the legal guardian/next of kin of the participants. Written informed consent was obtained from the legal guardian/next of kin of the individual(s) and minor(s), for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Z-JS and JL: study conception and design. Z-KC, T-GS, and X-MS: data collection. HC and Z-KC: data analysis and interpretation. Z-KC and LM: manuscript writing. Z-JS and JL: manuscript revision. All authors: manuscript review and final approval of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (82072996 and 82103333), the Fundamental Research Funds for the Central Universities (2042021kf0216), the Major Projects of Technological Innovation in Hubei Province (2019AEA170), and the Frontier Projects of Wuhan for Application Foundation (2019010701011381).

ACKNOWLEDGMENTS

The authors appreciate Zhi-Yuan Xiong and Hai-Yan Qi from the information center and all the staff from the Department of Oral Radiology and Department of Oral Pathology in the Hospital of Stomatology, Wuhan University for the assistance of collecting data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.793417/full#supplementary-material

REFERENCES

1. Wright JM, Vered M. Update From the 4th Edition of the World Health Organization Classification of Head and Neck Tumours: Odontogenic and Maxillofacial Bone Tumors. *Head Neck Pathol* (2017) 11:68–77. doi: 10.1007/s12105-017-0794-1
2. Luo HY, Li TJ. Odontogenic Tumors: A Study of 1309 Cases in a Chinese Population. *Oral Oncol* (2009) 45:706–11. doi: 10.1016/j.oraloncology.2008.11.001
3. Theodorou DJ, Theodorou SJ, Sartoris DJ. Primary non-Odontogenic Tumors of the Jawbones: An Overview of Essential Radiographic Findings. *Clin Imaging* (2003) 27:59–70. doi: 10.1016/s0899-7071(02)00518-1
4. Mendes RA, Carvalho JF, van der Waal I. Characterization and Management of the Keratocystic Odontogenic Tumor in Relation to its Histopathological and Biological Features. *Oral Oncol* (2010) 46:219–25. doi: 10.1016/j.oraloncology.2010.01.012
5. Sharif FN, Oliver R, Sweet C, Sharif MO. Interventions for the Treatment of Keratocystic Odontogenic Tumours. *Cochrane Database Syst Rev* (2015) 2015:Cd008464. doi: 10.1002/14651858.CD008464.pub3
6. McClary AC, West RB, McClary AC, Pollack JR, Fischbein NJ, Holsinger CF, et al. Ameloblastoma: A Clinical Review and Trends in Management. *Eur Arch Otorhinolaryngol* (2016) 273:1649–61. doi: 10.1007/s00405-015-3631-8
7. Vallejo-Rosero KA, Camolesi GV, de Sa PLD, Bernaola-Paredes WE. Conservative Management of Odontogenic Keratocyst With Long-Term 5-Year Follow-Up: Case Report and Literature Review. *Int J Surg Case Rep* (2020) 66:8–15. doi: 10.1016/j.ijscr.2019.11.023
8. Effiom OA, Ogundana OM, Akinshipo AO, Akintoye SO. Ameloblastoma: Current Etiopathological Concepts and Management. *Oral Dis* (2018) 24:307–16. doi: 10.1111/odi.12646
9. Omami G, Adel M. Width-To-Length Ratio Comparison Between Ameloblastomas and Odontogenic Keratocysts in the Body of the Mandible: A Preliminary Study. *Imaging Sci Dent* (2020) 50:319–22. doi: 10.5624/isd.2020.50.4.319
10. Safi AF, Kauke M, Timmer M, Grandoch A, Nickenig HJ, Gültekin E, et al. Does Volumetric Measurement Serve as an Imaging Biomarker for Tumor Aggressiveness of Ameloblastomas? *Oral Oncol* (2018) 78:16–24. doi: 10.1016/j.oraloncology.2018.01.002
11. Uehara K, Hisatomi M, Munhoz L, Kawazu T, Yanagi Y, Okada S, et al. Assessment of Hounsfield Unit in the Differential Diagnosis of Odontogenic Cysts. *Dentomaxillofac Radiol* (2021) 50:20200188. doi: 10.1259/dmfr.20200188
12. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vision* (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y
13. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the Game of Go With Deep Neural Networks and Tree Search. *Nature* (2016) 529:484–9. doi: 10.1038/nature16961
14. Sui H, Ma R, Liu L, Gao Y, Zhang W, Mo Z. Detection of Incidental Esophageal Cancers on Chest CT by Deep Learning. *Front Oncol* (2021) 11:700210. doi: 10.3389/fonc.2021.700210
15. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks. *Nature* (2017) 542:115–8. doi: 10.1038/nature21056
16. Hou R, Li X, Xiong J, Shen T, Yu W, Schwartz LH, et al. Predicting Tyrosine Kinase Inhibitor Treatment Response in Stage IV Lung Adenocarcinoma Patients With EGFR Mutation Using Model-Based Deep Transfer Learning. *Front Oncol* (2021) 11:679764. doi: 10.3389/fonc.2021.679764
17. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell* (2020) 181:1423–33.e11. doi: 10.1016/j.cell.2020.04.045
18. Liu Z, Liu J, Zhou Z, Zhang Q, Wu H, Zhai G, et al. Differential Diagnosis of Ameloblastoma and Odontogenic Keratocyst by Machine Learning of Panoramic Radiographs. *Int J Comput Assist Radiol Surg* (2021) 16:415–22. doi: 10.1007/s11548-021-02309-0
19. Poedjiastoeti W, Suebnukarn S. Application of Convolutional Neural Network in the Diagnosis of Jaw Tumors. *Healthc Inform Res* (2018) 24:236–41. doi: 10.4258/hir.2018.24.3.236
20. Yang H, Jo E, Kim HJ, Cha IH, Jung YS, Nam W, et al. Deep Learning for Automated Detection of Cyst and Tumors of the Jaw in Panoramic Radiographs. *J Clin Med* (2020) 9:1839. doi: 10.3390/jcm9061839
21. Dawood A, Patel S, Brown J. Cone Beam CT in Dental Practice. *Br Dent J* (2009) 207:23–8. doi: 10.1038/sj.bdj.2009.560
22. Scarfe WC, Farman AG. What is Cone-Beam CT and How Does it Work? *Dent Clin North Am* (2008) 52:707–30. doi: 10.1016/j.cden.2008.05.005

23. Lee JH, Kim DH, Jeong SN. Diagnosis of Cystic Lesions Using Panoramic and Cone Beam Computed Tomographic Images Based on Deep Learning Neural Network. *Oral Dis* (2020) 26:152–8. doi: 10.1111/odi.13223
24. El-Naggar AK, Chan JKC, Grandis JR, Takata T, Slootweg PJ. WHO Classification of Head and Neck Tumours. *Lyon: Int Agency Res Cancer* (2017). p. 203.
25. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. (2016). Rethinking the Inception Architecture for Computer Vision [Conference presentation]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, United States* (2016). Available at: https://arxiv.org/abs/1512.00567.
26. Han T, Liu C, Yang W, Jiang D. Learning Transferable Features in Deep Convolutional Neural Networks for Diagnosing Unseen Machine Conditions. *ISA Trans* (2019) 93:341–53. doi: 10.1016/j.isatra.2019.03.017
27. Petrovic ID, Migliacci J, Ganly I, Patel S, Xu B, Ghossein R, et al. Ameloblastomas of the Mandible and Maxilla. *Ear Nose Throat J* (2018) 97: E26–e32. doi: 10.1177/014556131809700704
28. Johnson NR, Gannon OM, Savage NW, Batstone MD. Frequency of Odontogenic Cysts and Tumors: A Systematic Review. *J Investig Clin Dent* (2014) 5:9–14. doi: 10.1111/jicd.12044
29. Bispo MS, Pierre Junior M, Apolinario AL Jr, Dos Santos JN, Junior BC, Neves FS, et al. Computer Tomographic Differential Diagnosis of Ameloblastoma and Odontogenic Keratocyst: Classification Using a Convolutional Neural Network. *Dentomaxillofac Radiol* (2021) 50:20210002. doi: 10.1259/dmfr.20210002
30. Zwahlen RA, Gratz KW. Maxillary Ameloblastomas: A Review of the Literature and of a 15-Year Database. *J Craniomaxillofac Surg* (2002) 30:273–9. doi: 10.1016/s1010-5182(02)90317-3
31. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging* (2016) 35:1285–98. doi: 10.1109/tmi.2016.2528162
32. Monkam P, Qi S, Xu M, Han F, Zhao X, Qian W. CNN Models Discriminating Between Pulmonary Micro-Nodules and Non-Nodules From CT Images. *BioMed Eng Online* (2018) 17:96. doi: 10.1186/s12938-018-0529-x
33. Ciompi F, de Hoop B, van Riel SJ, Chung K, Scholten ET, Oudkerk M, et al. Automatic Classification of Pulmonary Peri-Fissural Nodules in Computed Tomography Using an Ensemble of 2D Views and a Convolutional Neural Network Out-of-the-Box. *Med Image Anal* (2015) 26:195–202. doi: 10.1016/j.media.2015.08.001
34. Xu X, Zhou F, Liu B. Automatic Bladder Segmentation From CT Images Using Deep CNN and 3D Fully Connected CRF-RNN. *Int J Comput Assist Radiol Surg* (2018) 13:967–75. doi: 10.1007/s11548-018-1733-7

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Optimization of Cervical Cancer Screening: A Stacking-Integrated Machine Learning Algorithm Based on Demographic, Behavioral, and Clinical Factors

Lin Sun[1], Lingping Yang[1], Xiyao Liu[2], Lan Tang[3], Qi Zeng[1], Yuwen Gao[1], Qian Chen[1], Zhaohai Liu[4] and Bin Peng[1*]*

[1] School of Public Health and Management, Chongqing Medical University, Chongqing, China, [2] Department of Obstetrics, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [3] Department of Physical Examation, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [4] Information Section, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

OPEN ACCESS

Edited by:
Fu Wang,
Xi'an Jiaotong University, China

Reviewed by:
Suwei Chen,
Capital Medical University, China
Chong You,
Peking University, China

****Correspondence:***
Bin Peng
pengbin@cqmu.edu.cn
Zhaohai Liu
201447@hospital.cqmu.edu.cn

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *24 November 2021*
Accepted: *19 January 2022*
Published: *15 February 2022*

Citation:
Sun L, Yang L, Liu X, Tang L, Zeng Q, Gao Y, Chen Q, Liu Z and Peng B (2022) Optimization of Cervical Cancer Screening: A Stacking-Integrated Machine Learning Algorithm Based on Demographic, Behavioral, and Clinical Factors. Front. Oncol. 12:821453. doi: 10.3389/fonc.2022.821453

Purpose: The purpose is to accurately identify women at high risk of developing cervical cancer so as to optimize cervical screening strategies and make better use of medical resources. However, the predictive models currently in use require clinical physiological and biochemical indicators, resulting in a smaller scope of application. Stacking-integrated machine learning (SIML) is an advanced machine learning technique that combined multiple learning algorithms to improve predictive performance. This study aimed to develop a stacking-integrated model that can be used to identify women at high risk of developing cervical cancer based on their demographic, behavioral, and historical clinical factors.

Methods: The data of 858 women screened for cervical cancer at a Venezuelan Hospital were used to develop the SIML algorithm. The screening data were randomly split into training data (80%) that were used to develop the algorithm and testing data (20%) that were used to validate the accuracy of the algorithms. The random forest (RF) model and univariate logistic regression were used to identify predictive features for developing cervical cancer. Twelve well-known ML algorithms were selected, and their performances in predicting cervical cancer were compared. A correlation coefficient matrix was used to cluster the models based on their performance. The SIML was then developed using the best-performing techniques. The sensitivity, specificity, and area under the curve (AUC) of all models were calculated.

Results: The RF model identified 18 features predictive of developing cervical cancer. The use of hormonal contraceptives was considered as the most important risk factor, followed by the number of pregnancies, years of smoking, and the number of sexual partners. The SIML algorithm had the best overall performance when compared with other methods and reached an AUC, sensitivity, and specificity of 0.877, 81.8%, and 81.9%, respectively.

Conclusion: This study shows that SIML can be used to accurately identify women at high risk of developing cervical cancer. This model could be used to personalize the screening program by optimizing the screening interval and care plan in high- and low-risk patients based on their demographics, behavioral patterns, and clinical data.

Keywords: machine learning, cervical cancer, risk, artificial intelligence, personalized screening

INTRODUCTION

Cervical cancer is one of the most common malignant tumors in women worldwide (1). The 5-year survival rate for early-stage cervical cancer is high, ranging from 80% to 90% (2). However, the cure rate goes down to 10% for stage 4 disease (3). Cervical screening has, therefore, an important role in identifying the disease at an early stage and hence reduces the morbidity and mortality from the disease. The incidence and mortality from cervical cancer vary across different countries and tend to be lower in highly developed countries due to well-established screening and vaccination programs (4). However, underdeveloped regions often do not have sufficient medical resources allocated to screening. This implies that there is an increased need to identify women at a high risk of developing cervical cancer to optimize the screening interval and hence make better use of medical resources (5, 6).

Parametric prediction models can be used to better identify the early risk warning signs of cervical cancer (7–9). However, to our knowledge, there is currently no comprehensive risk prediction model based on demographic information, behavioral habits, and medical history for cervical cancer. Prediction models need to be able to make use of individual information to accurately predict the risk of developing the disease. Artificial intelligence (AI) and machine learning (ML) can be used to analyze large volumes of data to make accurate predictions and to identify hidden interactions (10, 11). Therefore, the use of AI and ML in the medical field has increased exponentially during the past few years. However, current risk prediction models for cervical cancer are based on former-generation algorithms, such as the decision tree model and random forest (RF) (12). Until recently, more powerful algorithms such as the stacking-integration machine learning (SIML) have yet to be fully explored. SIML's automatic large-scale integration strategy can effectively combat overfitting by adding regular items and transferring the integrated knowledge to a simple classifier, which is the best way to improve the effectiveness of machine learning.

This study aimed to develop an SIML that could be used to identify women at a high risk of developing cervical cancer based on their demographic, behavioral, and medical history and hence personalize the screening program according to their risk factors.

MATERIALS AND METHODS

Study Populations

These data were obtained from the public dataset provided by Kelvin Fernandes in the UCI database. The data were based on early screening data for cervical cancer collected at the *Hospital Universitario de Caracas*, Venezuela, from March 2012 to September 2013 (13). The majority of patients were of low socioeconomic status, low income, and low educational level. The patients were aged 13–84 years, with an average age of 27 years, and 88.6% of them had at least one pregnancy. The data collected included demographics, behavioral patterns, and medical histories of 858 patients. A total of 18 different potential risk variables were identified and coded, as shown in **Supplementary Table S1**. Due to missing variables for privacy concerns, not all patient variables were available for analysis. Feature datasets excluded variables with more than half loss rate or those that have all identical values. The original general data parameter index code is available in **Supplementary Table S1**, and the main content of the modeling is shown in **Figure 1**.

Dataset Preprocessing

The premise of an efficient and reliable disease risk prediction model was the accuracy of the data. Visualization of the data was first performed using the public packages related to ML in R, version 3.6.0 (The R Foundation for Statistical Computing, Vienna, Austria), while the PRISM software version 7.0 for Windows (GraphPad Software Inc., San Diego, CA, USA) was used to plot the data (**Supplementary Figure S1**).

Following visualization of the data, 18 high-risk prediction features (**Supplementary Table S1**) of a positive biopsy were identified. Continuous variables were categorized as follows. The ages of the patients were grouped into four categories: below 20 years, 20–29 years, 30–44 years, and 45–60 years, while the age of first sexual intercourse was grouped into five groups: below 13 years, 13–15 years, 16–17 years, 18–19 years, and above 20 years. Other classification variables were input according to the original characteristics.

Not all the data for each predictive feature were available. About 20%–30% of the clinical predictive data and about 0%–15% of the behavioral data were missing. The missing part of the data had to be estimated by using the information available in the existing data to replace the missing data with values (14). However, due to a large number of missing data, conventional mean and median filling methods could not be used in this case, since these techniques cannot guarantee data authenticity because the filling values are mostly unreal values, which will affect the accuracy of model construction. Therefore, nonparametric missing value imputation using RF (MissForest) was used to process missing data as suggested by Stekhoven et al.

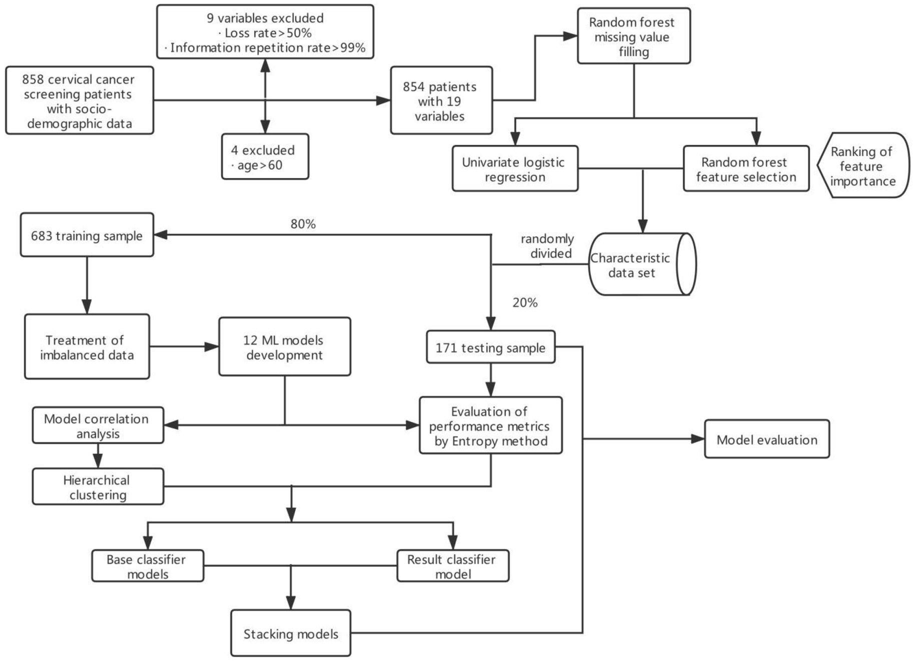

9 variables excluded
· Loss rate>50%
· Information repetition rate>99%
858 cervical cancer screening patients with socio-demographic data
854 patients with 19 variables
Random forest missing value filling
4 excluded
· age>60
Univariate logistic regression
Random forest feature selection
Ranking of feature importance
683 training sample
80%
randomly divided
Characteristic data set
20%
Treatment of imbalanced data
12 ML models development
171 testing sample
Model correlation analysis
Evaluation of performance metrics by Entropy method
Model evaluation
Hierarchical clustering
Base classifier models
Result classifier model
Stacking models

FIGURE 1 | Flowchart illustrating the development and validation of ML models. ML, machine learning.

(15). The parameters of the model were set as follows: the maximum iterations were set to 10. The number of trees was chosen to be 100.

Feature Selection

The model was designed to rely on a limited and effective set of features that do not require excessive input from patients. Using the RF model, a total of 18 predictors for developing cervical cancer were identified. The univariate logistic regression and feature selection model were then used to quantify the odds ratio (OR) and the contributing risk of each predictive value for developing cervical cancer. The analysis of feature selection was based on the RF classifier, whereby the importance of each predictive feature was sorted by using the error rate measurement. Specifically, for each tree in the RF, the error rate for classification of the out-of-bag portion of the data was recorded. The feature importance score was calculated by estimating improvement in the classification error rate of each feature. Finally, the importance scores of all trees in RF were averaged to get the final score of each feature (16). Nine important predictive features were finally identified.

Treatment of Imbalanced Data

Imbalanced data refer to the uneven distribution of data among different categories, whereby the main categories have a much larger representation (17). The imbalance ratio (IR) is expressed as the ratio of the number of large sample categories to the number of small sample categories. A large IR generally has a negative impact on the classification effect of the model and can lead to an inaccurate classification.

Two techniques were used to deal with imbalanced data in our study. The first method involved the use of resampling based on samples (oversampling, undersampling, and hybrids). The other method combines the use of resampling methods *via* the random oversampling example (ROSE) (18) and synthetic minority oversampling technique (SMOTE) (19) algorithms. In this study, five different resampling methods and RF were combined to build the models, and ultimately the best method was selected and integrated into the final SIML.

Model Development

Following class imbalance treatment, the cervical cancer screening data were randomly assigned to the training dataset (80% of data) and testing dataset (20% of the data). The training dataset was used to develop the algorithm, while the testing dataset was used to evaluate the performance of the algorithm. We then selected 12 widely used ML algorithms including RF, Stochastic Gradient Boosting (SGB), Bagged Classification and Regression Tree (TreeBag), eXtreme Gradient Boosting (XGBoost), Monotone Multi-Layer Perceptron Neural Network (MonMLP), Support Vector Machines with Radial Basis Function Kernel (SVMRadial), K-Nearest Neighbors (KNN), Gaussian Process with Radial Basis Function Kernel (GaussPrRadial), Regularized Logistic Regression (RgeLogistic), Stabilized Linear Discriminant (SLDA), AdaBoost Classification Trees (AdaBoost), and Logistic Model Trees (LMT). All of these supervised algorithms were implemented using the free and open-source library caret in R3.6.0. To adjust the optimal tuning parameters of each ML algorithm, we used 10-fold cross-validation and repeated three times on the training set.

This method involved dividing the training set into 10 sets and using nine sets for training and the remaining set was used for verification. This was performed 10 times, and the results of the different test sets were averaged, ensuring an independent result from the actual dataset subdivision (20).

RF, TreeBag, SGB, AdaBoost, and XGBoost are integrated algorithms that combine multiple simple tree models (21, 22) and are considered to be the most accurate for making predictions using various datasets for several applications. MonMLP is a feed-forward Artificial Neural Network (ANN) model, which maps multiple input datasets to a single output. As a popular ML algorithm, MonMLP has incomparable advantages in prediction accuracy. However, it requires tuning of many parameters and a large number of data for training (23). SVMRadial is an SVM model with Radial Basis Function, which constructs a decision curve in high-dimensional feature space to perform binary classification (24). KNN, GaussPrRadial, RegLogistic, and SLDA are relatively efficient and effective simple classification algorithms in data mining. Although these algorithms are relatively simple, they still perform very well and result in a model that is easier to interpret (21, 25). LMT is an algorithm generated by the combination of linear logistic regression and decision tree induction. It has been proven to be an accurate and simple classifier, which is also competitive with other advanced classifiers (such as RF) and easier to explain (26).

The performances of the algorithms were compared to select the optimal stacking algorithm. Stacking is a common integrated learning framework in the Kaggle competition, integrating many models to improve the result prediction accuracy. It is generally used to train a two-layer learning structure. The first layer (known as the learning layer) trains *n* different classifiers, and their predicted results are combined into a new feature set, which is then used as the input of the next layer classifier (27) (**Figure 2**). Stacking has the characteristics of distributing multiple classifiers while ensuring excellent performance. In summary, the stacking-integrated learning framework has two requirements for base classifiers: large differences between classifiers and high accuracy of classifiers. However, it is prone to overfitting (28). The features of the second layer come from learning the results of the first one. Thus, the original features should not be included in the features of the data of the second layer to reduce the risk of overfitting. The best choice of the second layer classifier is a relatively simple classifier. RegLogistic is a better method in Stacking (29), but LMT is more robust in overfitting (26), and can therefore be used instead.

Model Comparisons

The optimal tuning parameters of each ML algorithm were determined by cross-validation on the training samples after imbalance data processing. The models' internal verification scores were obtained from the training dataset, while the external validation scores were obtained from the test sets. External validation scores could be used to test the generalization power of the model. The performance evaluation of binary data (positive vs. negative) was mainly based on the sensitivity ($\frac{TP}{TP+FN}$) and specificity ($\frac{TN}{TN+FP}$), where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively. The area under the curve (AUC) was used to reflect the relationship between two performance variables. F1 scores and F2 scores were also used to measure the model's accuracy.

$F1 = \frac{2 * precision * recall}{precision + recall}$, $F2 = \frac{(1+2^2) * precision * recall}{2^2 * precision + recall}$, *in which* $precision = \frac{TP}{TP+FP}$ *and* $recall = \frac{TP}{TP+FN}$

Alternatively, the F1 score and F2 score were a kind of harmonic mean of model accuracy and recall (30), comparing different model performances in identifying true disease predictions when compared to false positives. The weight of the F2 score was more inclined to the recall value of the model and focuses on the sensitivity index of the model.

The entropy weight method was an objective weighting method that can be used to reduce the influence of human factors. After averaging the seven performance metrics of the 12 models, we calculated the weights of each metric using the entropy weight method (**Supplementary Table S2**).

The base models in the stacking structure were selected to be independent and weakly correlated. The correlation coefficients between the 12 models were calculated, and the correlation coefficient matrix was used to cluster the model by hierarchical clustering. Each cluster selected a classifier with the best performance as the base model.

RESULTS

Study Participants

The baseline characteristics of the participants are summarized in **Table 1**. Among the 858 screened patients, 4 (0.46%) were excluded, as they were over 60 years old. The majority of the included cases (46.14%) were aged between 20 and 29 years, 31.38% had their first sexual intercourse between 13 and 15 years old, 15.69% of the patients were smokers, 68.97% of patients took hormonal contraceptives, and 9.25% of the patients had sexually transmitted disease. However, only 6.44% of the performed biopsies were positive.

Predictors for a Positive Biopsy

The result of the univariate logistic regression analysis evaluating the relationship between behavioral habits, medical history, and positive biopsy is summarized in **Table 1**. The p values of age (p = 0.045), first sexual intercourse (age) (p = 0.061), number of pregnancies (p = 0.071), and use of hormonal contraceptives (years) (p = 0.007) were less than 0.1, suggesting a relationship to the occurrence of cervical cancer. Among them, the risk of cervical cancer was significantly higher in the 45–60 age group when compared with those under 20 years old (OR = 7.689, 95% CI: 1.952–30.281). Compared with those less than 13 years old for the first intercourse, the risk of cervical cancer was significantly lower in people who had sex for the first time after the age of 20 (OR = 0.132, 95% CI: 0.020–0.898). The longer use of hormonal contraceptives and a larger number of pregnancies were also features associated with an increased risk of developing cervical cancer, with ORs of 1.092 (95% CI: 1.024–1.165) and 1.180 (95% CI: 0.986–1.413) respectively.

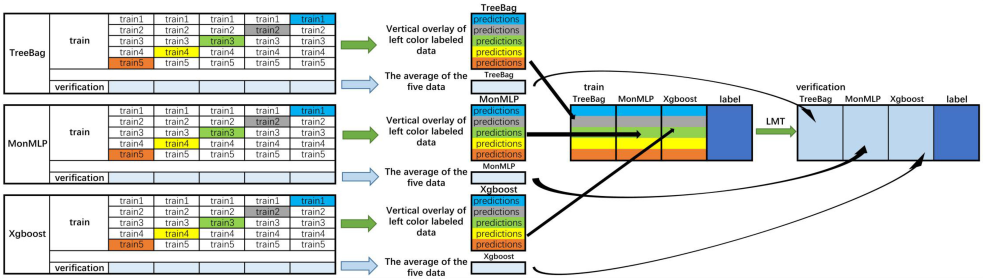


FIGURE 2 | Flowchart of the integrated stacking structure. 1) The training sets were divided into two groups of data: training and verification sets, and the training set is divided into five equal parts. 2) Take TreeBag as an example (The Figures above are Treebag, MonMLP, and XGBoost); train1, train2, train3, train4, and train5 are used as verification sets in proper sequence, and the rest are used as training sets. The model is trained by 5-fold cross-validation, and then predicted on the test set. Therefore, TreeBag can get five prediction results, which are vertically overlapped and merged into a matrix. The other two models are the same. 3) The predicted values of the three models are taken as three characteristic variables, and the resulting classifier LMT is used for fitting. Then, the reserved training set was averaged. The verification set of each characteristic variable was used to verify the performance of the LMT-stacking model. TreeBag, Bagged Classification and Regression Tree; MonMLP, Monotone Multi-Layer Perceptron Neural Network Random Over-Sampling Examples; XGBoost, eXtreme Gradient Boosting; LMT, Logistic Model Trees.

TABLE 1 | Sociodemographic factors associated with cervical cancer: univariate logistic regression analysis.

	Total (n = 854)	Biopsy negative (n = 799)	Biopsy positive (n = 55)	p	Odds ratio (95% CI)
Age, years				**0.045**	
<20	179 (20.96)	173 (21.65)	6 (10.91)		Referent
20–29	**394 (46.14)**	**366 (45.81)**	**28 (50.91)**	**0.085**	**2.206 (0.897–5.426)**
30–44	262 (30.68)	245 (30.66)	17 (30.91)	0.153	2.001 (0.773–5.178)
45–60	**19 (2.22)**	**15 (1.88)**	**4 (7.27)**	**0.004**	**7.689 (1.952–30.281)**
Number of sexual partners	2.00 (2.00–3.00)	2.00 (2.00–3.00)	2.00 (2.00–3.00)	0.986	1.001 (0.850–1.180)
First sexual intercourse(age), years				**0.061**	
<13	11 (1.29)	9 (1.13)	2 (3.64)		Referent
13–15	**268 (31.38)**	**256 (32.04)**	**12 (21.82)**	**0.063**	**0.211 (0.041–1.085)**
16–17	271 (31.73)	252 (31.54)	19 (34.55)	0.186	0.339 (0.068–1.683)
18–19	199 (23.30)	180 (22.53)	19 (34.55)	0.363	0.475 (0.096–2.361)
≥20	**105 (12.30)**	**102 (12.77)**	**3 (5.45)**	**0.038**	**0.132 (0.020–0.898)**
Num of pregnancies	**2.00 (1.00–3.00)**	**2.00 (1.00–3.00)**	**3.00 (1.00–4.00)**	**0.071**	**1.180 (0.986–1.413)**
Smoking, yes	134 (15.69)	123 (15.39)	11 (20.00)	0.365	1.374 (0.690–2.734)
	(n = 134)	**(n = 123)**	**(n = 11)**		
Smoking (years)	7.00 (2.00–11.00)	6.67 (2.00–11.00)	10.00 (3.00–15.00)	0.100	1.062 (0.988–1.141)
Smoking (packs/year)	1.38 (0.51–3.00)	1.35 (0.51–3.00)	2.00 (1.25–3.40)	0.169	1.017 (0.910–1.137)
Hormonal contraceptives, yes	589 (68.97)	553 (69.21)	36 (65.45)	0.561	0.843 (0.474–1.499)
	(n = 589)	**(n = 553)**	**(n = 36)**		
Hormonal Contraceptives(years)	**2.00 (1.00–5.00)**	**2.00 (1.00–4.50)**	**1.50 (0.50–9.50)**	**0.007**	**1.092 (1.024–1.165)**
IUD, yes	199 (23.30)	187 (23.40)	12 (21.82)	0.788	0.913 (0.472–1.768)
	(n = 199)	**(n = 187)**	**(n = 12)**		
IUD (years)	2.19 (1.60–3.77)	2.17 (1.56–3.65)	3.00 (2.50–4.88)	0.352	1.081 (0.918–1.272)
STDs, yes	79 (9.25)	67 (8.39)	12 (21.82)	0.395	2.000 (0.406–9.886)
	(n = 79)	**(n = 67)**	**(n = 12)**		
Number of STDs	2.00 (1.00–2.00)	2.00 (1.00–2.00)	2.00 (1.00–2.00)	0.926	0.958 (0.388–2.365)
STDs: condylomatosis	44 (55.70)	37 (55.22)	7 (58.33)	0.842	1.135 (0.327–3.941)
STDs: vaginal condylomatosis	4 (5.06)	4 (5.97)	0 (0.00)	/	/
STDs: vulvo-perineal condylomatosis	43 (54.43)	36 (53.73)	7 (58.33)	0.768	1.206 (0.347–4.183)
STDs: syphilis	18 (22.78)	18 (26.87)	0 (0.00)	/	/
STDs: HIV	18 (22.78)	13 (19.40)	5 (41.67)	0.100	2.967 (0.811–10.861)

Portions in bold represent p < 0.1. IUD, intrauterine device; STD, sexually transmitted disease.

The feature selection method using RF was applied. **Figure 3** demonstrated the relative importance of 18 variables in cervical cancer risk prediction. Based on this analysis, nine predictors had relative importance greater than one. The use of hormonal contraceptives (years) was identified as the most important risk factor, followed by the number of pregnancies, smoking (years), number of cigarette packets smoked annually, number of sexual partners, the use of an intrauterine device (IUD) (years), number of sexually transmitted diseases (STDs), human immunodeficiency virus (HIV), and age. These nine features were incorporated into the model and cross-validated. In contrast, in the univariate logistic regression, the number of

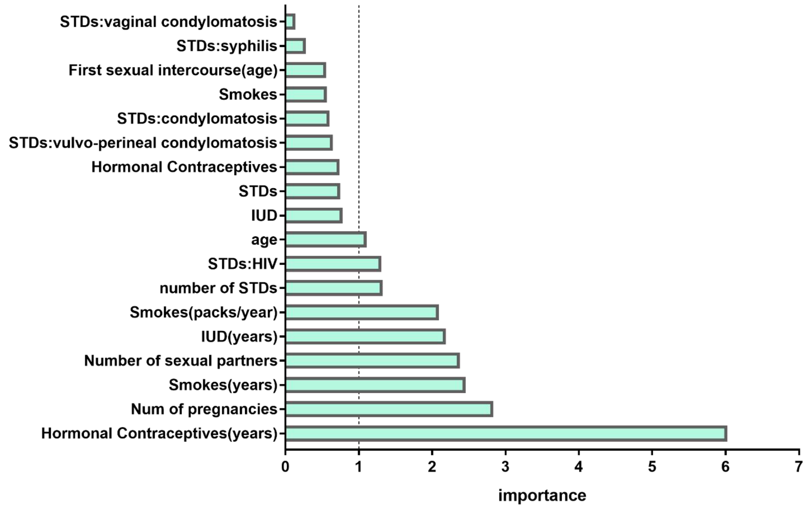


FIGURE 3 | Variable importance measures for each predictor of morbidity. IUD, intrauterine device; STD, sexually transmitted disease; HIV, human immunodeficiency virus.

sexual partners was not significantly correlated (p = 0.986) with the occurrence of cervical cancer.

These nine features were incorporated into the model and cross-validated.

Prediction Performance of the Sampling Method

Table 2 described the comparative performance scores of different sampling methods using RF. Each sampling model had been verified internally and externally. In the external validation, SMOTE-based RF performed best among all classifiers with an AUC of 0.849 and had the highest score in four of our seven performance metrics. The sensitivity and specificity were 90.9% and 73.1%, respectively, both higher than 70%. The accuracy, precision, F1 score, and F2 score were 74.2%, 18.9%, 0.312, and 0.195, respectively. SMOTE was therefore selected as the imbalance data processing algorithm for the final model.

Evaluation of the model performance using the receiver-operating characteristic (ROC) (**Supplementary Figure S2**) showed the comparison of the prediction ability of external and internal validation of the model under different sampling models. The curves modeled the sensitivity proportion of actual at-risk women identified at risk of developing cervical cancer to the specificity proportion of identified no-risk women in the models.

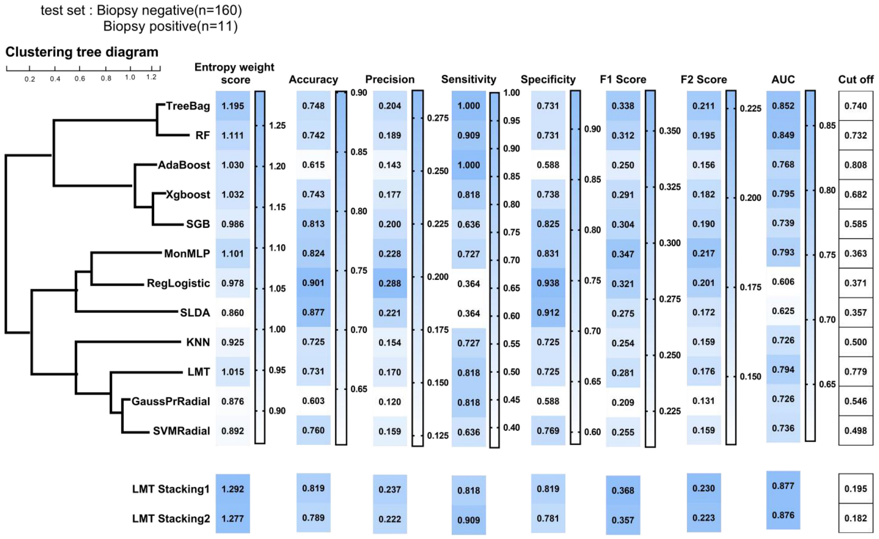
Prediction Performance of 12 Machine Learning Models

Toward at-risk patients of cervical cancer classification, **Figure 4** compared the performance metrics of 12 different models. According to the entropy weight score, TreeBag resulted in the best performance, with an AUC score of 0.852 for the test dataset. The sensitivity and specificity were 100% and 73.1%, respectively. Compared to RF, the performance of sensitivity and AUC was improved. As a whole, the tree-based models (TreeBag, RF, Adaboost, XGBoost, SGB, and LMT) performed better than other models, and the performance difference between the models was minor. Additionally, the performance of the deep learning model MonMLP ranked third, with an AUC of 0.793

TABLE 2 | Prediction performance of random forest algorithm on different sampling models.

Methods		Cutoff	Accuracy	Precision	Sensitivity	Specificity	F1 Score	F2 Score	AUC
Oversampling	Train set	0.703	0.978	0.749	1.000	0.977	0.857	0.535	0.997
	Test set	0.099	0.660	0.159	**1.000**	0.637	0.275	0.172	0.803
Undersampling	Train set	0.333	0.761	0.191	0.840	0.756	0.312	0.195	0.870
	Test set	**0.343**	**0.743**	0.163	0.727	**0.744**	0.267	0.167	0.739
Both sampling	Train set	0.672	0.947	0.550	0.977	0.945	0.704	0.440	0.988
	Test set	0.191	0.597	0.138	**1.000**	0.569	0.242	0.151	0.784
ROSE	Train set	0.270	0.773	0.171	0.659	0.781	0.272	0.170	0.733
	Test set	0.178	0.632	0.129	0.818	0.619	0.222	0.139	0.745
SMOTE	Train set	0.600	0.952	0.586	0.864	0.958	0.698	0.436	0.968
	Test set	0.268	0.742	**0.189**	0.909	0.731	**0.312**	**0.195**	**0.849**

The portions in bold represent the model is optimal in a single index. ROSE, random oversampling example; SMOTE, synthetic minority oversampling technique; AUC, area under the curve.



	Entropy weight score	Accuracy	Precision	Sensitivity	Specificity	F1 Score	F2 Score	AUC	Cut off
TreeBag	1.195	0.748	0.204	1.000	0.731	0.338	0.211	0.852	0.740
RF	1.111	0.742	0.189	0.909	0.731	0.312	0.195	0.849	0.732
AdaBoost	1.030	0.615	0.143	1.000	0.588	0.250	0.156	0.768	0.808
Xgboost	1.032	0.743	0.177	0.818	0.738	0.291	0.182	0.795	0.682
SGB	0.986	0.813	0.200	0.636	0.825	0.304	0.190	0.739	0.585
MonMLP	1.101	0.824	0.228	0.727	0.831	0.347	0.217	0.793	0.363
RegLogistic	0.978	0.901	0.288	0.364	0.938	0.321	0.201	0.606	0.371
SLDA	0.860	0.877	0.221	0.364	0.912	0.275	0.172	0.625	0.357
KNN	0.925	0.725	0.154	0.727	0.725	0.254	0.159	0.726	0.500
LMT	1.015	0.731	0.170	0.818	0.725	0.281	0.176	0.794	0.779
GaussPrRadial	0.876	0.603	0.120	0.818	0.588	0.209	0.131	0.726	0.546
SVMRadial	0.892	0.760	0.159	0.636	0.769	0.255	0.159	0.736	0.498
LMT Stacking1	1.292	0.819	0.237	0.818	0.819	0.368	0.230	0.877	0.195
LMT Stacking2	1.277	0.789	0.222	0.909	0.781	0.357	0.223	0.876	0.182

FIGURE 4 | Prediction performance of ML models on the test sample. ML, machine learning; TreeBag, Bagged Classification and Regression Tree; MonMLP, Monotone Multi-Layer Perceptron Neural Network Random Over-Sampling Examples; XGBoost, eXtreme Gradient Boosting; LMT, Logistic Model Trees; RF, random forest; SGB, Stochastic Gradient Boosting; SVMRadial, Support Vector Machines with Radial Basis Function Kernel; KNN, K-Nearest Neighbors; GaussPrRadial, Gaussian Process with Radial Basis Function Kernel; RgeLogistic, Regularized Logistic Regression; SLDA, Stabilized Linear Discriminant; AdaBoost, AdaBoost Classification Trees; AUC, area under the curve.

and sensitivity and specificity of 72.7% and 83.1%, respectively. The MonMLP model was significantly better than other models with top performance in terms of specificity. The tuned parameters of these models were listed in **Supplementary Table S3**.

According to the correlation results of the 12 models (**Supplementary Figure S4**), we divided the 12 models into 4 clusters (**Figure 4**) by using the hierarchical clustering method. The intra-cluster model prediction difference was small, while the inter-cluster model was large. In the first group, TreeBag and RF were included, and the correlation between them was as high as 0.80. Treebag was better than RF in predicting high-risk patients with cervical cancer. According to the hierarchical clustering results, AdaBoost, XGBoost, and SGB belonged to the tree model based on boosting integration and were divided into the second group. The correlation between the three models was greater than 0.50. The best model was XGBoost with an AUC, sensitivity, and specificity of 0.795, 81.8%, and 73.8%, respectively. The third group consisted of the MonMLP model and two simplistic models (RgeLogistic and SLDA). In terms of performance, MonMLP performed better than the other two models. This was partly due to the small number of positive biopsies, and therefore the two simplistic models could not learn enough logical relationships. In the fourth group, only LMT, KNN, GaussPrRadial, and SVMRadial performed well.

Prediction Performance of Stacking Models

In order to meet the two requirements of the stacking structure for the base classifier and improve the performance (27), we selected an optimal model from each group, namely, TreeBag, XGBoost, MonMLP, and LMT. The performance ranking of those models might be TreeBag > MonMLP > XGBoost > LMT. LMT model was a simpler model based on the Logistic and tree model, with high generalization and strong generalization robustness (26). Therefore, we chose LMT as the second layer structure of stacking (result classifier) and TreeBag, XGBoost, and MonMLP as the first layer (base classifier). Finally, two LMT-stacking models with different tuning parameters were built by training (**Supplementary Table S3**). The AUC, sensitivity, and specificity of the LMT-Stacking1 model were 0.877, 81.8%, and 81.9% (**Figure 4**), respectively, and 0.877, 81.8%, and 90.9%, respectively, for the LMT-Stacking2 model. The difference in AUC between the two models was only 0.1%, and the performance difference was not significant. Similar results were seen in the ROC curves for each of the models, as shown in **Supplementary Figure S5**.

DISCUSSION

AI and ML algorithms are increasingly used in healthcare to analyze large datasets and perform predictions. However, the use of these algorithms in identifying women at high risk of developing cervical cancer is limited and often based on former generation models, which have more limited accuracy than more advanced algorithms. In this study, we have proposed the use of SIML that integrates multiple algorithms to improve the prediction accuracy.

The findings of this study indicated that various ML algorithms could be used to predict women at high risk of developing cervical cancer based on demographic, behavioral, and clinical data. However, the SIML with TreeBag, XGBoost, and MonMLP as base classifier and LMT as result classifier provided the best overall performance. Compared with the LMT-Stacking1 model, the sensitivity of the LMT-Stacking2 model was highly improved, while the specificity decreased. However, because the data had few positive samples and the sensitivity varied significantly, the performance of the LMT-Stacking1 resulted in a better overall performance because it was more balanced.

Predictors for Developing Cervical Cancer

According to the feature selection based on RF, hormonal contraceptives (years), the number of pregnancies, smoking (years), the number of sexual partners, the use of IUD (years), and smoking (packs/year) were identified to be the most important influencing factors for the at-risk patient, especially the long-term use of hormone contraceptives. Human papillomavirus (HPV) infection was the leading cause of cervical cancer (31). According to Cox (32), the risk of developing an HPV infection was not only related to age but also increased with the increasing number of sexual partners, highlighting the need to improve awareness and improve vaccination campaigns. Co-infection with HIV might impair the ability of the immune system to control HPV infection. Additional risk factors included smoking, high parity, and long-term use of hormonal contraceptives (31). Exogenous hormones had been considered as auxiliary factors in the pathogenesis of cervical cancer caused by HPV. If the HPV-positive women took the hormone contraceptives for a long time, the risk of cervical squamous cell carcinoma tripled (33). Smoking was related to the development of squamous cell carcinoma and was an auxiliary factor and primary carcinogen in the development of cervical cancer (34). The use of IUD could create a potential malignant focus close to the cervical canal, eventually creating a transformation zone whereby preneoplastic lesions arise. The transformation zone was both targeted by HPV and a major effecter and inductive site for cell-mediated immune response (35).

Machine Learning and Cervical Cancer

Most studies on cervical cancer made use of ML to predict survival in cervical cancer (36). Although some studies had used generalized estimating equation regression models to predict the early risk probability of developing cervical cancer (34), their prediction accuracy remained limited. Our ML model utilized more features and could, therefore, improve the prediction accuracy. The Pittsburgh cervical cancer screening model consisted of 19 variables, including cytological examination and HPV test results. The incidence of cervical cancer was predicted by combining the case results, detailed medical history [including gender, HPV vaccination status, menstruation, contraception history, age, and race (37)]. The model could be used for risk stratification of patients only after screening. The advantage of our proposed model was that it provided a simple tool to identify high-risk groups before screening by combining behavioral data provided by patients with clinical data.

LIMITATIONS

The main limitation of this study was the limited sample size and population coverage. Compared with deep learning, SIML had the advantage of being suitable for small sample data, which only needed 80–560 samples. The specific sample size required depended on the dataset and sampling method (38). Therefore, the sample size in our study was sufficient to build a model. If the overall sample size was increased, the performance of the model could be improved significantly. Additionally, some potentially important parameters, such as previous screening information, were not considered in our study. Data on variation in behavioral patterns over time were not available, and therefore, we could not establish their impact on the model. Moreover, samples were obtained from the same institution, limiting the generalizability of the model. Although we used a combination of internal and external validation, we recommend the use of external datasets to further test the performance of this model.

CONCLUSIONS

This study shows that SIML can be used to accurately identify women at high risk of developing cervical cancer and performed better than other ML algorithms. This model could be used to personalize the screening program by optimizing the screening frequency and improving the care plan in high- and low-risk women based on their demographics, behavioral patterns, and clinical data. This will eventually reduce unnecessary screening in low-risk groups and hence reduce the screening costs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

BP and LS contributed to the protocol design. LS and QZ analyzed data. XL, YG, and QC contributed to the interpretation results. LS drafted the article, and LY proofread the article. BP, LT, and ZL revised the final version and are guarantors of this article. All authors made substantial contributions to the paper and read and approved the final article.

FUNDING

This study was funded by the National Key Research and Development Program of China (No. 2018YFC1311706).

ACKNOWLEDGMENTS

Here, we would like to thank Kelvin Fernandes for providing data (13). We also want to thank Professor Qiulin Shi, from the School of Public Health and Management at Chongqing Medical University, for her indispensable contributions to the statistical design and analysis. We would like to thank Xiyao Liu from the School of the First Clinical College at Chongqing Medical University for her comments on the article. We would like to thank Yan Zhan from the University of California, Berkeley, for his proofreading of the article and TopEdit (www.topeditsci.com) for the English language editing of this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.821453/full#supplementary-material

Supplementary Figure 1 | Visualization results before and after missing values was filled. The part of red color is the missing value, with each column as the standard. The larger the value is, the darker the color is. On the contrary, the smaller the value is, the lighter the color is.

Supplementary Figure 2 | Receiver operating characteristic curves for Random Forest prediction performance of difference Sampling models.

Supplementary Figure 3 | Receiver operating characteristic curves for 12 ML models.

Supplementary Figure 4 | Correlation coefficient diagrams of 12 ML models.

Supplementary Figure 5 | Receiver operating characteristic curves for LMT-stackingmodels.

REFERENCES

1. Wang L, Zhao Y, Xiong W, Ye W, Zhao W, Hua Y. MicroRNA-449a Is Downregulated in Cervical Cancer and Inhibits Proliferation, Migration, and Invasion. *Oncol Res Treat* (2019) 42(11):564–71. doi: 10.1159/000502122
2. Chao X, Li L, Wu M, Ma S, Tan X, Zhong S, et al. Efficacy of Different Surgical Approaches in the Clinical and Survival Outcomes of Patients With Early-Stage Cervical Cancer: Protocol of a Phase III Multicentre Randomised Controlled Trial in China. *BMJ Open* (2019) 9(7):e029055. doi: 10.1136/bmjopen-2019-029055
3. Choi JB, Lee WK, Lee SG, Ryu H, Lee CR, Kang SW, et al. Long-Term Oncologic Outcomes of Papillary Thyroid Microcarcinoma According to the Presence of Clinically Apparent Lymph Node Metastasis: A Large Retrospective Analysis of 5,348 Patients. *Cancer Manag Res* (2018) 10:2883–91. doi: 10.2147/CMAR.S173853
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2018) 68 (6):394–424. doi: 10.3322/caac.21492
5. Campos NG, Alfaro K, Maza M, Sy S, Melendez M, Masch R, et al. The Cost-Effectiveness of Human Papillomavirus Self-Collection Among Cervical Cancer Screening non-Attenders in El Salvador. *Prev Med* (2020) 131:105931. doi: 10.1016/j.ypmed.2019.105931
6. Cheung A, Figueredo C, Rinella ME. Nonalcoholic Fatty Liver Disease: Identification and Management of High-Risk Patients. *Am J Gastroenterol* (2019) 114(4):579–90. doi: 10.14309/ajg.0000000000000058
7. Huang J, Qian Z, Gong Y, Wang Y, Guan Y, Han Y, et al. Comprehensive Genomic Variation Profiling of Cervical Intraepithelial Neoplasia and Cervical Cancer Identifies Potential Targets for Cervical Cancer Early Warning. *J Med Genet* (2019) 56(3):186–94. doi: 10.1136/jmedgenet-2018-105745
8. Teoh D, Vogel RI, Langer A, Bharucha J, Geller MA, Harwood E, et al. Effect of an Electronic Health Record Decision Support Alert to Decrease Excess Cervical Cancer Screening. *J Low Genit Tract Dis* (2019) 23(4):253–8. doi: 10.1097/LGT.0000000000000484
9. van der Waal D, Bekkers RLM, Dick S, Lenselink CH, Massuger L, Melchers W, et al. Risk Prediction of Cervical Abnormalities: The Value of Sociodemographic and Lifestyle Factors in Addition to HPV Status. *Prev Med* (2020) 130:105927. doi: 10.1016/j.ypmed.2019.105927
10. Dinh A, Miertschin S, Young A, Mohanty SD. A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease With Machine Learning. *BMC Med Inform Decis Mak* (2019) 19(1):211. doi: 10.1186/s12911-019-0918-5
11. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* (2017) 376(26):2507–9. doi: 10.1056/NEJMp1702071
12. Weegar R, Sundstrom K. Using Machine Learning for Predicting Cervical Cancer From Swedish Electronic Health Records by Mining Hierarchical Representations. *PloS One* (2020) 15(8):e0237911. doi: 10.1371/journal.pone.0237911
13. Fernandes K, Cardoso JS, Fernandes J. (2017). Transfer Learning With Partial Observability Applied to Cervical Cancer Screening, in: *Paper presented at: Iberian Conference on Pattern Recognition and Image Analysis*. China: Springer.
14. Di Guida R, Engel J, Allwood JW, Weber RJ, Jones MR, Sommer U, et al. Non-Targeted UHPLC-MS Metabolomic Data Processing Methods: A Comparative Investigation of Normalisation, Missing Value Imputation, Transformation and Scaling. *Metabolomics* (2016) 12:93. doi: 10.1007/s11306-016-1030-9
15. Stekhoven DJ. *Missforest: Nonparametric Missing Value Imputation Using Random Forest.* (2013).
16. Breiman L. Random Forests. *Mach Learn* (2001) 45(1):5–32. doi: 10.1023/A:1010933404324
17. He HB, Garcia EA. Learning From Imbalanced Data. *IEEE Trans Knowl Data Eng* (2009) 21(9):1263–84. doi: 10.1109/TKDE.2008.239
18. Menardi G, Torelli N. Reducing Data Dimension for Cluster Detection. *J Stat Comput Simul* (2013) 83(11):2047–63. doi: 10.1080/00949655.2012.679032
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. *J Artif Intell Res* (2002) 16:321–57. doi: 10.1613/jair.953
20. Bernau C, Riester M, Boulesteix AL, Parmigiani G, Huttenhower C, Waldron L, et al. Cross-Study Validation for the Assessment of Prediction Algorithms. *Bioinformatics* (2014) 30(12):105–12. doi: 10.1093/bioinformatics/btu279
21. Olson RS, Cava W, Mustahsan Z, Varik A, Moore JH. Data-Driven Advice for Applying Machine Learning to Bioinformatics Problems. *Pac Symp Biocomput* (2018) 23:192–203. doi: 10.1142/9789813235533_0018
22. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proc 22nd ACM SIGKDD Int Conf Knowl Discovery and Data Mining* (2016). New York, United States: Association for Computing Machinery.
23. Sun WZ, Jiang MY, Ren L, Dang J, You T, Yin FF. Respiratory Signal Prediction Based on Adaptive Boosting and Multi-Layer Perceptron Neural Network. *Phys Med Biol* (2017) 62(17):6822–35. doi: 10.1088/1361-6560/aa7cd4
24. Cortes C, Vapnik V. Support Vector Networks. *Mach Learn* (1995) 20(3):273–97. doi: 10.1007/BF00994018
25. Zhang SC. Nearest Neighbor Selection for Iteratively kNN Imputation. *J Syst Software* (2012) 85(11):2541–52. doi: 10.1016/j.jss.2012.05.073
26. Landwehr N, Hall M, Frank E. Logistic Model Trees. *Mach Learn* (2005) 59(1-2):161–205. doi: 10.1007/s10994-005-0466-3

27. Seewald AK. (2002). How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness, in: *Paper presented at: Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)*, University of New South Wales, Sydney, Australia, July 8-12, 2002.
28. Zhou ZH. *Ensemble Methods - Foundations and Algorithms*. New York: Taylor & Francis (2012).
29. Džeroski SaŽ B. Is Combining Classifiers With Stacking Better Than Selecting the Best One? *Mach Learn* (2004) 54(3):255–73. doi: 10.1023/B:MACH.0000015881.36452.6e
30. Goutte C, Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, With Implication for Evaluation. *Lect Notes Comput Sci* (2005) 3408:345–59. doi: 10.1007/978-3-540-31865-1_25
31. Vesco KK, Whitlock EP, Eder M, Burda BU, Senger CA, Lutz K. Risk Factors and Other Epidemiologic Considerations for Cervical Cancer Screening: A Narrative Review for the U.S. Preventive Services Task Force. *Ann Intern Med* (2011) 155(10):698–705, W216. doi: 10.7326/0003-4819-155-10-201111150-00377
32. Cox JT. The Development of Cervical Cancer and its Precursors: What Is the Role of Human Papillomavirus Infection? *Curr Opin Obstet Gynecol* (2006) 18 Suppl 1:s5–13. doi: 10.1097/01.gco.0000216315.72572.fb
33. Moreno V, Bosch FX, Munoz N, Meijer CJ, Shah KV, Walboomers JM, et al. Effect of Oral Contraceptives on Risk of Cervical Cancer in Women With Human Papillomavirus Infection: The IARC Multicentric Case-Control Study. *Lancet* (2002) 359(9312):1085–92. doi: 10.1016/S0140-6736(02)08150-3
34. Fang JH, Yu XM, Zhang SH, Yang Y. Effect of Smoking on High-Grade Cervical Cancer in Women on the Basis of Human Papillomavirus Infection Studies. *J Cancer Res Ther* (2018) 14(Supplement):S184–9. doi: 10.4103/0973-1482.179190
35. Cortessis VK, Barrett M, Brown Wade N, Enebish T, Perrigo JL, Tobin J, et al. Intrauterine Device Use and Cervical Cancer Risk: A Systematic Review and Meta-Analysis. *Obstet Gynecol* (2017) 130(6):1226–36. doi: 10.1097/AOG.0000000000002307
36. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, et al. Survival Outcome Prediction in Cervical Cancer: Cox Models vs Deep-Learning Model. *Am J Obstet Gynecol* (2019) 220(4):381.e381–381.e314. doi: 10.1016/j.ajog.2018.12.030
37. Austin RM, Onisko A, Druzdzel MJ. The Pittsburgh Cervical Cancer Screening Model: A Risk Assessment Tool. *Arch Pathol Lab Med* (2010) 134(5):744–50. doi: 10.5858/134.5.744
38. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting Sample Size Required for Classification Performance. *BMC Med Inform Decis Mak* (2012) 12:8. doi: 10.1186/1472-6947-12-8

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Reveal the Heterogeneity in the Tumor Microenvironment of Pancreatic Cancer and Analyze the Differences in Prognosis and Immunotherapy Responses of Distinct Immune Subtypes

Xiaoqin Wang[1*†], *Lifang Li*[2†], *Yang Yang*[3†], *Linlin Fan*[1], *Ying Ma*[1] *and Feifei Mao*[4*]

[1] Department of Clinical Laboratory, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China, [2] Emergency Department, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China, [3] Department of Hepatobiliary and Pancreatic Surgery, The First People's Hospital of Changzhou, Changzhou, China, [4] Tongji University Cancer Center, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China

OPEN ACCESS

Edited by:
Fu Wang,
Xi'an Jiaotong University, China

Reviewed by:
Yi Xianfu,
Tianjin Medical University, China
Meng Wang,
University of California, San Francisco, United States

****Correspondence:***
Feifei Mao
maofeifei01@126.com
Xiaoqin Wang
xiaoqinwang2@outlook.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *10 December 2021*
Accepted: *19 January 2022*
Published: *17 February 2022*

Citation:
Wang X, Li L, Yang Y, Fan L, Ma Y and Mao F (2022) Reveal the Heterogeneity in the Tumor Microenvironment of Pancreatic Cancer and Analyze the Differences in Prognosis and Immunotherapy Responses of Distinct Immune Subtypes. Front. Oncol. 12:832715. doi: 10.3389/fonc.2022.832715

Purpose: The current clinical classification of pancreatic ductal adenocarcinoma (PDAC) cannot well predict the patient's possible response to the treatment plan, nor can it predict the patient's prognosis. We use the gene expression patterns of PDAC patients to reveal the heterogeneity of the tumor microenvironment of pancreatic cancer and analyze the differences in the prognosis and immunotherapy response of different immune subtypes.

Methods: Firstly, use ICGC's PACA-AU PDAC expression profile data, combined with the ssGSEA algorithm, to analyze the immune enrichment of the patient's tumor microenvironment. Subsequently, the spectral clustering algorithm was used to extract different classifications, the PDAC cohort was divided into four subtypes, and the correlation between immune subtypes and clinical characteristics and survival prognosis was established. The patient's risk index is obtained through the prognostic prediction model, and the correlation between the risk index and immune cells is prompted.

Results: We can divide the PDAC cohort into four subtypes: immune cell and stromal cell enrichment (Immune-enrich-Stroma), non-immune enrichment but stromal cell enrichment (Non-immune-Stroma), immune-enriched Collective but non-matrix enrichment (Immune-enrich-non-Stroma) and non-immune enrichment and non-stromal cell enrichment (Non-immune-non-Stroma). The five-year survival rate of immune-enrich-Stroma and non-immune-Stroma of PACA-CA is quite different. TCGA-PAAD's immune-enrich-Stroma and immune-enrich-non-Stroma groups have a large difference in productivity in one year. The results of the correlation analysis between the risk index and immune cells show that the patient's disease risk is significantly related to epithelial cells, megakaryocyte-erythroid progenitor (MEP), and Th2 cells.

Conclusion: The tumor gene expression characteristics of pancreatic cancer patients are related to immune response, leading to morphologically recognizable PDAC subtypes with prognostic/predictive significance.

Keywords: pancreatic cancer, immune subtypes, heterogeneity, prognosis, microenvironment

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is one of the lethal malignant neoplasms around the world (1–4), and its genetic and phenotypic heterogeneity makes generally effective therapies ineffective (5–9). The salient feature of pancreatic cancer is that it has an immunosuppressive microenvironment, the prognosis of patients is poor, and most of the patients' tumors will metastasize (10, 11). Research on the immune microenvironment of pancreatic cancer may help improve the therapeutic effect (12, 13). By detecting the expression of anti-tumor immune genes, markers that can predict patient response to treatment have been screened (14). In addition, mutations in genes such as PIK3CA, FGFR3, and TP53 have been shown to be related to tumor immune infiltration (15–18). Although we have a better understanding of the molecular mechanism and genetic background of pancreatic cancer, the 5-year survival rate for this disease is approximately 10% in the USA (19). Several phase III clinical trials that are effective for other cancers have not worked well in pancreatic cancer patients (7). Tumor heterogeneity and host differences will affect the characteristics of its tumor microenvironment. It is necessary to identify new biomarkers and explore new treatment approaches to provide more and more effective references for overcoming the immunosuppressive mechanism in the pancreatic cancer microenvironment.

The immune microenvironment plays an important role in tumor cell invasion and pancreatic cancer progression (20), and immune expression characteristics may affect the degree of inhibition of cancer cells. Invasive PDAC has epithelial-to-mesenchymal transition (EMT)-like characteristics and has been shown to be a poor prognostic factor for pancreatic cancer (21). The immune microenvironment with EMT-like tumors is conducive to tumor growth. Researchers reported on three subtypes of pancreatic cancer: classic, quasi-mesenchymal, and exocrine, and clarified the genetic markers of different subtypes, which may help to carry out more targeted treatments for patients (22). Other researchers have identified two tumor-specific subtypes based on gene expression: basal-like subtype and classical subtype (23). The classic subtype is consistent with the subtype described by Collisson et al. Tumor subtypes defined by exocrine-like genes have not been validated in its data set, and may be related to tissue contamination. Recently, researchers classified pancreatic cancer into four subtypes based on genomic studies—squamous cells, pancreatic progenitor cells, immunogenicity and abnormally differentiated endocrine and exocrine-identified the differences between pancreatic cancer subtypes and provided Different subtypes of treatment options (22, 24). Among them, squamous cells, pancreatic progenitor cells, and abnormally differentiated endocrine and exocrine (ADEX) subtypes correspond to the quasi-mesenchymal, classical, and exocrine-like subtypes reported by Collisson et al. (22). Recently, studies have shown that ADEX and immunogenic subtypes are related to the lower purity of the sample (24, 25). Although researchers have basically determined the characteristics of some pancreatic cancer subtypes, research conclusions about exocrine differentiation or immunogenic subtypes are still inconsistent.

Therefore, we aim to redefine the subtypes of PDAC and clarify its immune expression patterns, provide useful clues for exploring the different immunosuppressive mechanisms of PDAC, and use it in the stratification of patient clinical trials, so as to provide patients with PDAC more precise treatment.

RESULTS

Classification of Distinct Tumor Microenvironment Subtypes

Single sample gene set enrichment analysis (ssGSEA) defines an enrichment score to indicate the absolute enrichment degree of the gene set in each sample in a given data set. The enrichment score of each immune category can be found in the R package GSVA In the realization (26). Firstly, ssGSEA algorithm (27) was utilized to analyze the expression profiling database of the PACA-AU pancreatic cancer in the International Cancer Genome Consortium (ICGC). We obtained the immune enrichment of the tumor microenvironment of each patient's tumor tissue. And the tumor microenvironment-related genes come from the following references (**Table 1**).

Subsequently, we apply the spectral clustering algorithm to extract different categories based on the ssGSEA scores (**Figure 1A**). Meanwhile, we used t-distributed stochastic neighbor embedding (tSNE) to show the groups (**Figure 1B**), and revealed an immune-enriched subtype (Immune-enrich) exists in the cohort, and the rest are of the Non-immune type, that is, less immune infiltration (**Figure 1C**). In addition, even in the presence of a large population of immune cells, stromal cells also play vital roles in tumor immunity evasion. Therefore, we further dissected the enrichment of stromal cells in the patient's gene expression profile. Also using ssGSEA analysis, we found that the cohort had characteristics of activated stromal response (**Figure 1C**). Based on the above classification, we can divide the pancreatic cancer cohort into four subtypes: immune cell and stromal cell enrichment (Immune-enrich-Stroma), non-immune enrichment but stromal cell enrichment (Non-immune-Stroma), Immune enrichment but non-matrix enrichment (Immune-enrich-non-Stroma) and non-immune enrichment and non-stromal cell enrichment (Non-immune-non-Stroma). Immune-

TABLE 1 | Immune-related gene signatures and their references.

Signature name	Reference
Immune enrichment score	Yoshihara et al. Nat Commun. 2013 (28)
6-gene IFN-γsignature	Chow et al. J Clin Oncol. 2016 (suppl) (29)
Activated stroma	Moffitt et al. Nat Genet. 2015 (30)
Immune cell subsets	Cancer Genome Atlas Network. Cell. 2015 (31)
T cells	Bindea et al. Immunity. 2013 (32)
CD8 Tcells	Bindea et al. Immunity. 2013 (32)
T. NK. metagene	Alistar et al. Genome Med. 2014 (33)
B-cell cluster	Iglesia et al. Clin Cancer Res. 2014 (34)
Macrophages	Bindea et al. Immunity. 2013 (32)
Cytotoxic cells	Bindea et al. Immunity. 2013 (32)
Immunophenoscore	Charoentong et al. Cell Rep. 2017 (35)
T cell-inflamed GEP	Cristescu et al. Science. 2018 (36)
Expanded immune signature	Ayers et al. J Clin Invest. 2017 (37)
TGF-β-associated ECM	Chakravarthy et al. Nat Commun. 2018) (38)
MDSC	Yaddanapudi et al. Cancer Immunol Res. 2016 (39)
CAF	Calon et al. Cancer Cell. 2012 (40)
TAM M2/M1	Beyer et al. PLoS One. 2012 (41)
CD8 T cell exhaustion	Giordano et al. EMBO J. 2015 (42)
T cell exhaustion early/late stage	Philip et al. Nature. 2017 (43)
Nivolumab responsive	Riaz et al. Cell. 2017 (44)

enrich-Stroma subtypes mainly enrich tumor immune-related molecular signatures, including T cell-inflamed GEP, Expanded immune signature, Immunophenoscore, Immune enrichment score, CD8 T cell exhaustion, myeloid-derived suppressor cells (MDSC), cytotoxic cells, Immune cell subset, etc. At the same time, it also enriches PD1 and stroma related signatures, including anti-PD-1 resistant, nivolumab responsive and normal stroma. The signatures of Non-immune-Stroma subtypes mainly include anti-PD-1 resistant, activated stroma, CAF-stimulated, and normal stroma, while its immune-related family features are very low. Immune-enrich-non-Stroma subtypes mainly enrich tumor immune-related signatures, including T cell-inflamed GEP, Expanded immune signature and cytotoxic cells, etc., while its stromal signatures expression is very low. Non-immune-non-Stroma subtypes, as the name suggests, are rarely enriched in tumor immunity and stromal signatures.

Comparison of the Striking Differences in the Immune Microenvironment of the Four Subtypes

As follow, the four subtypes have the following immune differences (**Figure 2A**). Patients with immuno-enriched subtypes (**Figure 2A** red and light blue boxes) showed significant enrichment in the characteristics of recognizing immune cells or immune responses (all P <0.05). We further compared the difference in gene expression between immune-enriched and non-immune-enriched patients, mainly using the limma algorithm, and P<0.05 as the standard of significant difference (**Table S1**). At the same time, the significantly different genes of stromal cell enrichment and non-stromal enrichment was compared (**Table S2**).In order to verify the accuracy and consistency of the analysis method, we use the same strategy to predict the enrichment of other data. The first 50 genes that are differentially up-regulated are selected to construct a gene set, and the ssGSEA algorithm is used to predict the enrichment of other data. In addition, select significantly different immune activity or immune cell-related genes to verify their enrichment. The analysis results show that the GSE124231 data set (**Figure 2B**, n=48), the GSE131050 data set (**Figure 2C**, n=66), the PACA-CA data set (**Figure 2D**, n=234) and the TCGA-PAAD database (**Figure 2E**, n = 177) can be divided into immune enrichment and stromal cell enrichment groups. According to the constructed gene set, samples of different data sets can be divided into immune-enrich-Stroma, immune-enrich-non-stroma, non-immune-stroma and non-immune-non-stroma types. And immune enrichment type samples are mainly enriched for immune-related signatures, such as immune enrichment score, immunophenoscore, Immune cell subsets, etc. Stromal cell enrichment types mainly enrich stroma-related signatures, such as normal stroma, activated stromanivolumab responsive, etc. The above results show that the accuracy and consistency of our classification and research methods are trustworthy.

Four Immune Subtypes Are Related to Clinical Characteristics and Survival Prognosis

Based on the previous results, we have divided patients into 4 different subtypes of immune enrichment and stromal cell enrichment. Therefore, we need to further compare the clinical characteristics of different types and try to explore the relationship between each type and patient survival prognosis. Firstly, we sequentially compared the clinical information between different subtypes in the PACA-AU, PACA-CA and TCGA-PAAD cohorts. Statistics showed that there were significant differences among subtypes in the PACA-AU cohort, which included donor_sex, donor_vital_status, donor_relapse_type, donor_age_at_diagnosis and enrollment, donor_survival_time, donor_interval_up, donor_interval_up, donor_interval_up (**Table S3**). In the PACA-CAcohort, clinical markers such as donor_age_at_diagnosis and enrollment, donor_age_at_last_followup, donor_survival_time, donor_interval_of_last_followup are significantly different among subgroups (**Table S4**). Age_at_initial_pathologic_diagnosis, family_history_of_cancer (%), history_of_chronic_pancreatitis (%), history_of_diabetes (%) and other clinical indicators were significantly different among 4 subsets in the TCGA-PAAD cohort (**Table S5**).

Then, we successively explored the relationship between different subgroups in the cohort and the survival prognosis of patients. In the PACA-AU cohort, the 1-year (**Figure 3A**) and 5-year (**Figure 3C**) survival rates between different subtypes are significantly different (p.value <0.05), and the survival rate of the Immune_enrich_Stroma subgroup is higher than that of the other three groups. However, the difference in 3-year survival rates between patient groups was not significant (**Figure 3B**). Finally, we compared the survival rates of all PACA-AU patients (8 years) and found that the survival rates of different subgroups are still significantly different (p.value <0.05) (**Figure 3D**). Similarly, we compared the survival rates of patients in the PACA-CA cohort for 1 year (**Figure 3E**), 3 years (**Figure 3F**), 5 years (**Figure 3G**) and all patients (12 years) (**Figure 3H**) in detail, and found There are no significant differences between

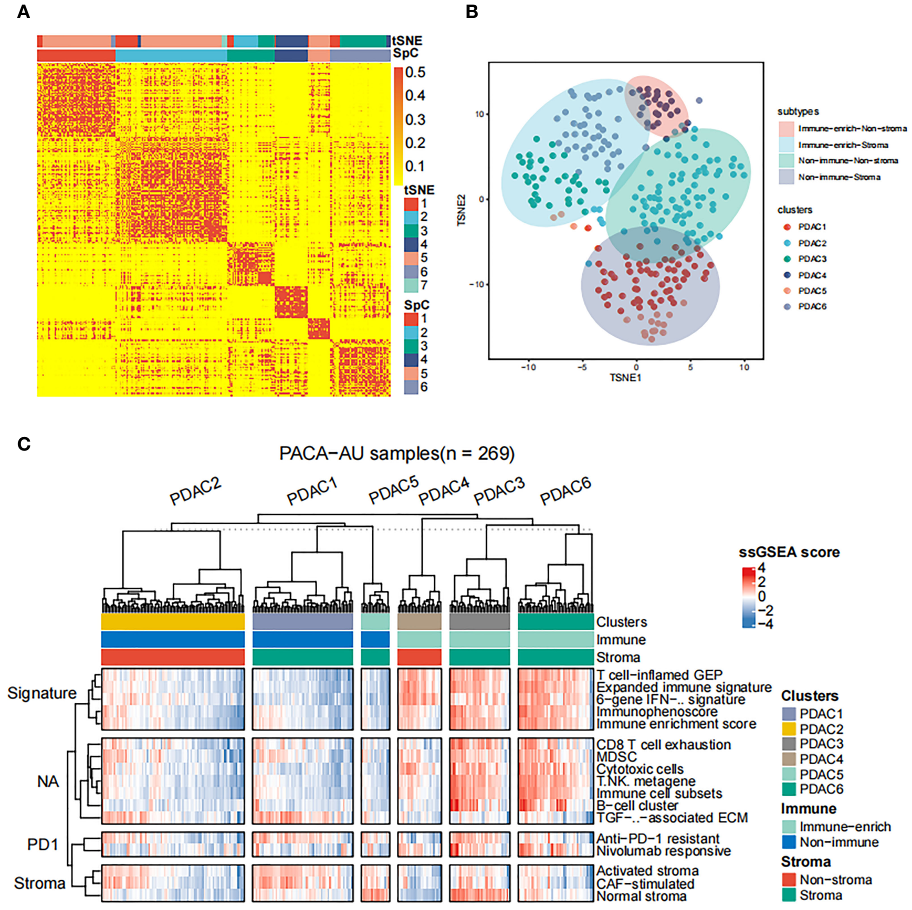


FIGURE 1 | Classification of distinct tumor microenvironment subtypes **(A)** Spectral classification of tumor microenvironment in PACA-AU alignment. This plot shows a heat map of the ssGSEA score, estimated using the gene set from the ICGC database. Based on tSNE cluster analysis, 7 subgroups were obtained, namely PDAC1, PDAC2, PDAC3, PDAC4, PDAC5, PDAC6, PDAC7. Based on Spectral classification, 6 subgroups were obtained, namely PDAC1, PDAC2, PDAC3, PDAC4, PDAC5, PDAC6. **(B)** tSNE classification of tumor microenvironment in PACA-AU cohort. **(C)** This figure shows the 4 immune subtypes of the PACA-AU cohort based on ssGSEA analysis and the main signatures of each subtype.

different subgroups. It is worth mentioning that there is a relatively large difference in the five-year survival rate between the immune-enrich-stroma and non-immune-stroma groups of PACA-CA (**Figure 3I**). The analysis results of the TCGA-PAAD cohort showed that the survival rates of patients in different subgroups were 1 year (**Figure 3J**), 3 years (**Figure 3K**), 5 years (**Figure 3L**) and all patients (8 years) (**Figure 3M**). It was found that there were no significant differences between the different subgroups. However, the one-year survival rate difference between immune-enrich-stroma and immune-enrich-non-stroma groups is relatively large (**Figure 3N**). In general, the classification of the PACA-AU cohort can provide an important reference for their clinical survival prognosis.

Prognostic Prediction Model Based on Signatures of Tumor Microenvironment

Since the subtype classification in the PACA-AU cohort has a strong correlation with survival prognosis, we use PACA-AU data as training data, and PACA-CA and TCGA-PAAD as test data to construct a prognostic prediction model. Firstly, PACA-AU data is treated as training data for parameter training of prediction models and selection of related gene sets. PACA-CA and TCGA-PAAD are regarded as testing data to test the parameters given by the training set and the predictive ability of the gene set. Then, use the cox regression algorithm to initially screen the genes that are significantly related to the patient's overall survival (P<0.05), and use the LASSO algorithm to further screen these genes. In the end, the best gene panel is obtained, and the forest diagram of the multivariate COX regression model is drawn (**Figure 4A**). In detail, those genes are KRT6C, PRR11, LTC4S, FGG, SERPINB3, CACNA2D3, FLT3LG, FDCSP, C5ORF46, FAM107A, CCL19, BLK, SLAMF1 and their multiple regression coefficients are 0.58, 0.89, -0.68, 0.69, 0.27, -0.56, -0.83, -0.54, 0.73, 0.97, -0.42, 0.62, 0.78. Subsequently, based on the expression level and multiple regression coefficients of gene panel obtained above, calculate their risk score. We further divided patients into high-risk groups and low-risk groups based on the risk index of the sample. Kaplan-Meier survival analysis was performed and showed in survival curve. There is a

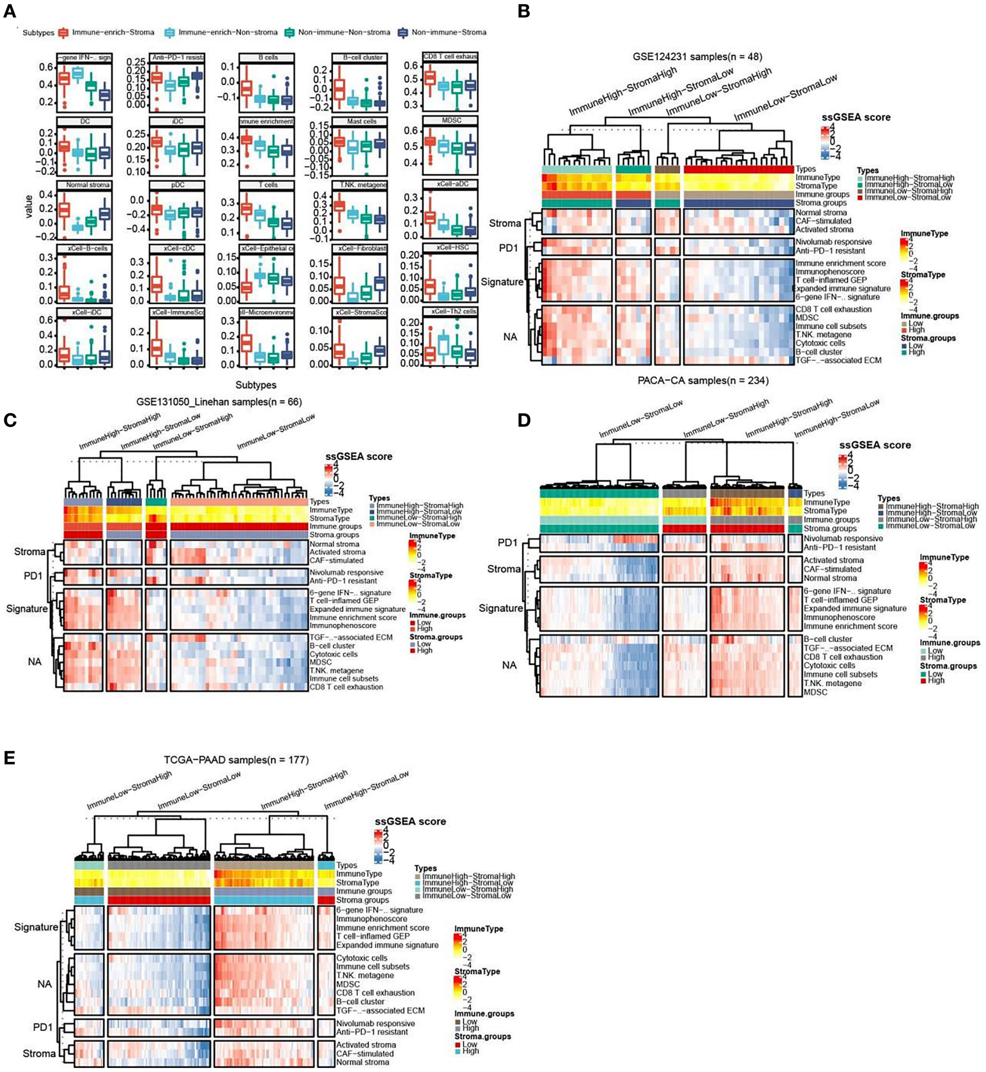


FIGURE 2 | Comparison of the striking differences in the immune microenvironment of the four subtypes. **(A)** Comparison of the striking differences in the immune microenvironment of the four subtypes. Red represents immune-enrich-stroma subtype, Light_blue represents immune-enrich-non-stroma subtype, Green represents non-immune-non-stroma subtype, and Navy blue represents non-immune-stroma subtype. **(B)** Immune-enrich-Stroma, Immune-enrich-non-Stroma, Non-immune-Stroma and Non-immune-non-Stroma types in the GSE124231 data set (n=48). **(C)** Four types in the GSE131050 data set (n=66). **(D)** Four types in the PACA-CA data set (n=234). **(E)** Four types in the TCGA-PAAD database (n = 177).

significant difference in survival probability between the high-risk group and the low-risk group in PACA-AU cohort (p <0.05) (**Figure 4B**). At the same time, we drew the ROC curve of the one-year, three-year, and five-year survival period of the patients in the training set based on the risk index (**Figure 4C**). However, there was no significant difference in the survival probability between the high-risk group and the low-risk group in the TCGA-PAAD testing set.

At the same time, we drew the ROC curve of patient survival in PACA-CA (**Figure 4E**) cohorts based on the risk index. The ROC curve of the prediction model of the PACA-CA training set shows that the prediction model is relatively ideal, and the prediction of the 1-year survival period is slightly better than the 3-year and 5-year survival periods. In addition, the prediction effect of the PACA-CA testing set (**Figure 4D**) is slightly inferior to that of the PACA-AU training set, except for the 5-year survival period of the TCGA-PAAD testing set. Overall, the prognosis prediction model can better predict the grouping of patients based on the risk index, which provides guidance for the prognosis prediction of patients.

Immune Cells Related to Risk Index

Based on the previous results, we want to know which immune cells are specifically related to the risk index of the PACA-AU cohort. Therefore, we used the sample risk index to make further correlation analysis with the expression of various immune cells and immune molecules. The results showed that the patient's risk

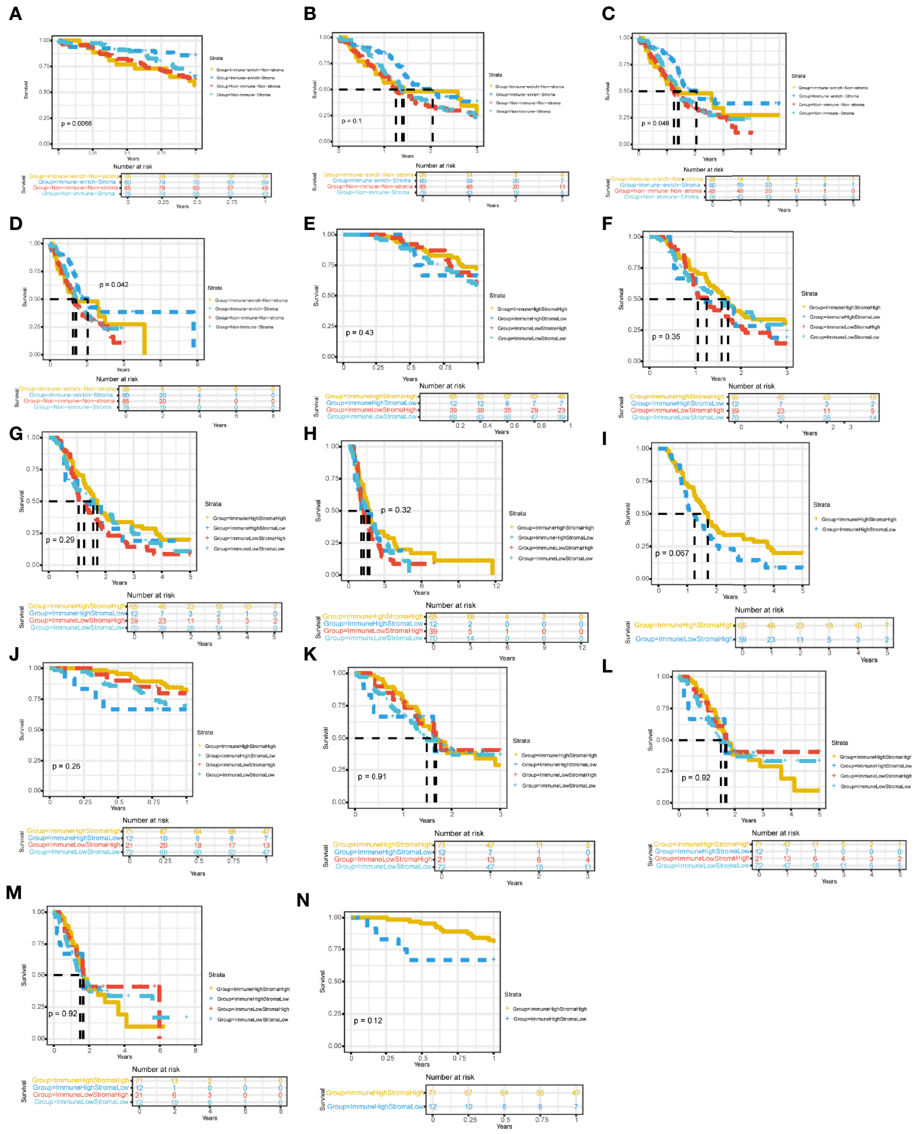


FIGURE 3 | Four immune subtypes are related to clinical characteristics and survival prognosis Comparison of survival rates between subgroups in different cohorts **(A)** Comparison of 1-year survival rate of PACA-AU cohort. **(B)** Comparison of 3-year survival rate of PACA-AU cohort. **(C)** Comparison of 5-year survival rate of PACA-AU cohort. **(D)** Comparison of survival rates of all PACA-AU cohort. **(E)** Comparison of 1-year survival rate of PACA-CA cohort. **(F)** Comparison of 3-year survival rate of PACA-CA cohort. **(G)** Comparison of 5-year survival rate of PACA-CA cohort. **(H)** Comparison of survival rates of all PACA-CA cohort. **(I)** Comparison of survival rates of the Immune-enrich-Stroma and Non-immune-Stromasubtypes. **(J)** Comparison of 1-year survival rate of TCGA-PAAD cohort. **(K)** Comparison of 3-year survival rate of TCGA-PAAD cohort. **(L)** Comparison of 5-year survival rate of TCGA-PAAD cohort. **(M)** Comparison of survival rates of all TCGA-PAAD cohort. **(N)** Comparison of survival rates of the Immune-enrich-Stroma and Immune-enrich-non-Stroma subtypes.

index and epithelial cells, megakaryocyte-erythroid progenitor (MEP), and Th2 cells showed a positive correlation with p<0.01. In addition, T cells, NK cells, memory B-cells, mast cells and other immune cells have a negative correlation with p <0.01 (**Figure 5**).

DISCUSSION

In this study, we used the ssGSEA algorithm to calculate the ssGSEA scores of PACA-AU pancreatic cancer patients, and then combined the Spectral clustering algorithm to extract the 4 subtypes in the cohort. We further compared the differences in the immune microenvironment of the four subtypes, and screened the immune enrichment and stromal enrichment molecular markers. Genes with significant differences are mostly related to immunity in (**Table S1**). For example, changes in the expression of PTPRCAP affect the survival rate of cancer patients (45), and the single nucleotide polymorphism (SNP) of PTPRCAP is associated with the susceptibility of gastric cancer (46). Natural killer cell granule protein 7 (NKG7) is

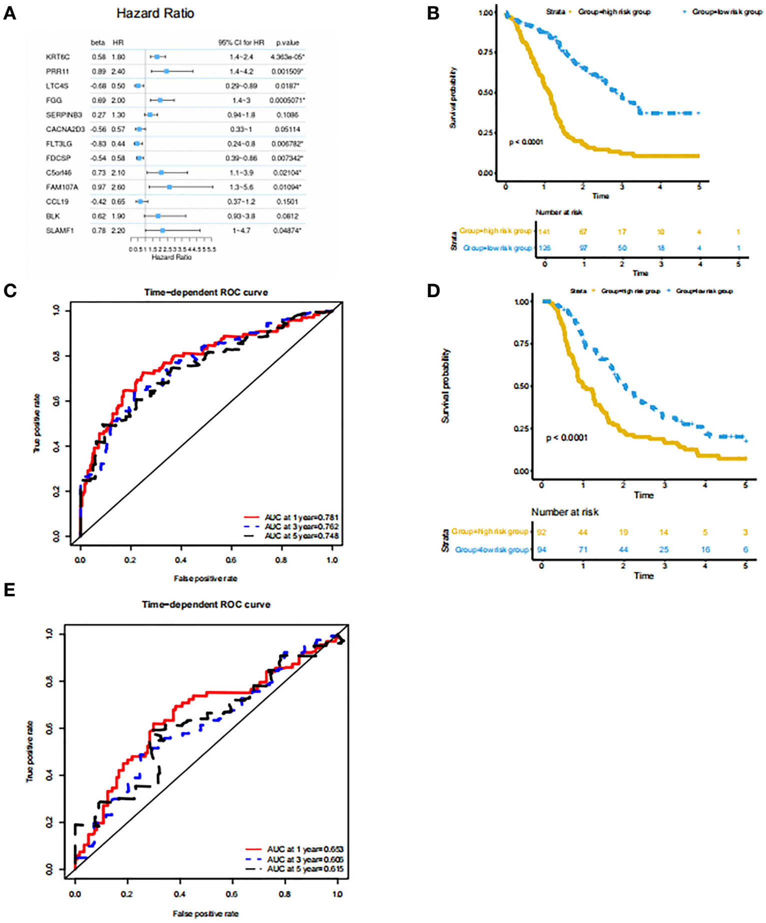


FIGURE 4 | Prognostic prediction model based on signatures of tumor microenvironment **(A)** features: significant factor name; multi_beta: Cox multiple regression coefficient; multi_HR: Cox multiple regression risk ratio; multi 95% CI for HR: Cox multiple regression risk ratio 95% confidence interval; Forest diagram: horizontal line shows the confidence interval interval, and the dot represents the hazard ratio; multi_p.value: Cox multiple regression proportional hazard hypothesis test P value. **(B)** Survival curve of the high and low risk groups in the training set. The horizontal axis represents time (unit: day), the vertical axis represents survival rate. A flat curve represents a high survival rate or a longer survival period, and a steep curve represents a low survival rate or a shorter survival period. **(C)** ROC curve of the training set prediction model. The horizontal axis is the false positive rate FP, and the vertical axis is the true positive rate TP. The legend in the upper left corner corresponds to the AUC value of the ROC curve for different survival periods. **(D)** Survival curves of the high- and low-risk groups in the PACA-CA testing set. **(E)** ROC curve of PACA-CA test set prediction model. The horizontal axis is the false positive rate FP, and the vertical axis is the true positive rate TP. The legend in the upper left corner corresponds to the AUC value of the ROC curve for different survival periods. The horizontal axis is the false positive rate (FP), and the vertical axis is the true positive rate (TP). The legend in the upper left corner corresponds to the AUC value of the ROC curve for different survival periods.

related to inflammatory diseases (47), and its lack will result in a significant reduction in IFN-γ produced by T cells and NK cells. In addition, NKG7 is related to the cytotoxic degranulation of CD8+ T cells (48). Researchers have discovered that CD96 can serve as a new immune checkpoint receptor target for T cells and natural killer cells (49). Similarly, we observed the top genes and their stromal functions in (**Table S2**). For example, Slits3 is expressed in primary bone marrow stromal and bone marrow-derived endothelial cells and stromal cell lines, and plays a role in *in vitro* migration and *in vivo* homing of hematopoietic stem and progenitor cells (50). SPARC is a stromal cell protein, which can be produced by cells associated with tumor stromal cells and has high expression levels in many cancers. It plays an important role in the fibroproliferative reaction of tumors (51).

Using the same research method, it was verified in the GSE124231 (n=48), GSE131050_Linahan (n=66), PACA-CA (n=234), TCGA-PAAD (n=177) cohorts, and the typing was accurate in different cohorts. Further compare the clinical information of patients in the cohort, and in-depth exploration of the difference in survival of patients with different subgroups. We found that in the PACA-AU cohort, the 1-year, 5-year, and 8-year survival times of different subsets patients were significantly correlated. Next, cox regression combined with Lasso algorithm was performed to construct a multivariate COX model. Calculate the patient's risk index based on gene expression level and multiple regression coefficients, and divide the patients into high-risk groups and low-risk groups based on the risk index. Interestingly, the PACA_AU and PACA-CA risk

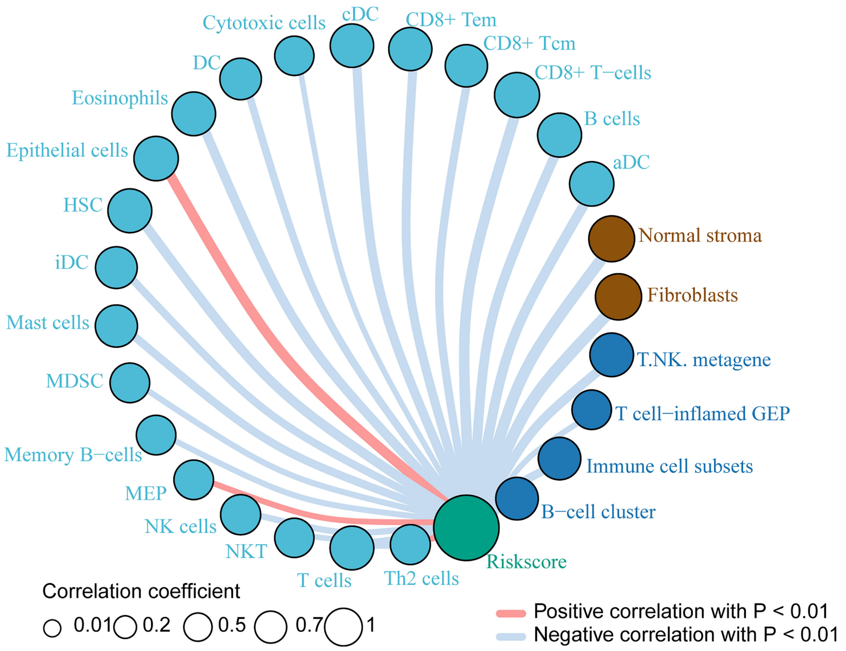


FIGURE 5 | Immune cells related to risk index Immune cells associated with the risk index of PACA-AU patients. The red line indicates a positive correlation between the risk index and immune cells, and the gray line indicates a negative correlation between the risk index and immune cells. The size of the circle indicates different correlation coefficients, and the larger the area of the circle, the larger the correlation coefficient.

indexes are significantly correlated with the survival level of patients.

In PADA-AC and TCGA-PAAD, the survival time difference between different immune subgroups is not significant. Only the five-year survival of immune-enrich-stroma and non-immune-stroma group in PACA-CA cohort and the one-year survival of immune-enrich-Stroma and immune-enrich-non-Stroma group in TCGA-PAAD cohort are relatively large. On the one hand, the cohort clustering algorithm may not cover all patients in the cohort, on the other hand, it may also be because the cohort samples are not large enough, and the representativeness of the statistical results needs to be further improved.

We initially explored the types of immune cells related to the risk index, and we identified immune cells that are positively and negatively related to the risk index. This research lays the foundation for the subsequent in-depth exploration of the correlation mechanism between immune cells and patient disease risk. However, only analyzing the types of immune cells is insufficient for the study of the mechanism. In the later stage, we will conduct more in-depth analysis and verification of important immune cells and their molecular signatures.

METHODS

Project and Sample

Dataset of 461 PACA-AU donors were downloaded from ICGC database (https://dcc.icgc.org/projects/PACA-AU) with detailed clinical information. The independent datasets used for verification come from GSE124231, GSE131050_Linehan, PACA-CA and TCGA-PAAD projects, including 48, 66, 234 and 177 donors respectively. Moreover, patients in the PACA-CA and TCGA-PAAD cohorts had detailed clinical information.

Bioinformatics Analysis

1) ssGSEA algorithm: Use the R package "GSVA" and use ssGSEA to explore the PACA-AU pancreatic cancer expression profile data of the ICGC database, and analyze the immune enrichment of each patient's tumor microenvironment. Additionally, the gene expression of all samples were took as the input and ssGSEA algorithms were occupied to determine the proportion of the various immune cells of all PDAC samples. The immune gene signatures were listed in the **Table 1**. According to the immune enrichment status of PACA-AU samples, they are divided into immune cells and stromal cell enriched (immune-enrich-stroma), non-immune enrichment but stromal cell enrichment (non-immune-stroma), and immune-enriched but Non-matrix enrichment (immune-enrich-non-stroma) and non-immune enrichment and non-stromal cell enrichment (non-immune-non-stroma). According to the ssGSEA score obtained by each sample, the Spectral clustering algorithm is used to extract different classifications. In addition, the R package "limma" was used to analyze immuno-enriched and non-immune-enriched patients, as well as the significantly different genes of stromal cell enrichment and non-matrix enrichment, and P<0.05 was taken as the significant difference.

2) The unsupervised clustering of the data set was performed mainly based on tSNE which embedded in t-distributed random

neighborhoods (45). In this study, we use tSNE to show the different subgroups of the PACA-AU cohort.

3) We performed Kaplan-Meier survival analysis on the samples and plotted survival curves. Survival analysis divided the samples into high-index groups and low-index groups based on the median. Data visualization is mainly done in the R environment (version 4.1.0). Kaplan-Meier survival analysis relies on the use of the "survival" package. The ROC curve is drawn based on the'survivalROC' package.

4) Prognosis prediction model establishment process: a). Use the training set to perform unit cox regression on each gene to initially screen disease-related genes; b). After obtaining all cox significant genes in all units, perform 1000X LASSO regression to calculate the frequency of each gene and rank it; c). According to the sorting result of the previous step, build the gene set incrementally. Use each gene set to perform multiple cox regression to get the contribution of each gene; d). Obtain the optimal gene set according to the gene contribution degree, and perform multiple cox regression analysis on these genes. Finally, we determined the regression coefficient of each gene; e). Calculate the death risk score of each patient through regression coefficients; f). The death risk score model is tested in the training set (comparing the predicted situation with the actual situation); g). The same model is tested in the independent testing set at the beginning (comparison of the predicted situation with the actual situation).

5) Construct the optimal multivariate COX model based on the Lasso algorithm. This analysis uses the LASSO algorithm for gene screening: In the field of statistics and machine learning, Lasso algorithm (least absolute shrinkage and selection operator, also translated as minimum absolute shrinkage and selection operator, lasso algorithm) is a regression analysis method that simultaneously performs feature selection and regularization (mathematics).It aims to enhance the predictive accuracy and interpretability of statistical models. Lasso adopts the linear regression method of L1-regularization, so that the weight of some learned features is 0, so as to achieve the purpose of sparseness, selection of variables, and construction of the best model. The characteristic of LASSO regression is to perform variable selection and regularization while fitting a generalized linear model. Therefore, regardless of whether the target dependent variable (dependent/response variable) is continuous, binary or discrete, it can be modeled by LASSO regression and then predicted.

6) We use the Lasso algorithm (glmnet package) to select the best gene model based on the COX multiple regression model, and finally draw the unit cox regression model forest diagram based on the gene Panel as follows: We calculate the risk score (Risk Score) of each patient based on the expression of the gene Panel and the multiple regression coefficient. The formula is as follows:

$$\mathrm{Riskscore} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \beta \mathrm{i} * \mathrm{xi}$$

xi represents the expression level of each gene in the Panel, βi is the multivariate COX regression beta value (multi_beta) corresponding to each gene.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

FM and XW conceived this project. XW, LL, LF, and YM collected the data. YY and FM analyzed and interpreted the data. XW and FM performed the statistical analyses and wrote the manuscript. All authors have reviewed the manuscript and approved the final version.

FUNDING

This work was financially supported by National Natural Science Foundation of China (82002480).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.832715/full#supplementary-material

REFERENCES

1. Ansari D, Tingstedt B, Andersson B, Holmquist F, Sturesson C, Williamsson C, et al. Pancreatic Cancer: Yesterday, Today and Tomorrow. *Future Oncol* (2016) 12(16):1929–46. doi: 10.2217/fon-2016-0010
2. Goral V. Pancreatic Cancer: Pathogenesis and Diagnosis. *Asian Pac J Cancer Prev* (2015) 16(14):5619–24. doi: 10.7314/APJCP.2015.16.14.5619
3. Gupta R, Amanam I, Chung V. Current and Future Therapies for Advanced Pancreatic Cancer. *J Surg Oncol* (2017) 116(1):25–34. doi: 10.1002/jso.24623
4. Hidalgo M, Cascinu S, Kleeff J, Labianca R, Lohr JM, Neoptolemos J, et al. Addressing the Challenges of Pancreatic Cancer: Future Directions for Improving Outcomes. *Pancreatology* (2015) 15(1):8–18. doi: 10.1016/j.pan.2014.10.001
5. Schlesinger Y, Yosefov-Levi O, Kolodkin-Gal D, Granit RZ, Peters L, Kalifa R, et al. Single-Cell Transcriptomes of Pancreatic Preinvasive Lesions and Cancer Reveal Acinar Metaplastic Cells' Heterogeneity. *Nat Commun* (2020) 11(1):4516. doi: 10.1038/s41467-020-18207-z
6. Dąbkowski K, Bogacka B, Tarnowski M, Starzyńska T. Pancreatic Cancer Microenvironment. *Pol Merkur Lekarski* (2016) 41((246):296–302.
7. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade. *Science* (2017) 357(6349):409–13. doi: 10.1126/science.aan6733
8. Knudsen ES, Vail P, Balaji U, Ngo H, Botros IW, Makarov V, et al. Stratification of Pancreatic Ductal Adenocarcinoma: Combinatorial Genetic, Stromal, and Immunologic Markers. *Clin Cancer Res* (2017) 23 (15):4429–40. doi: 10.1158/1078-0432.CCR-17-0162

9. Dreyer SB, Chang DK, Bailey P, Biankin AV. Pancreatic Cancer Genomes: Implications for Clinical Management and Therapeutic Development. *Clin Cancer Res* (2017) 23(7):1638–46. doi: 10.1158/1078-0432.CCR-16-2411
10. Ren B, Cui M, Yang G, Wang H, Feng M, You L, et al. Tumor Microenvironment Participates in Metastasis of Pancreatic Cancer. *Mol Cancer* (2018) 17(1):108. doi: 10.1186/s12943-018-0858-1
11. Lin QJ, Yang F, Jin C, Fu DL. Current Status and Progress of Pancreatic Cancer in China. *World J Gastroenterol* (2015) 21(26):7988–8003. doi: 10.3748/wjg.v21.i26.7988
12. Chronopoulos A, Robinson B, Sarper M, Cortes E, Auernheimer V, Lachowski D, et al. ATRA Mechanically Reprograms Pancreatic Stellate Cells to Suppress Matrix Remodelling and Inhibit Cancer Cell Invasion. *Nat Commun* (2016) 7:12630. doi: 10.1038/ncomms12630
13. Sunami Y, Kleeff J. Immunotherapy of Pancreatic Cancer. *Prog Mol Biol Transl Sci* (2019) 164:189–216. doi: 10.1016/bs.pmbts.2019.03.006
14. Yamazaki K, Masugi Y, Effendi K, Tsujikawa H, Hiraoka N, Kitago M, et al. Upregulated SMAD3 Promotes Epithelial-Mesenchymal Transition and Predicts Poor Prognosis in Pancreatic Ductal Adenocarcinoma. *Lab Invest* (2014) 94(6):683–91. doi: 10.1038/labinvest.2014.53
15. Sivaram N, McLaughlin PA, Han HV, Petrenko O, Jiang YP, Ballou LM, et al. Tumor-Intrinsic PIK3CA Represses Tumor Immunogenecity in a Model of Pancreatic Cancer. *J Clin Invest* (2019) 129(8):3264–76. doi: 10.1172/JCI123540
16. Siemers NO, Holloway JL, Chang H, Chasalow SD, Ross-MacDonald PB, Voliva CF, et al. Genome-Wide Association Analysis Identifies Genetic Correlates of Immune Infiltrates in Solid Tumors. *PloS One* (2017) 12(7): e0179726. doi: 10.1371/journal.pone.0179726
17. Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer. *Nature* (2016) 531(7592):47–52. doi: 10.1038/nature16965
18. Hashimoto S, Furukawa S, Hashimoto A, Tsutaho A, Fukao A, Sakamura Y, et al. *ARF6 and AMAP1 are Major Targets of KRAS and TP53 Mutations to Promote Invasion, PD-L1 Dynamics, and Immune Evasion of Pancreatic Cancer. Proc Natl Acad Sci USA* (2019) 116(35):17450–9. doi: 10.1073/pnas.1901765116
19. Mizrahi JD, Surana R, Valle JW, Shroff RT. Pancreatic Cancer. *Lancet* (2020) 395((10242):2008–20. doi: 10.1016/S0140-6736(20)30974-0
20. Fabris L, Perugorria MJ, Mertens J, Bjorkstrom NK, Cramer T, Lleo A, et al. The Tumour Microenvironment and Immune Milieu of Cholangiocarcinoma. *Liver Int* (2019) 39(Suppl 1):63–78. doi: 10.1111/liv.14098
21. Wartenberg M, Zlobec I, Perren A, Koelzer VH, Gloor B, Lugli A, et al. Accumulation of FOXP3+T-Cells in the Tumor Microenvironment Is Associated With an Epithelial-Mesenchymal-Transition-Type Tumor Budding Phenotype and Is an Independent Prognostic Factor in Surgically Resected Pancreatic Ductal Adenocarcinoma. *Oncotarget* (2015) 6(6):4190–201. doi: 10.18632/oncotarget.2775
22. Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of Pancreatic Ductal Adenocarcinoma and Their Differing Responses to Therapy. *Nat Med* (2011) 17(4):500–3. doi: 10.1038/nm.2344
23. O'Kane GM, Grunwald BT, Jang GH, Masoomian M, Picardo S, Grant RC, et al. GATA6 Expression Distinguishes Classical and Basal-Like Subtypes in Advanced Pancreatic Cancer. *Clin Cancer Res* (2020) 26(18):4901–10. doi: 10.1158/1078-0432.CCR-19-3724
24. Puleo F, Nicolle R, Blum Y, Cros J, Marisa L, Demetter P, et al. Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology* (2018) 155(6):1999–2013 e3. doi: 10.1053/j.gastro.2018.08.033
25. Cancer Genome Atlas Research Network and Electronic address: andrew_aguirre@dfci.harvard.edu; Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* (2017) 32(2):185–203.e13. doi: 10.1016/j.ccell.2017.07.007
26. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1. *Nature* (2009) 462((7269):108–12. doi: 10.1038/nature08460
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. *Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proc Natl Acad Sci USA* (2005) 102(43):15545–50. doi: 10.1073/pnas.0506580102
28. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. *Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data*. 2013: Nat Commun. doi: 10.1038/ncomms3612
29. Chow LQM, Haddad R, Gupta S, Mahipal A, Mehra R, Tahara M, et al. Antitumor Activity of Pembrolizumab in Biomarker-Unselected Patients With Recurrent and/or Metastatic Head and Neck Squamous Cell Carcinoma: Results From the Phase Ib KEYNOTE-012 Expansion Cohort. *J Clin Oncol* (2016) 34(32):3838–45. doi: 10.1200/JCO.2016.68.1478
30. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SG, Hoadley KA, et al. Virtual Microdissection Identifies Distinct Tumor- and Stroma-Specific Subtypes of Pancreatic Ductal Adenocarcinoma. *Nat Genet* (2015) 47 (10):1168–78. doi: 10.1038/ng.3398
31. Cancer Genome Atlas Research Network. *The Molecular Taxonomy of Primary Prostate Cancer*. *Cell* (2015) 163(4):1011–25. doi: 10.1016/j.cell.2015.10.025
32. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* (2013) 39(4):782–95. doi: 10.1016/j.immuni.2013.10.003
33. Alistar A, Chou JW, Nagalla S, Black MA, D'Agostino RJr., Miller LD. Dual Roles for Immune Metagenes in Breast Cancer Prognosis and Therapy Prediction. *Genome Med* (2014) 6(10):80. doi: 10.1186/s13073-014-0080-8
34. Iglesia MD, Vincent BG, Parker JS, Hoadley KA, Carey LA, Perou CM, et al. Prognostic B-Cell Signatures Using Mrna-Seq in Patients With Subtype-Specific Breast and Ovarian Cancer. *Clin Cancer Res* (2014) 20(14):3818–29. doi: 10.1158/1078-0432.CCR-13-3368
35. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* (2017) 18(1):248–62. doi: 10.1016/j.celrep.2016.12.019
36. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-Tumor Genomic Biomarkers for PD-1 Checkpoint Blockade-Based Immunotherapy. *Science* (2018) 362(6411):eaar3593. doi: 10.1126/science.aar3593
37. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-Gamma-Related Mrna Profile Predicts Clinical Response to PD-1 Blockade. *J Clin Invest* (2017) 127(8):2930–40. doi: 10.1172/JCI91190
38. Chakravarthy A, Khan L, Bensler NP, Bose P, De Carvalho DD. TGF-Beta-Associated Extracellular Matrix Genes Link Cancer-Associated Fibroblasts to Immune Evasion and Immunotherapy Failure. *Nat Commun* (2018) 9 (1):4692. doi: 10.1038/s41467-018-06654-8
39. Yaddanapudi K, Rendon BE, Lamont G, Kim EJ, Al Rayyan N, Richie J, et al. MIF is Necessary for Late-Stage Melanoma Patient MDSC Immune Suppression and Differentiation. *Cancer Immunol Res* (2016) 4(2):101–12. doi: 10.1158/2326-6066.CIR-15-0070-T
40. Calon A, Espinet E, Palomo-Ponce S, Tauriello DV, Iglesias M, Cespedes MV, et al. Dependency of Colorectal Cancer on a TGF-Beta-Driven Program in Stromal Cells for Metastasis Initiation. *Cancer Cell* (2012) 22(5):571–84. doi: 10.1016/j.ccr.2012.08.013
41. Beyer M, Mallmann MR, Xue J, Staratschek-Jox A, Vorholt D, Krebs W, et al. High-Resolution Transcriptome of Human Macrophages. *PloS One* (2012) 7 (9):e45466. doi: 10.1371/journal.pone.0045466
42. Giordano M, Henin C, Maurizio J, Imbratta C, Bourdely P, Buferne M, et al. Molecular Profiling of CD8 T Cells in Autochthonous Melanoma Identifies Maf as Driver of Exhaustion. *EMBO J* (2015) 34(15):2042–58. doi: 10.15252/embj.201490786
43. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, et al. Chromatin States Define Tumour-Specific T Cell Dysfunction and Reprogramming. *Nature* (2017) 545(7655):452–6. doi: 10.1038/nature22367
44. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and Microenvironment Evolution During Immunotherapy With Nivolumab. *Cell* (2017) 171(4):934–49. doi: 10.1016/j.cell.2017.09.028
45. van der Maaten L. *Visualizing Data Using T-SNE. J Mach Learn Res* (2008) 1:1–48.
46. Ju H, Lim B Fau - Kim M, Kim M Fau - Kim YS, Kim Ys Fau - Kim WH, Kim Wh Fau - Ihm C, Ihm C Fau - Noh S-M, et al. A Regulatory Polymorphism at

Position -309 in PTPRCAP is Associated With Susceptibility to Diffuse-Type Gastric Cancer and Gene Expression. *Neoplasia* (2009) 11(12):1340–7. doi: 10.1593/neo.91132

47. van Es LA, de Heer E, Vleming LJ, van der Wal A, Mallat M, Bajema I, et al. GMP-17-Positive T-Lymphocytes in Renal Tubules Predict Progression in Early Stages of Iga Nephropathy. *Kidney Int* (2008) 73(12):1426–33. doi: 10.1038/ki.2008.66
48. Ng SS, De Labastida Rivera F, Yan J, Corvino D, Das I, Zhang P, et al. The NK Cell Granule Protein NKG7 Regulates Cytotoxic Granule Exocytosis and Inflammation. *Nat Immunol* (2020) 21(10):1205–18. doi: 10.1038/s41590-020-0758-6
49. Dougall WC, Kurtulus S, Smyth MJ, Anderson AC. TIGIT and CD96: New Checkpoint Receptor Targets for Cancer Immunotherapy. *Immunol Rev* (2017) 276(1):112–20. doi: 10.1111/imr.12518
50. Geutskens SB, Andrews WD, van Stalborch AM, Brussen K, Holtrop-de Haan SE, Parnavelas JG, et al. Control of Human Hematopoietic Stem/Progenitor Cell Migration by the Extracellular Matrix Protein Slit3. *Lab Invest* (2012) 92 (8):1129–39. doi: 10.1038/labinvest.2012.81
51. Framson PE, Sage EH. SPARC and Tumor Growth: Where the Seed Meets the Soil? *J Cell Biochem* (2004) 92(4):679–90. doi: 10.1002/jcb.20091

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.

The handling editor declared a shared parent affiliation with several of the authors LL, LF, YM, and XW at time of review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Oncology

ORIGINAL RESEARCH
published: 11 March 2022
doi: 10.3389/fonc.2022.814667

Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test

Hanqiang Ouyang[1,2,3†], Fanyu Meng[4,5†], Jianfang Liu[6†], Xinhang Song[4], Yuan Li[6], Yuan Yuan[6], Chunjie Wang[6], Ning Lang[6], Shuai Tian[6], Meiyi Yao[4,5], Xiaoguang Liu[1,2,3], Huishu Yuan[6], Shuqiang Jiang[4*] and Liang Jiang[1,2,3*]*

[1] Department of Orthopaedics, Peking University Third Hospital, Beijing, China, [2] Engineering Research Center of Bone and Joint Precision Medicine, Beijing, China, [3] Beijing Key Laboratory of Spinal Disease Research, Beijing, China, [4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, [5] University of Chinese Academy of Sciences, Beijing, China, [6] Department of Radiology, Peking University Third Hospital, Beijing, China

Background: Recently, the Turing test has been used to investigate whether machines have intelligence similar to humans. Our study aimed to assess the ability of an artificial intelligence (AI) system for spine tumor detection using the Turing test.

Methods: Our retrospective study data included 12179 images from 321 patients for developing AI detection systems and 6635 images from 187 patients for the Turing test. We utilized a deep learning-based tumor detection system with Faster R-CNN architecture, which generates region proposals by Region Proposal Network in the first stage and corrects the position and the size of the bounding box of the lesion area in the second stage. Each choice question featured four bounding boxes enclosing an identical tumor. Three were detected by the proposed deep learning model, whereas the other was annotated by a doctor; the results were shown to six doctors as respondents. If the respondent did not correctly identify the image annotated by a human, his answer was considered a misclassification. If all misclassification rates were >30%, the respondents were considered unable to distinguish the AI-detected tumor from the human-annotated one, which indicated that the AI system passed the Turing test.

Results: The average misclassification rates in the Turing test were 51.2% (95% CI: 45.7%–57.5%) in the axial view (maximum of 62%, minimum of 44%) and 44.5% (95% CI: 38.2%–51.8%) in the sagittal view (maximum of 59%, minimum of 36%). The misclassification rates of all six respondents were >30%; therefore, our AI system passed the Turing test.

Conclusion: Our proposed intelligent spine tumor detection system has a similar detection ability to annotation doctors and may be an efficient tool to assist radiologists or orthopedists in primary spine tumor detection.

Keywords: spine tumor, Turing test, deep learning, MRI, primary tumor

Edited by:
*Abhishek Mahajan,
Tata Memorial Hospital, India*

Reviewed by:
*Mirza Pojskic,
University Hospital of Giessen and Marburg, Germany
Khandkar Ali Kawsar,
Edinburgh Royal Infirmary, United Kingdom*

****Correspondence:***
*Huishu Yuan
huishuy@bjmu.edu.cn
Shuqiang Jiang
sqjiang@ict.ac.cn
Liang Jiang
jiangliang@bjmu.edu.cn*

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *14 November 2021*
Accepted: *16 February 2022*
Published: *11 March 2022*

Citation:
Ouyang H, Meng F, Liu J, Song X, Li Y, Yuan Y, Wang C, Lang N, Tian S, Yao M, Liu X, Yuan H, Jiang S and Jiang L (2022) Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test. Front. Oncol. 12:814667. doi: 10.3389/fonc.2022.814667

INTRODUCTION

Magnetic resonance imaging (MRI) is commonly used to diagnose spine disorders (e.g., myelopathy, spine canal stenosis, and traumatic injury). Spine tumors may cause spine fractures, instability, neurological deficits, or even paralysis. However, they are rarely observed because of their low incidence. Thus, it is difficult for junior radiologists or orthopedists to accumulate diagnostic experience, and they may not be capable of detecting different spine tumors on MRI. Deep learning (DL)—a class of artificial intelligence (AI)—is now prevalent in computer vision tasks. For spine imaging, especially MRI, DL, and other AI systems are being applied as diagnostic imaging technologies (1–5). Hallinan et al. (6) used a DL model for automated detection of the central canal, lateral recess, and neural foraminal stenosis in lumbar spine MRI; Huang et al. (7) utilized a DL-based fully automated program for vertebrae and disc quantifications on lumbar spine MRI; Merali et al. (8) developed a DL model for the detection of cervical spinal cord compression in MRI scans, and Ito et al. (9) developed the DL-based automated detection of spinal schwannomas in MRI. However, evaluation measures for AI methods are lacking because conventional radiology assessment systems do not meet the requirements of DL models. Thus, in this study, we applied the Turing test, a classical evaluation method in AI, on primary spine tumor DL detection on MR images.

Alan Turing, a British mathematician and theoretical computer scientist, is widely regarded as the founding father of AI. Alan Turing's paper in 1950 entitled *"Computing Machinery and Intelligence"* had considered the question "Can machines think?" (10). Subsequently, he replaced the question with a significantly more practical scenario, namely, the Turing imitation game. The game has now become widely known, particularly in the clinical domain, as the Turing test. The Turing test (11, 12) is proposed to assess if a machine can think like a human, which reframed his question as follows: Can a machine display intelligence *via* imitation? Although this proposal is complex, a common operation of the Turing test requires an interrogator to communicate electronically with a subject to judge whether the subject is a human or machine (13, 14). The machine performed well if the interrogator makes an incorrect identification as often as a correct one. When evaluating the automated detection ability of a DL model, the gold standard is a comparison with the manual annotation of the same images by radiologists or orthopedists. However, the use of manual spine tumor annotations as the gold standard has been questioned because annotations themselves are subjective. For example, when a patient's spine tumor is annotated by two different doctors, their annotations will hardly denote the same exact square, thereby reflecting inter- and intra-observer variability. Thus, the first purpose of using the Turing test was to confirm whether the automated detection ability of our DL models could achieve a clinically applicable standard compared with that of manual annotations at a tertiary university hospital. To this end, we proposed a simple interface program with choice questions to assess automated detection versus manual annotation based on position, shape, and area overlap. We hypothesized that if a clinical respondent is unable to distinguish the different bounding boxes drawn by an automated detection system and those produced manually by a spine expert, then it is likely that this DL model will be considered adequate for clinical application, which may assist orthopedists to find the primary spine tumors efficiently in the future. The second aim of this study was to assess the accuracy rate, false-positive, and false-negative results of the manual annotations from radiologists and orthopedists. It's important to note that in this study, we mostly focused on the primary tumors located in the skeletal spine structures, thus we did not collect the intradural or intramedullary nervous system tumors.

METHODS

Patient and Image Acquisition

We reviewed consecutive spinal tumor patients histologically diagnosed with primary spine tumors at our hospital between January 2012 and December 2020. Although primary spine tumors are rarely observed because of their low incidence, Peking University Third Hospital (PUTH) is a famous spine center in North China, and we can collect enough primary spine tumors patients in this study. The MR images of intradural or intramedullary nervous system tumors and ones acquired from other hospitals were excluded. Our database contained 508 patients, 226 women and 282 men (mean age, 49.0 [range, 3–84] years), including 19532 MR images with tumors. We used 12179 images from 321 patients to develop AI detection systems and 6635 images from 187 patients as a test set. For the Turing test, 100 patients were randomly selected from 187 patients in the test set. Sagittal and axial images were selected as representatives for manual annotation and training for the automated detection model because they span a wider range of spine regions, crucial for training the DL models for automated detection. Thus, the remaining 718 coronal images in the database did not participate in the training and testing process.

Preoperative MRI scans were performed on Discovery MR750 3.0T or Optima MR360 1.5T (GE Healthcare; Piscataway, NJ, USA). Conventional MRI scanning sequences included axial T2-weighted imaging (T2WI), sagittal T2WI, coronal T2WI, T1-weighted imaging (T1WI), and fat-suppressed T2WI scans. For axial and sagittal reconstruction, the scans were performed with the following parameters: field of view = 320 mm × 320 mm; matrix = 94 × 94; flip angle = 90; slice thickness = 3.0 mm; slice spacing = 3.3 mm; FS-T2WI turbo spin echo, repetition time (TR) = 2500–4000 msec, and echo time (TE) = 50–120 msec; and T1WI, TR = 400–800 msec, and TE = 10–30 msec.

Turing Test of Spine Tumors Detection

The study was approved by the PUTH Medical Science Research Ethics Committee review board, which waived the need for informed consent as this was a retrospective review of a previous prospective study.

In our case, the Turing test was carried out with a choice question, each choice question featured four similar MR images as candidates, and three of the candidates were results predicted by DL models, one of them was annotated by doctors. Among them, the results predicted by the DL model were obtained by a DL-based tumor detection system with Faster Region-Convolutional Neural Network (Faster R-CNN) (15) architecture in our study, and only one was manually annotated by one of the five annotation doctors A-E (four radiologists and one orthopedist). One hundred patients and 200 choice questions (one axial choice question and one sagittal choice question for each patient), were randomly selected from our database for the Turing test. Without knowing which were human annotations, every choice question was shown to six respondent doctors F-K (four radiologists and two orthopedists) to select which one (reasonable candidate) among the four MR images was annotated by the annotation doctor. Since the DL-based tumor AI detection system is designed to react similarly to human intelligence, we considered the doctor's lesion annotation as the correct option. Therefore, if the respondent did not correctly identify the image annotated by a human, his answer was considered a misclassification. The AI system passed the Turing test if the misclassification rates of the six respondents were all >30%. **Figure 1** shows the flow of the Turing test, which introduces the specific steps of the Turing test.

$$\text{Misclassification Rate} = \frac{F}{T + F}$$

where T represented the respondent correctly identifying the image annotated by a human, and F represented the respondent did not correctly identify the image annotated by a human.

Manual Annotation Database

Spine MRI data from Digital Imaging and Communications in Medicine files were exported in Joint Photographic Experts Group (JPEG) format from the picture archiving and communication systems of our hospital. These JPEG images were manually annotated using software *Labelme*, an image labeling tool developed in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. *Labelme* is capable of creating customized labeling tasks or performing image labeling; we annotated the images by manually inputting a minimal bounding box containing every tumor lesion on each sagittal or axial MRI slice to generate JPEG images for the automated detection training (**Figure 2**). Taken together, four radiologists and one orthopedist (doctors A–E) annotated 19532 MRI slices. To ensure that each tumor was recognized by the DL model under different conditions, all slices on T1W1 and T2W1 MR images were annotated.

Manual Annotation Assessment by Doctors

Before testing whether automated detection was sufficiently similar to manual annotation (namely, indistinguishable when judged by a blinded respondent), we randomly assessed the manual annotations to reduce inter- and intra-observer variability. The other three senior radiologists, except doctor F-K in our hospital, randomly and independently examined and verified the annotation images of doctors A–E. Based on the evaluation of the manual annotations, the computer engineers calculated the ultimate accuracy rate, false-positive rate, and false-negative rate of their labels by utilizing the confusion matrix. Clinical information of patients was not provided for any of the doctors to ensure a fair comparison between humans and DL models.

Architecture of Deep Learning-Based Automated Detection

In this study, we trained the automated DL detection model using the locations and bounding box labels of spine tumors as training data. The automated detection model was trained and validated using a computer equipped with a Quadro P6000 graphics processing unit (NVIDIA; Santa Clara, CA), a Xeon E5-2667 v4 3.2 GHz CPU (Intel; Santa Clara, CA), and 64 GB of RAM.

We used PyTorch, a suitable framework for DL, to train a neural network model applied to the spine tumor dataset of MR images. A two-stage DL system with Faster R-CNN (15) architecture was used as the training model and consisted of a

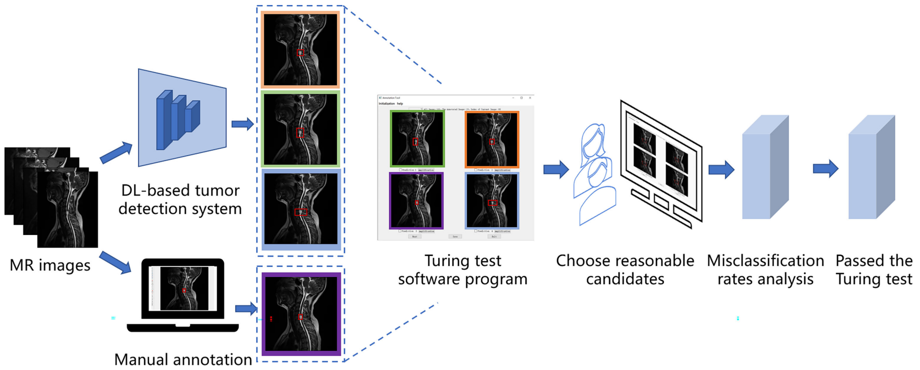

DL-based tumor detection system
MR images
Manual annotation
Turing test software program
Choose reasonable candidates
Misclassification rates analysis
Passed the Turing test

FIGURE 1 | The flow of the Turing test. This figure shows the specific steps of the Turing test.

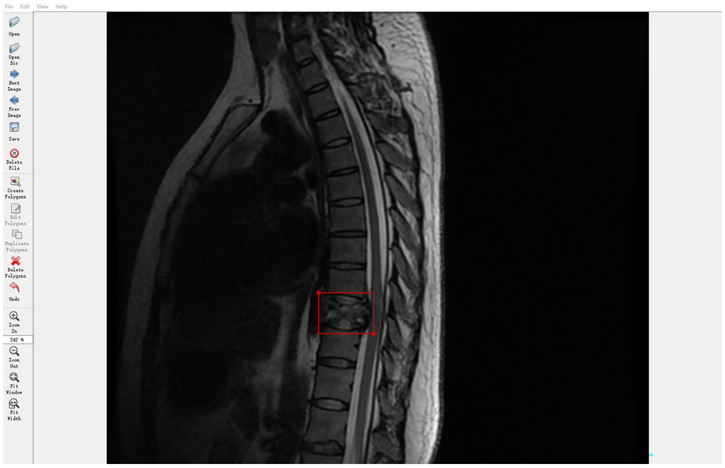


FIGURE 2 | Labeling tool used by doctors to annotate tumor coordinates: Labelme. The Labelme displays the currently annotated image. The red annotation box indicates that the current location is a tumor. The annotation tool will automatically generate the coordinates of the upper left point and the lower right point.

region proposal network (RPN) and region regression. The RPN was used to generate many anchors to get region proposals. It used SoftMax to recognize whether the anchors were positive or negative, the lesions are generally considered to be in positive anchors. Then the region regression could correct the positive anchors to obtain accurate proposals. Three different backbones of the proposed model were used to extract MR image feature maps, like ResNet-50 (16), ResNet-101 (16), and ResNet-152 (16), consisting of 50, 101, and 152 convolutional, pooling, and activation layers, respectively. These feature maps were shared for the RPN layer and region regression. And Feature Pyramid Networks (17) were also used in the model to solve the multi-scale problem in object detection.

The first-stage inputs were the MRI spine data of the three different backbones; the outputs were the different regions and activation maps, which were subsequently used as second-stage inputs. In the second stage, the region of interest (ROI) pooling layer collected the input feature maps and proposals, combining the information to extract proposal feature maps. Subsequently, a small network (i.e., multiple fully connected layers) was constructed with a regression branch to obtain the final precise positions of the lesion area. For efficient computing, all-region features were fed to the same regressor. Finally, we obtained the output of the three models. We call the Faster R-CNN framework with the backbone ResNet50, ResNet101, and ResNet152 as CNN1 (convolutional neural network 1), CNN2, and CNN3, respectively. **Figure 3** shows the automated detection framework of our Turing test.

Evaluation Measurement in Artificial Intelligence

The tumor detection performance was evaluated from the aspect of the class label and position accuracy, which could be measured

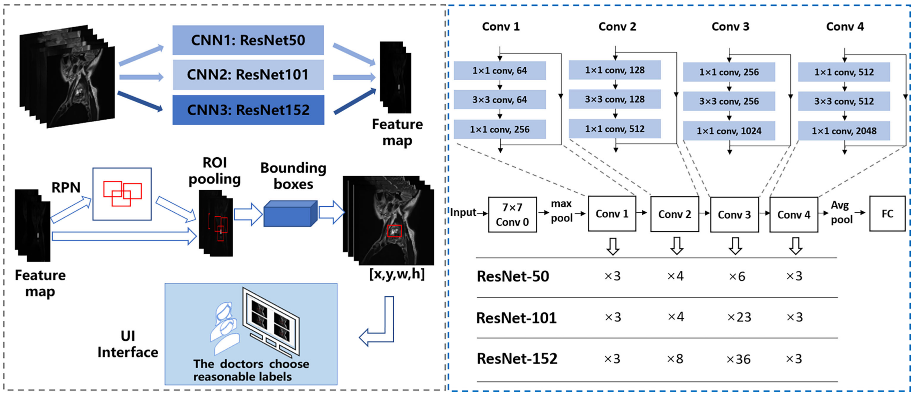


FIGURE 3 | The framework of automated detection of spine tumors utilizing the Turing Test. Faster RCNN is used as the framework, and ResNet50, ResNet101, and ResNet152 are used to extract image features respectively.

with average precision. Compared with the ground truth annotated by doctors, when the intersection of union (IoU) went over the threshold, the prediction was considered correct. The IoU formula is as follows:

$$IoU = \frac{b_{pred} \quad \cap \quad b_{gt}}{b_{pred} \quad \cup \quad b_{gt}}$$

where, b_{pred} and b_{gt} represent a bounding box of predictions and ground truth, respectively.

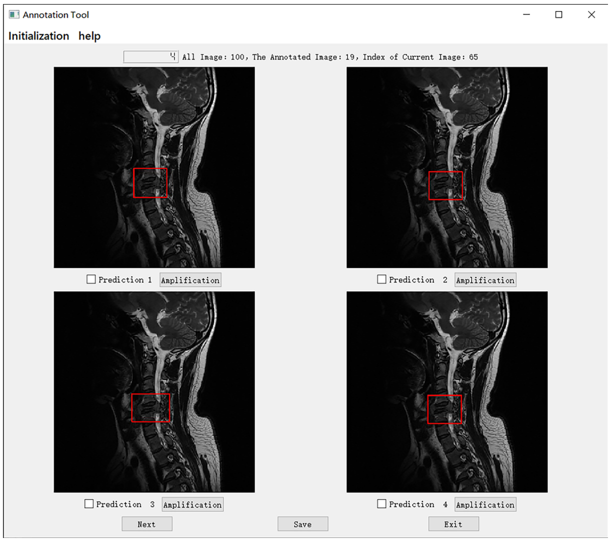
Training Implementation Details

In the training stage, the input images were divided into mini-batches. Each mini-batch contained eight images per GPU, and each image had 2000 region of interest samples with a ratio of one positive to three negatives. Specifically, anchors with an IoU >0.7 with the annotated bounding boxes) were set as positive examples; those with an IoU <0.3 were set as negative examples. The RPN anchors spanned five scales {16, 32, 64, 128, 256} and three aspect ratios {0.5, 0.8, 1.3}, totaling 15 anchors. The threshold of the non-maximum suppression layer was set to 0.5. We trained on one GPU with SGD for 10 epochs with a learning rate of 0.01, which was decreased by 0.5 every epoch. We used a weight decay of 0.0005 and momentum of 0.9. Due to the similarity between medical pictures, having more training pictures helps the DL model to better extract features, which can enhance the generalization of the model. Therefore, for better performance, axial and sagittal images were trained together for the MRI dataset. Similarly, T1W1 and T2W1 were trained together as a training set.

Turing Test Software Program

To complete the Turing test, the annotated images were reviewed by a team of 6 respondents, including two radiologists and one orthopedist who worked at our hospital for approximately 10 years, and other two radiologists and one orthopedist who worked there >20 years. The six respondents (doctors F–K) specialized in spine tumors and had not performed the annotation previously (doctors A–E). The six respondents' answering processes were double blindly designed to ensure no communication with any other people occurred.

We set up a Turing test software program with choice questions (**Figure 4**). In every choice question, the respondents were shown an interface with four MR images of an identical tumor; three featured bounding boxes were generated by DL models, whereas only one featured a bounding box drawn by an annotation doctor. The four images with correspondent bounding boxes featured were randomly ordered in each question. The respondents would be asked, "Which one is annotated by a human?" Each respondent reviewed approximately 200 choice questions (sagittal and axial figures) from 100 patients, randomly selected from a pool of 6635 annotated images of the test set. Figures of the interfaces were presented for assessment in questions only once owing to the random nature of the selection process. The display could be adjusted to a standard window, and a magnifying tool was provided to enable a detailed image inspection. Additionally, the software program documented respondent responses provided for each question and the time required to choose each respondent.



4 All Image: 100, The Annotated Image: 19, Index of Current Image: 65

FIGURE 4 | The choice interface of a Turing test software program. This software displays four options of one choice question, including amplification, timing, and technical functions. The user can click the next button to continue to the next question. After confirming the answer to the current question, click the save button to save the answer. After all the questions are completed, click the exit button to exit the program. Of these four options, prediction 3 is the result annotated by one of the doctors A-E, and prediction1, 2, and 4 is the result of the tumor location predicted by the model CNN 3, CNN1, and CNN2 respectively.

Specifically, in the selection of questions shown in **Figure 4**, prediction 3 is the result annotated by one of the doctors A-E, and predictions 1, 2, and 4 is the tumor location predicted by models CNN 3, CNN1, and CNN2, respectively. The network depth of the models CNN 1, CNN 2, and CNN3 differed. Compared with CNN 1, CNN 2 and CNN3 have a sequentially increasing number of network layers; the more the layers of the network mean the richer the abstract features of different levels that can be extracted. Moreover, the deeper the network, the more abstract the features, and the more semantic information.

Statistical and Data Analyses

All statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS, version 26.0; IBM Corporation, Armonk, N.Y., USA). Results were obtained for the fivefold cross-validation of object detection. The Mann–Whitney U test and chi-squared test were used for comparisons between groups for continuous and categorical variables, respectively. A P-value <0.05 was considered significant. The criteria of true detection and false detection were calculated for the DL-based automated tumor detection on MR images and the annotation team.

RESULTS

Patient Characteristics and Data Split

We obtained the MRI dataset of the primary spine tumors to train and evaluate our model. We trained together in these two views and tested them separately. Concerning primary spine tumors, 19532 images from 508 patients were included in the MRI dataset. We chose 12179 images from 321 patients to develop AI detection systems, including 7788 images (2346 axial; 5442 sagittal) randomly selected from 193 patients as the training set and 4391 images (1199 axial; 3192 sagittal) from 128 patients randomly selected as the validation set. The validation set was used to determine the network structure and help train a better model. Moreover, 6635 images (1835 axial; 4800 sagittal) from the other 187 patients were randomly selected as test set and Turing test data source. The dataset only contained axial and sagittal views, and the remaining 718 coronal images were not used for the training or the Turing test.

Evaluation Measurement Among Doctors

The five doctors annotated primary spine tumors on MR images; the total number of annotated images for each doctor was 4527 (A), 4159 (B), 3910 (C), 3727 (D), and 3209 (E). As **Figure 5** shows, there were a total of 26 tumor histological categories in our dataset, such as schwannoma, myeloma, and chordoma, among others. In the dataset, there were 3758 schwannoma images and only 25 ganglion neurofibroma images. The evaluations of the five annotation doctors are listed in **Figures 6**, **7** and include detailed accuracy rate, false-positive rate, and false-negative rate of the spine tumors MRI manual annotations for each doctor. In the training set of primary spine tumors, the five doctors' MRI annotations accuracy rates were 94.44% (A), 98.16% (B), 92.20% (C), 97.84% (D), and 87.99% (E); the false-positive rates were 1.40% (A), 0.00% (B), 5.50% (C), 0.00% (D), and 0.00% (E); the false-negative rates were 4.16% (A), 1.83% (B), 2.30% (C), 2.16% (D), and 12.00% (E). The average accuracy rate, false-positive rate, and false-negative rate of doctors A-E were 94.13%, 1.38%, and 4.49% respectively. In the test group of primary spine tumors, the five doctors' MRI annotations accuracy rates were 97.90% (A), 97.90% (B), 98.40% (C), 98.75% (D), and 96.43% (E); the false-positive rates were 0.50% (A), 0.00% (B), 0.00% (C), 0.00% (D), and 0.00% (E); the false-negative rates were 1.60% (A), 2.10% (B), 1.60% (C), 1.25% (D), and 3.57% (E). The average accuracy rate, false-positive rate, and false-negative rate of doctors A-E were 97.88%, 0.10%, and 2.02% respectively. **Tables 1** and **2** show the details of the precision, recall, F1-score, specificity, and sensitivity in the training and testing sets.

Evaluation With the Turing Test

The mean Average Precision (mAP) results of CNN1, CNN2, and CNN3 were 79.1%, 79.8%, and 80.6% respectively in the axial view, and 84.5%, 85.2%, and 86.1% in the sagittal view, respectively when IoU was over 0.3. These three models were used for Turing testing. The Turing test contained 100 choice questions in the axial view and another 100 choice questions in the sagittal view. **Figure 8** shows the overall percentage of annotation images incorrectly identified by each respondent when asked the following: "Which one was drawn by a human?" in axial and sagittal views. The misclassification rates for the respondents were 44% (F), 52% (G), 62% (H), 59% (I), 46% (J), and 44% (K) in the axial view question, and the average misclassification rate was 51.2% (95% CI: 45.7–57.5%). Among the results of doctors who wrongly selected the prediction of the DL model but did not correctly select the annotations of the doctors A-E, 47.6% chose the prediction by CNN3, 27.4% by CNN2, and 25.0% by CNN1 in the axial view question. Moreover, the misclassification rates for the respondents were 46% (F), 36% (G), 51% (H), 59% (I), 36% (J), and 39% (K) in the sagittal view question, and the average misclassification rate was 44.5% (95% CI: 38.2–51.8%). Among the

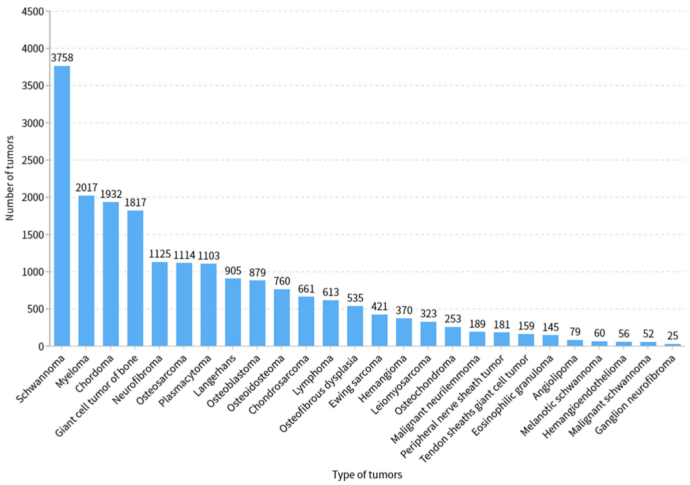


FIGURE 5 | The number of images of different tumor categories.

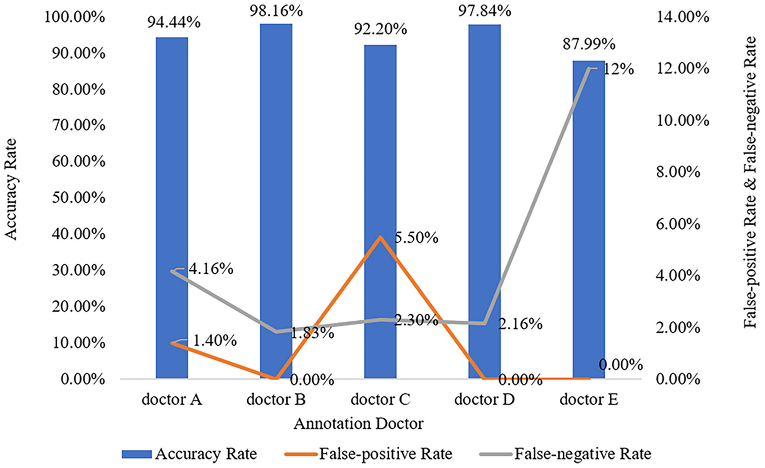


FIGURE 6 | Manual annotation results on MRI primary spine tumor dataset in training set. This figure shows in detail the accuracy rate, false-positive rate, and false-negative rate of the training set annotated by doctors A-E.

results of doctors who wrongly selected the prediction of the DL model but did not correctly select the annotations of the doctors A-E, 48.4% chose the prediction by CNN3, 26.2% by CNN2, and 25.4% by CNN1 in the sagittal view question. According to the results selected by doctors F-K, CNN 3 performed better and the predictions were closer to the manual annotations. Among the six respondents, the lowest misclassification was achieved by an expert radiologist with 25 years of experience. The misclassification rates of the respondents during the Turing test represented an inability to distinguish the annotation source between a human and a computer. The misclassification rates were all >30%, indicating that the DL models passed the Turing test. Therefore, the automated detection of spine tumors by our DL model was equal to that of annotation doctors in our hospital.

The complete raw results from the Turing test are provided as Supplemental Material (see file "TuringTestResults"). **Figure 9** shows an MRI scan in which all doctors chose the DL prediction in both axial and sagittal views, which indicated their failure. **Figure 10** shows an MRI scan in which all doctors F-K correctly selected the annotations of doctors A-E in axial and sagittal views, respectively.

Table 3 shows the assessment time required by each respondent for each multiple-choice question in the Turing test. In the axial view, the average time per question needed by each respondent for the Turing test was 10.72 s (F), 12.08 s (G), 15.73 s (H), 9.46 s (I), 5.69 s (J), and 9.01 s (K). For the 100 choice questions in the axial view, the mean time for each question was 10.45 (range: 5–70) s; therefore, the entire assessment took

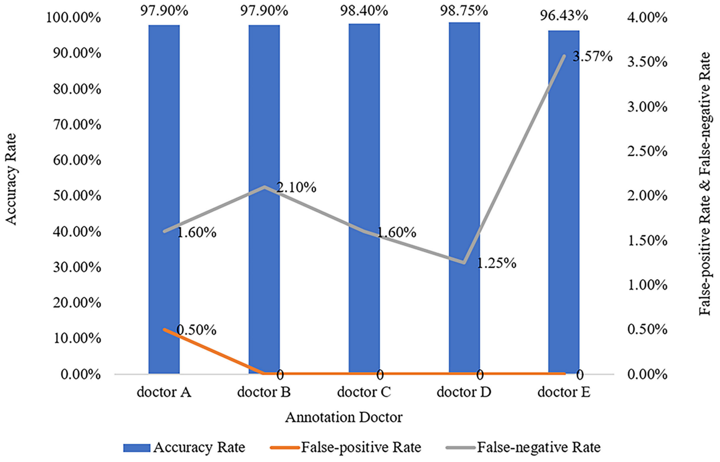


FIGURE 7 | Manual annotation results on MRI primary spine tumor dataset in testing set. This figure shows in detail the accuracy rate, false-positive rate, and false-negative rate of the testing set annotated by doctors A-E.

TABLE 1 | The details of precision, recall, f1 score, specificity, and sensitivity of the training set annotated by the doctor A-E.

	Doctor A	Doctor B	Doctor C	Doctor D	Doctor E
Precision	98.60%	100.00%	94.50%	100.00%	100.00%
Recall	95.95%	98.20%	97.62%	97.89%	89.29%
F1 score	97.26%	99.09%	96.04%	98.93%	94.34%
Specificity	98.56%	100.00%	94.67%	100.00%	100.00%
Sensitivity	95.95%	98.20%	97.62%	97.89%	89.29%

*Precision = TP/(TP+FP); Recall = TP/(TP+FN); F1 score= (2*Precision*Recall)/(Precision + Recall);*
Specificity = TN/(FP+TN); Sensitivity = TP/(TP+FN).
TP = true-positive: It is actually a lesion area, and the doctor annotated it as a lesion area;
FP = false-positive: It is actually not a lesion area, but the doctor annotated it as a lesion area;
FN = false-negative: It is actually a lesion area, but the doctor annotated it is not a lesion area;
TN = true-negative: It is actually not a lesion area, and the doctor annotated it is not a lesion area.

approximately 17 min 25 s per participant (range: 9 min 29 s–26 min 13 s). Moreover, in the sagittal view, the average time per question taken by each respondent for the Turing test was 9.25 s (F), 13.92 s (G), 10.04 s (H), 9.66 s (I), 7.41 s (J), and 9.02 s (K). For the 100 choice questions in the sagittal view, the mean time for each question was 9.88 (range: 4–54) s; therefore, the entire assessment took approximately 16 min 28 s per participant (range: 12 min 21 s–23 min 12 s). No correlation was observed between the time required and the level of accuracy of the assessment. All time results are provided as Supplemental Material (see file "TimingResults").

DISCUSSION

Rather than assessing the performance of our DL model, this study aimed primarily to evaluate whether our DL model for the automated detection of primary spine tumors was as good as that of standard manual annotation methods using the Turing test (18, 19). Although it is doubtful whether AI will ever pass the Turing test for various complex clinical scenarios, it is easy to misunderstand the role of AI in future medical development. AI should complement rather than replace medical professionals. One of our primary aims in using the DL model was to develop a novel method of detecting primary spine tumors from MR images, which is likely to assist orthopedists to find the spine tumors efficiently and reduce the burden on them in the future. The results showed that the accuracy of our DL automatic detection was comparable to that of annotation doctors in radiology or orthopedics. Despite some reports on the applications of AI systems for the spine (20–28), especially on MRI (29) and tumor (30), few studies used the Turing test to evaluate the automatic detection of primary spine tumors in MR images based on DL.

Regardless of symptoms and physical observations, AI facilitates the diagnosis of spine tumors over humans (31). Bluemke et al. (32) reviewed AI radiology research to make a brief guide for authors, reviewers, and interrogators. Wang et al. (33) made a multi-resolution approach for spine metastasis detection using deep Siamese neural networks. Liu et al. (34) compared radiomics with machine learning in the prediction of high-risk cytogenetic status in multiple myeloma based on MRI. The performance of our proposed automatic detection model is not only comparable to that of actual radiologists or orthopedists but also helps to minimize the possibility of overlooking tumors. Massaad et al. (35) used machine learning algorithms to assess the performance of the metastatic spine tumor frailty index. Furthermore, the application of this model can reduce the delay in diagnosing spine tumors because it responds significantly more quickly than humans. Additionally, due to time constraints, radiologists or orthopedists could not evaluate all MR images on their own; sometimes other surgeons or physicians must assess MR spine images. Fortunately, the detection rate of this system is comparable to that of annotation doctors, and the possibility of missing tumors becomes significantly less. Consequently, patients with primary spine tumors can be referred to spine tumor surgeons earlier and more safely.

TABLE 2 | The details of precision, recall, f1 score, specificity, and sensitivity of the testing set annotated by the doctor A-E.

	Doctor A	Doctor B	Doctor C	Doctor D	Doctor E
Precision	99.50%	100.00%	100.00%	100.00%	100.00%
Recall	98.42%	97.94%	98.43%	98.77%	96.55%
F1 score	98.96%	98.96%	99.21%	99.38%	98.25%
Specificity	99.48%	100.00%	100.00%	100.00%	100.00%
Sensitivity	98.42%	97.94%	98.43%	98.77%	96.55%

*Precision = TP/(TP+FP); Recall = TP/(TP+FN); F1 score= (2*Precision*Recall)/(Precision + Recall);*
Specificity = TN/(FP+TN); Sensitivity = TP/(TP+FN).
TP = true-positive: It is actually a lesion area, and the doctor annotated it as a lesion area;
FP = false-positive: It is actually not a lesion area, but the doctor annotated it as a lesion area;
FN = false-negative: It is actually a lesion area, but the doctor annotated it is not a lesion area;
TN = true-negative: It is actually not a lesion area, and the doctor annotated it is not a lesion area.

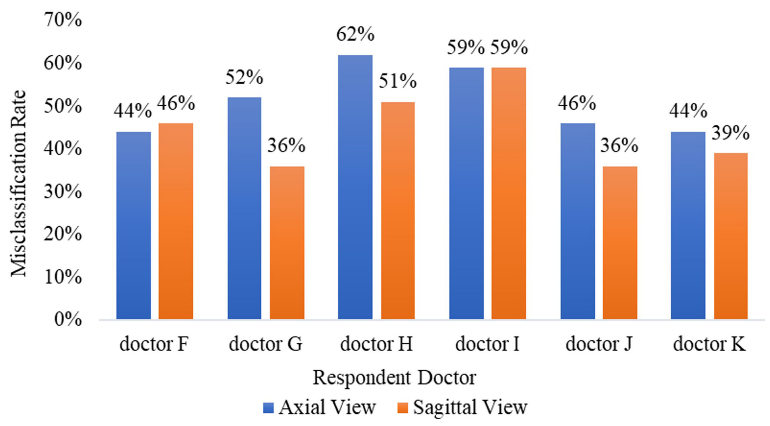
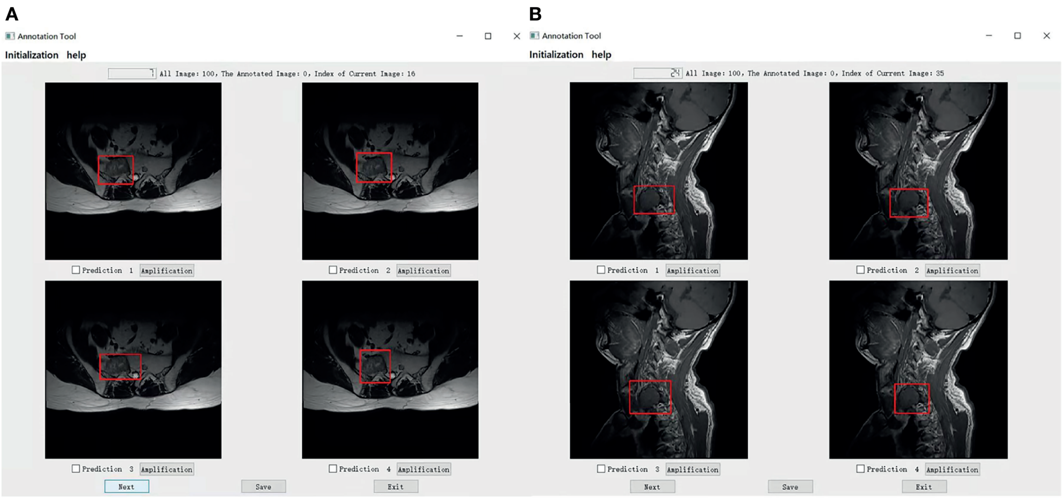


FIGURE 8 | The misclassification rates of all six respondents in axial and sagittal views.

Although the MR images used in this study corresponded to various spine tumor types, the object detection model achieved high accuracy. However, there are always exceptions in clinical settings. For instance, sometimes, it is difficult to identify spine tumors because of signal intensity, location, configuration, or tumor shape. Therefore, the differentiation of spine tumors in neuroimaging is not always reliable. Nevertheless, the use of MRI has facilitated the diagnosis of spine tumors. Another drawback is that if the patient is allergic to contrast agents and/or experiences renal insufficiency, an enhanced MRI scan cannot be performed. In this case, if our proposed system is used to detect spine tumors, we can determine whether other imaging modalities, such as positron emission tomography-computed tomography, should be performed. If MRI cannot be performed owing to renal dysfunction, the proposed system allows for MRI to be performed as minimally as possible.

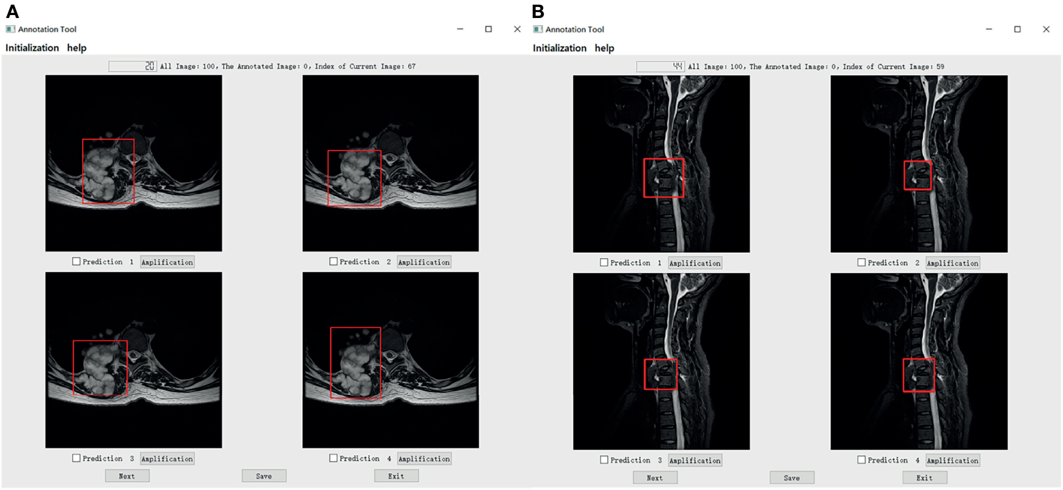
Some individuals believe that passing the Turing test suggests that human-level intelligence can be achieved by machines. However, achieving human-level AI is still far from reality (36, 37). This study, compared to other Turing test studies to date, is one of a few to include a large number of patients with primary spine tumors and a large set of marked spine tumor MR images. The human respondents in this study had only a fair level of agreement with one another, averaging approximately 51.17% accuracy for selecting the human annotation. In a prior report from Scheuer et al. (38), the skilled human interrogators in their



All Image: 100, The Annotated Image: 0, Index of Current Image: 16

All Image: 100, The Annotated Image: 0, Index of Current Image: 35

FIGURE 9 | **(A)** Shows that all doctors F-K have selected prediction 2 which is predicted by the model CNN1 instead of prediction 4 annotated by one of the doctors A-E in axial. **(B)** Shows that four of all doctors F-K have selected the predictions from the models instead of the prediction 3 annotated by one of the doctor A-E in the signal. Among them, four of all doctors F-K chose prediction 4 from the model CNN3, they were doctors F, H, I, and J And doctor G chose prediction 1 predicted by CNN1, and doctor K chose the prediction 2 predicted by CNN2.


A
Annotation Tool
Initialization help
20
□ Prediction 1 Amplification
□ Prediction 2 Amplification
□ Prediction 3 Amplification
□ Prediction 4 Amplification
Next
Save
Exit
B
Annotation Tool
Initialization help
44
□ Prediction 1 Amplification
□ Prediction 2 Amplification
□ Prediction 3 Amplification
□ Prediction 4 Amplification
Next
Save
Exit

All Image: 100, The Annotated Image: 0, Index of Current Image: 67

All Image: 100, The Annotated Image: 0, Index of Current Image: 59

FIGURE 10 | **(A)** Shows that all doctors F-K have correctly selected the prediction 4 annotated by doctors A-E in the axial. And **(B)** shows that all doctors F-K have correctly selected the prediction 1 annotated by doctors A-E in sagittal.

TABLE 3 | The Average Time (second) Per Question Taken by Each Respondent in the Turing test.

	Doctor F	Doctor G	Doctor H	Doctor I	Doctor J	Doctor K
Axial view(s)	10.72	12.08	15.73	9.46	5.69	9.01
Sagittal view(s)	9.25	13.92	10.04	9.66	7.41	9.02

study had a higher sensitivity (45%) for electroencephalography spike events marked by three neurologists. However, the longer the Turing test, the bigger the challenge for a machine to satisfactorily pretend to be a human. In our test of 100 choice questions in the axial or sagittal view, which took approximately 17 min, it would be extremely difficult for a machine to mislead a clinical respondent. Additionally, one of the major challenges in clinical studies dealing with bounding box lesion annotation is to define a "gold standard." Gooding et al. (13) made an evaluation of auto contouring in clinical practice using the Turing test, and Sathish et al. (18) compared lung segmentation and nodule detection between convolutional neural network and humans using the Turing test. Using the choice monitor, the respondents assumed the human's label as the golden standard; hence, they tried to judge the best labels as objectively as possible. This study has demonstrated that with training, the DL model can improve its ability at tumor annotation and mislead the respondents' judgments. In several studies, DL technology has been shown to have a reasonable ability to discriminate between abnormal construct and normal construct in the spine.

Despite a design to limit selection and respondent biases, this study has some limitations. First, the spine tumor MR images were all obtained from a single center, drawn from a cohort of documented patients, and the number of MR images utilized in this study was significantly limited. Hence, it is necessary to improve the accuracy of our system by incorporating multi-center MRI data. Despite the limited number of images, we were able to amplify the training datasets by applying random transformations (e.g., flipping and scaling) to the images. This technique has proven valuable for DL with small datasets. Another limitation was that the proposed system only analyzed and detected the location and approximate outline of spine tumors. Other relevant characteristics, such as whether a spine tumor was benign or malignant, were not recognized in our DL model. Therefore, further research of methods to identify other spine tumor characteristics is necessary. Furthermore, only axial and sagittal images were obtained in our study; hence, the addition of coronal images would improve the model's performance. In addition, to help doctors with image annotation and follow-up, we converted the DICOM into an easy-to-read JPEG. The average misclassification rate of doctors in our current Turing test was over 35%. Despite these limitations, we believe that in the future, our system, with its high accuracy and comparable performance to clinical experts, could be applied to different settings and conditions.

CONCLUSION

In conclusion, this study proposed an AI primary spine tumor detection system that passed the Turing test; respondents were unable to distinguish between our DL model and annotation doctors. The present results show that our DL model may be an efficient tool to assist radiologists or orthopedists in primary spine tumors detection, increasing efficiency and sparing time.

In the future, larger multi-center datasets are necessary to increase the accuracy of our system and validate our model.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

This study was approved by the Medical Science Research Ethics Committee, and it waived the need for informed consent.

AUTHOR CONTRIBUTIONS

HO, FM, and JL, conceptualization, investigation, registration, data curation, visualization, formal analysis, and writing - original draft. XS, investigation, registration, data curation, formal analysis, and writing - original draft. YL, YY, NL, ST, and MY, conceptualization, investigation, methodology, and manual annotation. XL, conceptualization, methodology, project administration, resources, supervision, and validation. HY, SJ, and LJ, conceptualization, funding acquisition, project administration, resources, supervision, validation, writing - review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Beijing Municipal Natural Science Foundation (Z190020), Capital's Funds for Health Improvement and Research (2020-4-40916), Clinical Medicine Plus X - Young Scholars Project, Peking University, the Fundamental Research Funds for the Central Universities (PKU2021LCXQ005), National Natural Science Foundation of China (82102638, 82171927, 81971578), and Peking University Third Hospital Clinical Key Project (BYSYZD2021040, BYSY2018003, BYSYZD2019005).

ACKNOWLEDGMENTS

We appreciate the help on data collection (Beng Wang and Yuqing Zhao) and knowledge support (Hong Liu and Menglei Jiao) from our colleagues.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.814667/full#supplementary-material

REFERENCES

1. Karhade AV, Schwab JH. Introduction to the Special Issue of The Spine Journal on Artificial Intelligence and Machine Learning. *Spine J* (2021) 21 (10):1601–3. doi: 10.1016/j.spinee.2021.03.028
2. Suri A, Jones BC, Ng G, Anabaraonye N, Beyrer P, Domi A, et al. A Deep Learning System for Automated, Multi-Modality 2D Segmentation of Vertebral Bodies and Intervertebral Discs. *Bone* (2021) 149:115972. doi: 10.1016/j.bone.2021.115972
3. Yang HS, Kim KR, Kim S, Park JY. Deep Learning Application in Spinal Implant Identification. *Spine (Phila Pa 1976)* (2021) 46(5):E318–24. doi: 10.1097/BRS.0000000000003844
4. Cina A, Bassani T, Panico M, Luca A, Masharawi Y, Brayda-Bruno M, et al. 2-Step Deep Learning Model for Landmarks Localization in Spine Radiographs. *Sci Rep* (2021) 11(1):9482. doi: 10.1038/s41598-021-89102-w
5. Kim DH, Jeong JG, Kim YJ, Kim KG, Jeon JY. Automated Vertebral Segmentation and Measurement of Vertebral Compression Ratio Based on Deep Learning in X-Ray Images. *J Digit Imaging* (2021) 34(4):853–61. doi: 10.1007/s10278-021-00471-0
6. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI. *Radiology* (2021) 300(1):130–8. doi: 10.1148/radiol.2021204289
7. Huang J, Shen H, Wu J, Hu X, Zhu Z, Lv X, et al. Spine Explorer: A Deep Learning Based Fully Automated Program for Efficient and Reliable Quantifications of the Vertebrae and Discs on Sagittal Lumbar Spine MR Images. *Spine J* (2020) 20(4):590–9. doi: 10.1016/j.spinee.2019.11.010
8. Merali Z, Wang JZ, Badhiwala JH, Witiw CD, Wilson JR, Fehlings MG. A Deep Learning Model for Detection of Cervical Spinal Cord Compression in MRI Scans. *Sci Rep* (2021) 11(1):10473. doi: 10.1038/s41598-021-89848-3
9. Ito S, Ando K, Kobayashi K, Nakashima H, Oda M, Machino M, et al. Automated Detection of Spinal Schwannomas Utilizing Deep Learning Based on Object Detection From Magnetic Resonance Imaging. *Spine (Phila Pa 1976)* (2021) 46(2):95–100. doi: 10.1097/BRS.0000000000003749
10. Turing AM. Computing Machinery and Intelligence. *Mind* (1950) 59:433–60. doi: 10.1093/mind/LIX.236.433
11. Bush JT, Pogany P, Pickett SD, Barker M, Baxter A, Campos S, et al. A Turing Test for Molecular Generators. *J Med Chem* (2020) 63(20):11964–71. doi: 10.1021/acs.jmedchem.0c01148
12. Powell J. Trust Me, I'm a Chatbot: How Artificial Intelligence in Health Care Fails the Turing Test. *J Med Internet Res* (2019) 21:e16222. doi: 10.2196/16222
13. Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, van der Stoep J, et al. Comparative Evaluation of Autocontouring in Clinical Practice: A Practical Method Using the Turing Test. *Med Phys* (2018) 45(11):5105–15. doi: 10.1002/mp.13200
14. Warwick K, Shah H. Passing the Turing Test Does Not Mean the End of Humanity. *Cognit Comput* (2016) 8:409–19. doi: 10.1007/s12559-015-9372-6
15. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* (2017) 39(6):1137–49. doi: 10.1109/TPAMI.2016.2577031
16. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *CVPR* (2016), 770–8. doi: 10.1109/CVPR.2016.90
17. Lin T-Y, Dollár P, Girshick RB, He KM, Hariharan B, Belongie SJ. Feature Pyramid Networks for Object Detection. *CVPR* (2017) 2117–25. doi: 10.1109/CVPR.2017.106
18. Sathish R, Sathish R, Sethuraman R, Sheet D. Lung Segmentation and Nodule Detection in Computed Tomography Scan Using a Convolutional Neural Network Trained Adversarially Using Turing Test Loss. *Annu Int Conf IEEE Eng Med Biol Soc* (2020) 2020:1331–4. doi: 10.1109/EMBC44109.2020.9175649
19. Webster J, Amos M. A Turing Test for Crowds. *R Soc Open Sci* (2020) 7 (7):200307. doi: 10.1098/rsos.200307
20. Yeh YC, Weng CH, Huang YJ, Fu CJ, Tsai TT, Yeh CY. Deep Learning Approach for Automatic Landmark Detection and Alignment Analysis in

Whole-Spine Lateral Radiographs. *Sci Rep* (2021) 11(1):7618. doi: 10.1038/s41598-021-87141-x

21. Jakubicek R, Chmelik J, Ourednicek P, Jan J. Deep-Learning-Based Fully Automatic Spine Centerline Detection in CT Data. *Annu Int Conf IEEE Eng Med Biol Soc* (2019) 2019:2407–10. doi: 10.1109/EMBC.2019.8856528
22. Maki S, Furuya T, Yoshii T, Egawa S, Sakai K, Kusano K, et al. Machine Learning Approach in Predicting Clinically Significant Improvements After Surgery in Patients With Cervical Ossification of the Posterior Longitudinal Ligament. *Spine (Phila Pa 1976)* (2021) 46(24):1683–9. doi: 10.1097/BRS.0000000000004125
23. Shin Y, Han K, Lee YH. Temporal Trends in Cervical Spine Curvature of South Korean Adults Assessed by Deep Learning System Segmentation, 2006-2018. *JAMA Netw Open* (2020) 3(10):e2020961. doi: 10.1001/jamanetworkopen.2020.20961
24. Galbusera F, Niemeyer F, Wilke HJ, Bassani T, Casaroli G, Anania C, et al. Fully Automated Radiological Analysis of Spinal Disorders and Deformities: A Deep Learning Approach. *Eur Spine J* (2019) 28(5):951–60. doi: 10.1007/s00586-019-05944-z
25. Wang KY, Suresh KV, Puvanesarajah V, Raad M, Margalit A, Jain A. Using Predictive Modeling and Machine Learning to Identify Patients Appropriate for Outpatient Anterior Cervical Fusion and Discectomy. *Spine (Phila Pa 1976)* (2021) 46(10):665–70. doi: 10.1097/BRS.0000000000003865
26. Korez R, Putzier M, Vrtovec T. A Deep Learning Tool for Fully Automated Measurements of Sagittal Spinopelvic Balance From X-Ray Images: Performance Evaluation. *Eur Spine J* (2020) 29(9):2295–305. doi: 10.1007/s00586-020-06406-7
27. Han SS, Azad TD, Suarez PA, Ratliff JK. A Machine Learning Approach for Predictive Models of Adverse Events Following Spine Surgery. *Spine J* (2019) 19(11):1772–81. doi: 10.1016/j.spinee.2019.06.018
28. Al Arif SMMR, Knapp K, Slabaugh G. Fully Automatic Cervical Vertebrae Segmentation Framework for X-Ray Images. *Comput Methods Programs BioMed* (2018) 157:95–111. doi: 10.1016/j.cmpb.2018.01.006
29. Won D, Lee HJ, Lee SJ, Park SH. Spinal Stenosis Grading in Magnetic Resonance Imaging Using Deep Convolutional Neural Networks. *Spine (Phila Pa 1976)* (2020) 45(12):804–12. doi: 10.1097/BRS.0000000000003377
30. Chianca V, Cuocolo R, Gitto S, Albano D, Merli I, Badalyan J, et al. Radiomic Machine Learning Classifiers in Spine Bone Tumors: A Multi-Software, Multi-Scanner Study. *Eur J Radiol* (2021) 137:109586. doi: 10.1016/j.ejrad.2021.109586
31. DiSilvestro KJ, Veeramani A, McDonald CL, Zhang AS, Kuris EO, Durand WM, et al. Predicting Postoperative Mortality After Metastatic Intraspinal Neoplasm Excision: Development of a Machine-Learning Approach. *World Neurosurg* (2021) 146:e917–24. doi: 10.1016/j.wneu.2020.11.037
32. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the Radiology Editorial Board. *Radiology* (2020) 294(3):487–9. doi: 10.1148/radiol.2019192515
33. Wang J, Fang Z, Lang N, Yuan H, Su MY, Baldi P. A Multi-Resolution Approach for Spinal Metastasis Detection Using Deep Siamese Neural Networks. *Comput Biol Med* (2017) 84:137–46. doi: 10.1016/j.compbiomed.2017.03.024
34. Liu J, Zeng P, Guo W, Wang C, Geng Y, Lang N, et al. Prediction of High-Risk Cytogenetic Status in Multiple Myeloma Based on Magnetic Resonance Imaging: Utility of Radiomics and Comparison of Machine Learning Methods. *J Magn Reson Imaging* (2021) 54(4):1303–11. doi: 10.1002/jmri.27637
35. Massaad E, Williams N, Hadzipasic M, Patel SS, Fourman MS, Kiapour A, et al. Performance Assessment of the Metastatic Spinal Tumor Frailty Index Using Machine Learning Algorithms: Limitations and Future Directions. *Neurosurg Focus* (2021) 50(5):E5. doi: 10.3171/2021.2.FOCUS201113
36. Warwick K, Shah H. Can Machines Think? A Report on Turing Test Experiments at the Royal Society. *J Exp Theor Artif Intell* (2016) 28:989–1007. doi: 10.1080/0952813x.2015.1055826
37. Barone P, Bedia MG, Gomila A. A Minimal Turing Test: Reciprocal Sensorimotor Contingencies for Interaction Detection. *Front Hum Neurosci* (2020) 14:102. doi: 10.3389/fnhum.2020.00102
38. Scheuer ML, Bagic A, Wilson SB. Spike Detection: Inter-Reader Agreement and a Statistical Turing Test on a Large Data Set. *Clin Neurophysiol* (2017) 128(1):243–50. doi: 10.1016/j.clinph.2016.11.005

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Oncology

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/fonc.2022.853281

Kidney Tumor Segmentation Based on FR2PAttU-Net Model

Peng Sun[1], Zengnan Mo[2], Fangrong Hu[1], Fang Liu[3], Taiping Mo[1], Yewei Zhang[4] and Zhencheng Chen[1*]*

[1] School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, China, [2] Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, China, [3] College of Life and Environment Science, Guilin University of Electronic Technology, Guilin, China, [4] Hepatopancreatobiliary Center, The Second Affiliated Hospital of Nanjing Medical University, Nanjing, China

The incidence rate of kidney tumors increases year by year, especially for some incidental small tumors. It is challenging for doctors to segment kidney tumors from kidney CT images. Therefore, this paper proposes a deep learning model based on FR2PAttU-Net to help doctors process many CT images quickly and efficiently and save medical resources. FR2PAttU-Net is not a new CNN structure but focuses on improving the segmentation effect of kidney tumors, even when the kidney tumors are not clear. Firstly, we use the R2Att network in the "U" structure of the original U-Net, add parallel convolution, and construct FR2PAttU-Net model, to increase the width of the model, improve the adaptability of the model to the features of different scales of the image, and avoid the failure of network deepening to learn valuable features. Then, we use the fuzzy set enhancement algorithm to enhance the input image and construct the FR2PAttU-Net model to make the image obtain more prominent features to adapt to the model. Finally, we used the KiTS19 data set and took the size of the kidney tumor as the category judgment standard to enhance the small sample data set to balance the sample data set. We tested the segmentation effect of the model at different convolution and depths, and we got scored a 0.948 kidney Dice and a 0.911 tumor Dice results in a 0.930 composite score, showing a good segmentation effect.

Keywords: kidney tumor segmentation, FR2PAttU-Net, KiTS19, data augmentation, CT

OPEN ACCESS

Edited by:
Fu Wang,
Xi'an Jiaotong University, China

Reviewed by:
P.K. Vinod,
International Institute of Information Technology, Hyderabad, India
Enrico Checcucci,
IRCCS Candiolo Cancer Institute, Italy

****Correspondence:***
Zhencheng Chen
zhenchchen@163.com
Yewei Zhang
zhangyewei@njmu.edu.cn

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *12 January 2022*
Accepted: *17 February 2022*
Published: *17 March 2022*

Citation:
Sun P, Mo Z, Hu F, Liu F, Mo T, Zhang Y and Chen Z (2022) Kidney Tumor Segmentation Based on FR2PAttU-Net Model. Front. Oncol. 12:853281. doi: 10.3389/fonc.2022.853281

INTRODUCTION

In recent years, the incidence rate of kidney tumors has increased (1–3). If we rely on artificial ways to process medical image data of patients, it will waste a lot of time. And because of the difference in medical experience, some small and challenging methods to find tumors are easily ignored by doctors, and subjective factors lead to misjudgment. Therefore, how to use the deep learning model to segment kidney tumors is a challenging task (4). However, most kidney image analysis is usually based on kidney segmentation rather than tumor segmentation or two deep models: the first to segment the kidney and the second to segment the tumor on the kidney (5, 6). Among many current research schemes, they get scored about 0.97 kidney Dice and 0.85 tumor Dice (7). These methods can provide higher values from the extracted features by pre-analyzing the information provided by the image; they play a role in the early detection and diagnosis of abnormalities. However, new research in this field is still significant because effective and accurate segmentation always has room for improvement, especially considering ignoring minor medical errors (8, 9). In these cases, the segmentation task of kidney and kidney tumors becomes more complex (10). Therefore, it is necessary to study the application of more in-depth learning methods in kidney tumors without manual intervention, improve the analysis efficiency, and reduce workload of experts to improve the segmentation effect of tumors.

This paper proposes an automatic segmentation method of kidney and tumor in CT image to support the diagnosis of kidney disease of experts: a flexible model that can segment kidneys and tumors simultaneously. In the design of our improved model, we consider the primary shortcomings of the existing deep learning model and develop a new, efficient and automatic kidney segmentation method. In this article, we emphasize the following contributions:

(1) We use the cascade network model. The first model is used to coarse segment the kidney and tumor ROI (the kidney without tumor is not segmented). The second model is used to finely segment the tumor in CT images to improve the segmentation effect of the tumor.
(2) We propose to reconstruct labeled CT images based on tumor size to balance the kidney tumor data set and reduce the impact of category imbalance.
(3) We propose the FR2PAttU-Net model and verify it in the KiTS19 data set. Finally, it can segment tumors with high precision, even when kidney tumors are unclear.

Therefore, we believe that the proposed FR2PAttU-Net model provides an effective kidney tumor segmentation method, improving the segmentation effect and diagnosis rate of kidney tumors.

The overall structure of this paper is as follows. Section 2 introduces the relevant research and findings; Section 3 discusses the methods; Section 4 reports the experiments carried out to verify our research, the comparative analysis of the corresponding results and other similar studies, and Section 5 gives the discussion and conclusions.

Related Work

The task of kidney segmentation has not only recently started. Several methods have been developed in the past few years, and more and more expressive results have been obtained to solve this problem.

In 2015, Ronneberger et al. (11) proposed the U-Net model to realize the segmentation of medical images. The U-Net model is one of the earliest algorithms for semantic segmentation using a Fully Convolutional Network. The symmetric U-shaped structure that contains the compression path and the expansion path in the paper was very innovative at the time. Due to its relatively simple task, U-Net has achieved a meager error rate through 30 pictures, supplemented by a data expansion strategy, and won the championship's championship. First, it established the position of the U-Net model in medical image segmentation. Then a variant algorithm based on the U-Net model is applied in multiple directions of medical image segmentation.

Since U-Net, a series of algorithms have been derived for medical image segmentation. For example, Yang et al. (12) proposed a method for measuring lung parenchymal parameters based on the ResU-Net model based on lung window CT images, and analyzed the relationship between lung volume and CT value or density, and concluded that lung volume is negatively correlated with CT value or density. Oktay et al. (13) proposed a new attention gate (AG) model for medical imaging, which can automatically learn to focus on target structures of different shapes and sizes, and use the model trained by AGs to implicitly learn to suppress outside areas in the input image while highlighting salient features useful for specific tasks. The experimental results show that, while maintaining computational efficiency, AGs consistently improve the prediction performance of U-Net under different data sets and training scales. Alom et al. (14) proposed a U-Net-based recurrent convolutional neural network (RCNN) and a U-Net model-based recurrent, residual convolutional neural network (RRCNN), named RU-Net and R2U-Net, respectively. The proposed model utilizes the capabilities of the U network, residual network, and RCNN. The experimental results show that compared with the equivalent model, including U-Net and residual U-Net (ResU-Net), the model has the advantages of segmentation tasks. Better performance. Wang et al. (15) used U-net combined with the recurrent residual and attention models to segment the image. Experiments show that they can obtain better results.

Since 2020, the segmentation of kidney and kidney tumors based on the U-Net model has gradually increased. Isensee et al. (16) introduced nnU-Net ('no-new-Net'), which eliminated many of the powerful reasons for the unnecessary bells and whistles in the proposed network design, and instead focused on the remaining aspects of the performance and versatility of the composition method. nnU-Net achieved the highest average dice score in the challenge online leaderboard. Da Cruz et al. (17) used U-Net 2D for initial segmentation and delineated the kidney (CT) image. In the KiTS19 challenge, its average Dice coefficient is 93.03%. Turk et al. (18) used the superior characteristics of the existing V-Net model to propose a new hybrid model, which improved the previously unapplied encoder and decoder stages and obtained 97.7% kidney Dice and 86.5% tumor Dice.

In 2021, Heller et al. (19) released the KiTS19 challenge and published the top five methods and segmentation effects in the article: The fifth place was made by Ma (20). A 3D U-Net is used as the main architecture which is based on nnU-Net implementation. Compared to the original 3D U-Net, the notable changes are padding convolutions, instance normalization, and leaky-ReLUs. This submission scored a 0.973 kidney Dice, and a 0.825 tumor Dice resulting in a 0.899 composite score. The fourth place was made by Hou et al. (21). They use a cascaded volumetric convolutional network for kidney tumor segmentation from CT volumes. There are two steps in this model, and one is coarse location, the other is fine predictions. This submission scored a 0.974 kidney Dice and a 0.831 tumor Dice resulting in a 0.902 composite score. The third place was made by Mu et al. (22). They used multi-resolution VB-nets for segmentation of kidney tumor, and they scored a 0.973 kidney Dice and a 0.832 tumor Dice resulting in a 0.903 composite score. The second place was made by Hou et al. (23). They used cascaded semantic segmentation for kidney and tumor. This cascaded approach had three stages. Stage 1 performed a coarse segmentation of all kidneys in the image. The second stage is run for each rectangular kidney region that is found by the first stage, and in the third stage of the model, a fully convolutional net is used to segment the tumor voxels from the kidney voxels. This submission scored a 0.967 kidney Dice and a 0.845 tumor Dice resulting in a 0.906 composite score. The first place was made by Isensee et al. (24). Three 3D U-Net architectures were tested using five-fold cross validation, and this submission scored a 0.974 kidney Dice and a 0.851 tumor Dice resulting in a 0.912 composite score.

Based on the above analysis, we find that most algorithms in the field of medical image segmentation take the U-Net architecture as the starting point for further development and derive a series of improved and variant algorithms from realizing the task of medical image segmentation. Although most models can achieve good results, there is always room for effective and accurate segmentation improvement. Furthermore, although multiple networks will increase the time cost, they can improve the segmentation effect simultaneously. Therefore, in this work, we propose the FR2PAttU-Net model to improve the segmentation performance of kidney tumor CT images.

MATERIALS AND METHODS

This section will introduce the overall scheme of kidney tumor segmentation. The first section introduces the structure of the FR2PAttU-Net model for kidney and tumor segmentation. The second section presents the steps of kidney tumor segmentation, namely, data preparation, coarse segmentation, and fine segmentation. We will explain each piece in detail next.

FR2PAttU-Net

We propose the FR2PAttU-Net model, where F, R2, P, and Att are the abbreviations for Fuzzy set, Recurrent Residual, Parallel, and Attention, respectively. The "U"-shaped architecture of the standard U-Net is used in our network. **Figure 1** shows the architecture and layers that make up our network, with the contraction path defined on the left of the model and the symmetrical expansion path specified on the right. All convolutional layers are modified from consecutive 3 × 3 kernels to parallel kernels, and we will introduce the specific structures and functions of F, R2, P, and Att step by step. Furthermore, we use the activation function Leaky-ReLU.

Image Enhancement Based on Fuzzy Set (F)

Image enhancement emphasizes or sharpens certain features of an image, such as edges, contours, contrast, etc., for display, observation, or further analysis and processing. The processed image is transformed through specific image processing into an image of better visual quality and effect or more "useful" for a particular application. Fuzzy sets provide a form of loose processing information. For example, using fuzzy sets to enhance images of kidneys and kidney tumors can make the entire kidney more clearly delineated, making it more adaptable to the network.

Image enhancement based on fuzzy sets mainly includes three steps: image fuzzy feature extraction, membership function value correction, and fuzzy domain inverse transformation (25). Define Z as an object set, where z represents a type of element in Z. A fuzzy set A in Z is mainly characterized by a degree of membership $\mu_A(z)$. In this regard, the fuzzy set A is composed of z-values and membership

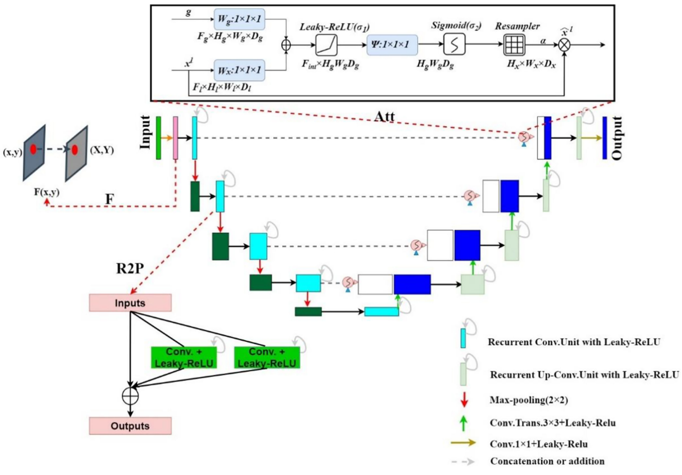


FIGURE 1 | FR2PAttU-Net Model.

We use fuzzy sets to perform a gray-scale transformation to enhance the image. Then, we stipulate the following fuzzy rules:

R1: IF one pixel is dark, THEN makes this pixel darker;

R2: IF one pixel is gray, THEN keeps it gray;

R3: IF one pixel is bright, THEN makes this pixel brighter;

This rule represents our approach. But, of course, the pixels in the IF condition are dark (either gray or bright), and this concept is blurred. In the same way, the darker (or staying gray, or merrier) in the THEN conclusion is also fuzzy. To this end, we need to establish a membership function to determine the membership of a pixel to three conditions (26).

The determination of the membership function is very complicated. However, here we try to make it simple. First, a pixel is dark (fuzzy), then the approximate shape of its membership function is that the domain membership is 1 when it is lower than a certain value z1. After the gray level crosses a specific value, z2, its membership degree is 0. So, of course, z1 ≠ z2. Then we perform linear interpolation between z1 and z2, and then we can get the membership function of R1. Similarly, R2 and R3 are the same.

For pixel Z_0, it is necessary to calculate the corresponding membership degrees $\mu_{\mathrm{dark}}(Z_0)$, $\mu_{\mathrm{gray}}(Z_0)$, and $\mu_{\mathrm{bright}}(Z_0)$ according to the rules R1, R2, and R3. This process is called fuzzification. The function (or the corresponding relationship) used to fuzz an input quantity is the knowledge base.

After fuzzification, the three membership degrees $\mu_{\mathrm{dark}}(Z_0)$, $\mu_{\mathrm{gray}}(Z_0)$, and $\mu_{\mathrm{bright}}(Z_0)$ corresponding to a pixel can be deblurred. There are many de-obfuscation algorithms, and Equation (1) is the center of gravity method.

$$v_0 = \frac{\mu_{dark}(z_0) \times v_d + \mu_{gray}(z_0) \times v_g + \mu_{bright}(z_0) \times v_b}{\mu_{dark}(z_0) + \mu_{gray}(z_0) + \mu_{bright}(z_0)} \tag{1}$$

Among them, v_d, v_g, and v_b are the single output values. Then, pixel Z_0 must calculate the corresponding membership degrees $\mu_{\mathrm{dark}}(Z_0)$, $\mu_{\mathrm{gray}}(Z_0)$, and $\mu_{\mathrm{bright}}(Z_0)$ according to R1, R2, and R3. Finally, we obtain a weighted maturity estimate, which is the most output value. At this point, the output v_0 is obtained.

The specific transformation result can be obtained by Equations (2) and (3).

$$m = image[x][y] \tag{2}$$

$$f(x) = \begin{cases} 0,\ 0 \le m < 0.15 \\ (m - 0.15)/0.28 \times 127,\ 015 \le x < 0.43 \\ (m - 0.45)/0.28 \times 255 + (0.71 - m)/0.28 \times 127,\ 0.43 \le x < 0.71 \\ 255, else \end{cases} \tag{3}$$

image[*x*][*y*] is the pixel value at point (x, y), This article takes m values 0.15, 0.43, 0.71, 1, respectively, and divides the entire pixel value into four regions to complete the pixel conversion.

The effect of the fuzzy set enhancement algorithm is shown in the **Figure 2**.

Recurrent-Residual-Parallel Convolutional Network (R2P)

The residual network enables the training of deeper networks, and the recurrent residual convolutional layer allows the network to extract better features. The network provides for the network to deepen and avoid the inability to learn the gradient under the same amount of parameters, resulting in better performance. As shown in **Figure 3**, the model uses the recurrent residual block instead of the traditional Conv + ReLU layer in the encoding and decoding process, which can train a deeper network. All convolution layers are composed of successive convolution (convolution kernel 3 × 3) are modified to parallel convolutional network, and we tested parallel convolutions (convolution kernel = 3 × 3), and parallel convolutions (convolution kernel = 3 × 3 and 5 × 5), and perform parallel convolution operations on the image, stitching all outputs into one deep feature map. Different convolution and pooling operations can obtain more information about the input image, and processing these operations in parallel and combining all the

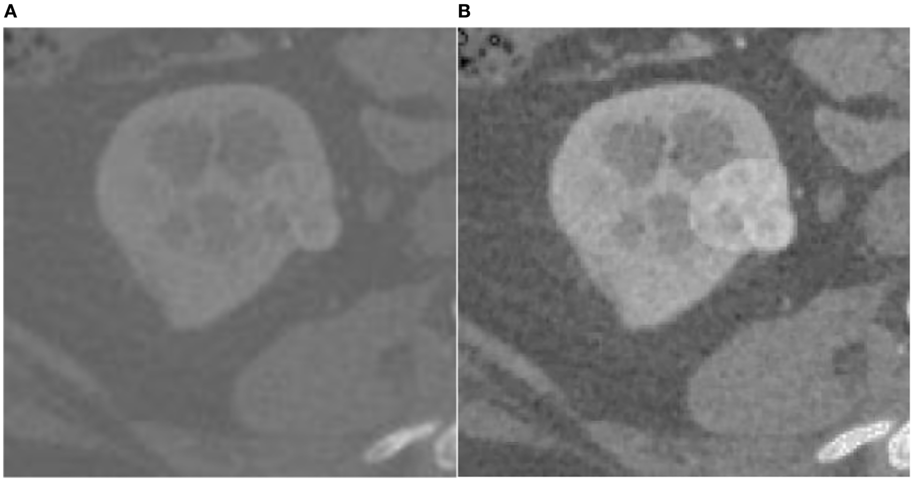


FIGURE 2 | Result of fuzz set enhancement algorithm, **(A)** is the original CT image, **(B)** is the image enhanced by the fuzzy set.

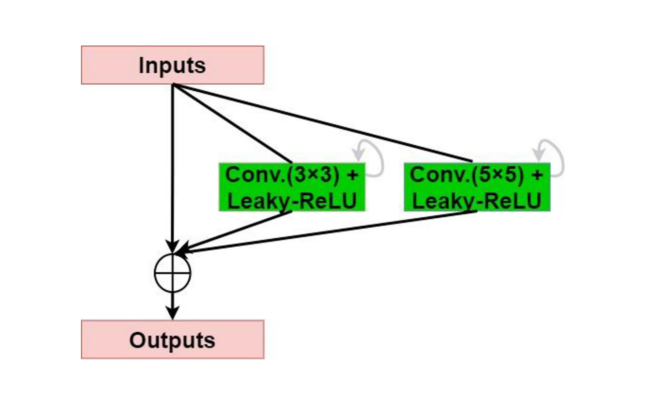


FIGURE 3 | Recurrent-residual-parallel convolutional network (R2P).

results will yield a better image representation. We use different convolution kernels for image feature extraction, which fully increases width of the model, increases the receptive field, and improves the robustness of the network, thereby improving the ability of the model to adapt to features of different scales in the image. Then, summation of features at other time steps is used to obtain a more expressive quality, which helps extract lower-level features; finally, skip connections are not cut off in the original U-Net but are cascaded operate.

Attention Gate (Att)

An attention gate is added to the model, which automatically learns to distinguish the shape and size of objects. **Figure 4** shows the calculation method of the attention gate. First, the output g corresponding to the decoder part is upsampled + convolved, and then 1-dimensional convolution is used to reduce the dimension of g and x (from the encoder at the same level). As a result, the number of channels becomes 1/2 of the original. Then the two parts of the results are added; after the activation function Leaky-ReLU and one-dimensional convolution, the number of channels is reduced to 1. Then through the Sigmod function, a 1-dimensional attention map with the same size as x is obtained, and the original x is used as element-wise multiplication to get a weighted vector.

Leaky-ReLU

Furthermore, the Leaky-ReLU activation function and batch normalization follow closely (27). The difference from ReLU is that the negative axis of Leaky-ReLU retains a tiny constant leak so that when the input information is less than 0, the information is not wholly lost, and the corresponding retention is carried out. That is, ReLU has no gradient when the value is less than zero, and Leaky-ReLU gives a slight incline when the value is less than 0. It is equivalent to allowing backpropagation of gradients corresponding to intervals less than 0 rather than direct interception.

Segmentation Scheme

This paper mainly segments kidneys and tumors from three parts. In the first part, kidney data is collected and preprocessed. We picked out the slice range containing the kidney from the CT images and discarded the invalid area that did not include the kidney and tumor. The second part is coarse segmentation. We use the first model to segment the approximate size of the kidney and tumor. This step is only used to locate the initial location of the kidney and tumor, select the ROI, and do not segment. The third part analyzes ROIs and reconstructs CT images with labels to balance the kidney tumor segmentation dataset. Then we use the second model for fine segmentation of kidneys and tumors, where the ROI region is used as the input image to improve the segmentation effect. The segmentation scheme is shown in **Figure 5**. Each of these steps is described in detail in the subsections that follow.

Data Preparation

In this study, we downloaded the available data set from the homepage of the KiTS19 data set and did not use additional data. A total of 210 scans with high-quality ground truth segmentations were downloaded from the KiTS19 data set, publicly available on GitHub (https://github.com/neheller/kits19). The homepage of the KiTS19 data set provides other instructions on the preparation of the data set and the ethics committee (28). Manual segmentation may cause many errors in subsequent kidney or kidney tumor monitoring. In addition, it is very time-consuming and may degrade system performance (29). Despite these adverse effects, we still used the KiTS19 dataset because of the lack of available datasets in the literature. Patients with cysts and tumor thrombi were excluded from the KiTS19 dataset because in these patients, the tumor was beyond what we thought was the primary site and the appropriate boundaries were unclear. Therefore, we only selected kidneys with tumor lesions in this study to construct

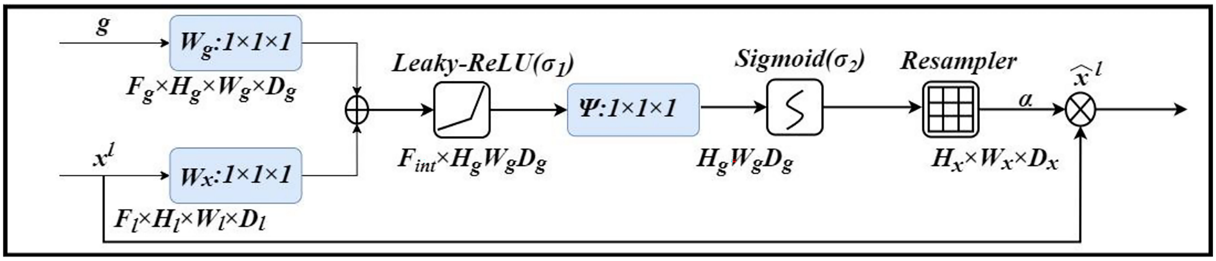

g
W_g:1×1×1
$F_g \times H_g \times W_g \times D_g$
x^l
W_x:1×1×1
$F_l \times H_l \times W_l \times D_l$
Leaky-ReLU(σ_1)
$F_{int} \times H_g W_g D_g$
Ψ:1×1×1
Sigmoid(σ_2)
$H_g W_g D_g$
Resampler
α
$H_x \times W_x \times D_x$
$\hat{x}^l$

FIGURE 4 | Attention Gate (Att).

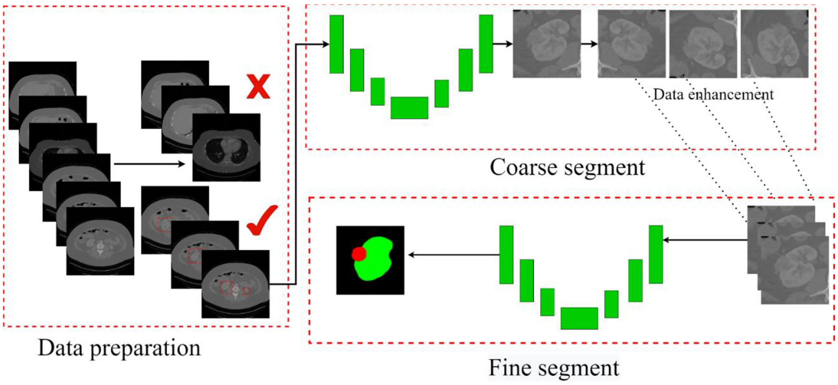


FIGURE 5 | The overall process. Contains three parts: data preparation, coarse segmentation, and fine segmentation. Section *Data Preparation* introduces data preparation in detail, Section *Coarse Segmentation Based on FR2PAttU-Net Model* introduces coarse segmentation and Section *Fine Segmentation Based on FR2PAttU-Net Model* introduces fine segmentation.

training and test datasets. The task is the segmentation of kidneys and kidney tumors in contrast-enhanced abdominal CT without judging the type of tumor.

To make the data satisfy our network model, we cut the 3D data into several 2D images with 512 × 512 pixels. In addition, all the slices without kidney markers are discarded. Processing the original CT images before sending them to the network is a crucial step for practical training. The first aspect to consider is the presence of unexpected substances that may appear in the body of the patient. In particular, the metal artifacts have a significant negative impact on the quality of CT images, which is a well-known fact. The main problem with artifacts is that the areas generated in the image have abnormal intensity values or are much higher or lower than the intensity values of pixels corresponding to organic tissues. Since deep learning algorithms are based on data-driven models, abnormal voxels corresponding to non-organic artifacts can significantly affect learning. To reduce the impact of non-organic artifacts, we uniformly process the complete data set, namely, training and test data. We only consider the effective intensity range between 0.5 and 99.5% in all images and tailor the outliers accordingly. After preprocessing, data is normalized with the normal foreground mean and standard deviation to improve the training effect of the network.

Coarse Segmentation Based on FR2PAttU-Net Model

Since some organs in the abdomen in CT images are similar in shape and texture to the kidney, they will also segment them at the end, so it is necessary to coarse segment and extracts the kidney ROI. Coarse segmentation based on FR2PAttU-Net is performed on each slice, thus constructing a 2D segmentation of kidney tumors. The model is trained from CT images with an original size of 512 × 512 pixels. The tumor and the kidney are regarded as the same type to make a label to construct a binary segmentation model. That is, the label only includes the background and the kidney. After the model segmented the tumor and kidney area, the ROI area smaller than 128 × 128 was expanded to 128 × 128 and expanded the ROI area larger than 128 × 128 to 256 × 256, it was better to obtain the kidney, tumor, and background information. Through the coarse segmentation of the kidney, the kidney region is separated, which reduces the scope of the problem and increases the chance of successful segmentation of kidney tumors. **Figure 6** shows coarse segmentation results of CT images ranging from 512 × 512 pixels to 128 × 128 pixels.

Fine Segmentation Based on FR2PAttU-Net Model

Coarse segmentation can reduce the range of the segmented image and save the entire computing resources of the model. Since the fine segmentation needs to use the coarsely segmented ROI area as training data, to avoid the impact of the imbalanced distribution of the data in training set on the tumor segmentation results, this paper needs to enhance the small sample data to balance the sample data set. This paper calculates and counts the tumor size in the training set. There are 4,691 ROI images containing tumors. The area size distribution of the connected regions is shown in **Figure 7**.

Analyzing the data in **Figure 7**, we found that the tumor size distribution in the training set was not even, where the tumor area differed by about a factor of 2 between 0–500 and 2,000–3,000. Therefore, we must reconstruct the data to balance the kidney tumor segmentation dataset. For fewer datasets, we adopted data augmentation methods such as flipping, rotating, shifting, and mirroring and extended them to more data to balance the kidney tumor dataset. **Figure 8** shows several commonly used data augmentation functions.

We use the second model to accurately segment kidneys and tumors after balancing the dataset in the ROI region. Here, the input image is the kidney ROI region, all pixels predicted to be background are set to 0, and kidney and tumor are represented by different pixels.

EXPERIMENTAL RESULTS

In this section, we detail the experimental results validating the proposed method. First, we introduce the metrics used for performance validation and then discuss the results obtained

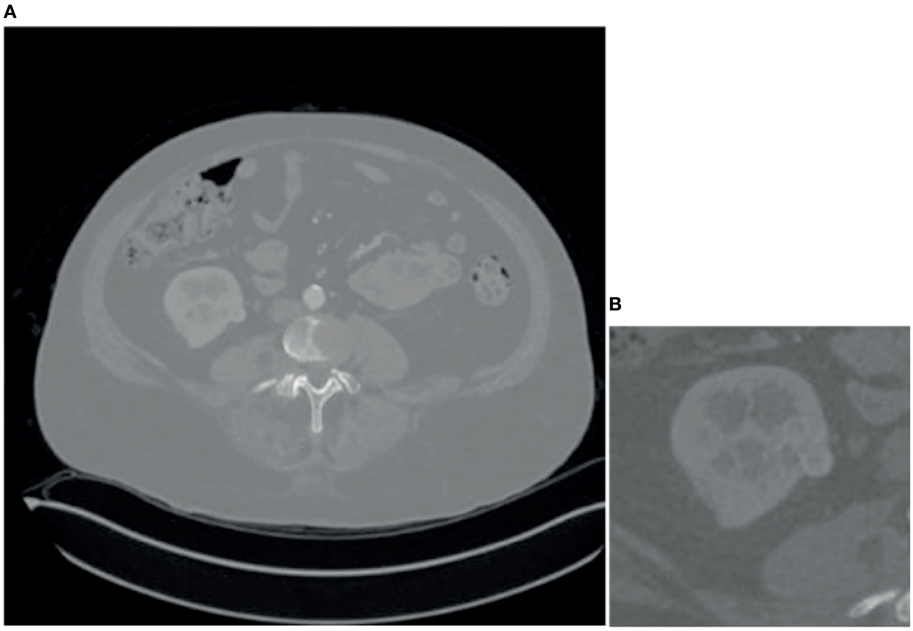


FIGURE 6 | Result of coarse segmentation on CT image of 512 × 512 pixels to 128 × 128 pixels, **(A)** is CT image of 512 × 512 pixels, **(B)** is CT image of 128 × 128 pixels.

in each step of the proposed process in detail. In addition to this, we also provide a series of case studies and a comparative analysis with the relevant literature.

Evaluation Indicators

To measure the accuracy of our method, we use metrics commonly used in CAD/CADx systems to evaluate the classification and segmentation methods of medical images (30). The metric used is the Dice similarity coefficient. It measures the spatial similarity or overlap between two segments and is commonly used to evaluate the ground truth and segmentation performance of the medical images. Equation (4) and **Figure 9** shows the calculation method of DSC.

$$\overline{DSC} = \frac{1}{n}\Sigma_{i=1}^{n}\frac{2|A_i \cap B_i|}{|A_i| + |B_i|} i = 1,\ldots 2, n \tag{4}$$

This article randomly selected 200 CT images for testing, and the rest was used as the training set. To avoid that a particular image area is equal to 0 and cannot calculate the formula, we add 1 to the numerator and denominator of the calculation formula (4). Therefore, the Dice calculation method is changed to Equation (5):

$$\overline{Kidney(Tumor)\ Dice} = \frac{1}{200}\sum_{i=0}^{199}\frac{2|A_i \cap B_i| + 1}{|A_i| + |B_i| + 1} i = 0, 1, 2\ldots, 199 \tag{5}$$

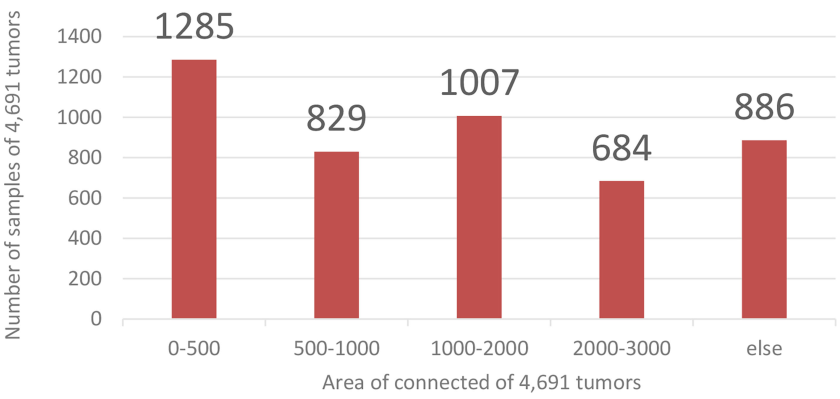


FIGURE 7 | Training data distribution. The abscissa is the Area of connection of 4,691 tumors, and the ordinate is the number of samples of 4,691 tumors.

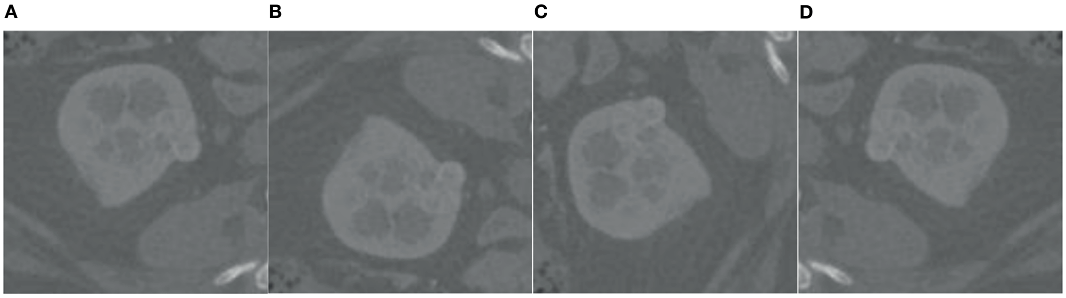


FIGURE 8 | Data enhancement. **(A)** is the original kidney ROI image, **(B)** is the result of horizontal flipping, **(C)** is the result of vertical flipping, and **(D)** is the result of rotating.

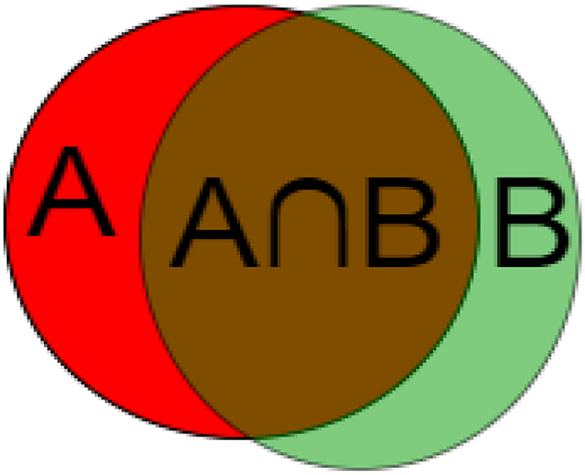


FIGURE 9 | Calculation method of DSC. **(A)** segmentation result, **(B)** label.

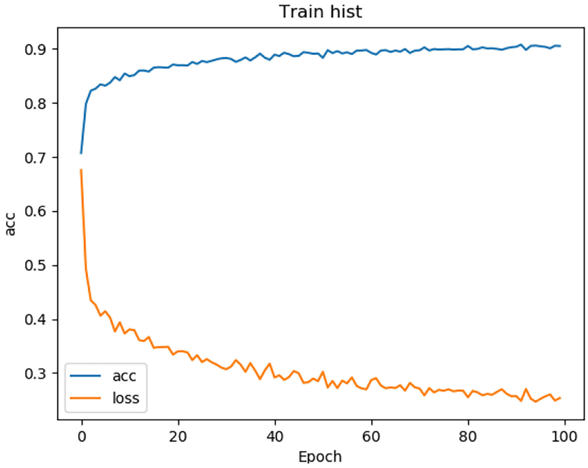


FIGURE 10 | Training result (U-Net).

Among them, A_i is the *i*-th segmented area, and B_i is the *i*-th label image, $\overline{Kidney(Tumor)\ Dice}$ is the average of n results.

Experimental Results

In our experiments, we used the CT data described in *Data Preparation*. Our model is trained by an Adam optimizer, and the learning coefficient is set to 0.001. The batch size is set to 8 and the total epoch is formed to 500,000 (steps_per_epoch = 500, epochs = 100). This model is trained on NVIDIA GeForce RTX 3060 (12GB) graphics processing unit (GPU).

We tested the renal tumor segmentation results of multiple models on the same dataset to verify the effectiveness of the FR2PAttU-Net model for image segmentation. The U-Net model training and segmentation results are saved in **Figure 10** and **Table 1**, and the R2AttU-Net model training and segmentation results are saved in **Figure 11** and **Table 2**. **Figures 12**, **13** are the training results of FR2PAttU-Net using various convolutions, and **Tables 3**, **4** are the segmentation results of FR2PAttU-Net using various convolutions.

Tables 1–**4** are results of fine segmentation. That is, the input image is 128 × 128. Each table has six columns, input image pixel size, last layer image pixel size, total training time, kidney Dice, tumor Dice, and Composite score. With the deepening of the network, the image pixels of the previous layer gradually decrease until the GPU Terminates the experiment when out of memory is displayed. Comparing **Tables 1**–**4**, we find that with the deepening of the model, the training time of the model will be longer and longer, but our model can still extract better feature information. Furthermore, performing multiple convolution operations on the image in parallel can obtain different information about the input image than consecutive convolution operations; processing these operations in parallel and combining all the results will result in better image representation, resulting in a better tumor segmentation.

Figure 14 shows the overall segmentation effect based on the FR2PAttU-Net model (convolution kernel = 3 × 3 and 5 × 5) on the kidney CT images of three patients. Each patient shows five pictures, among which, A is the original image, B is the label, C is the coarse segmentation result, D is the label of ROI, and E is the fine segmentation result. **Figure 14-1** is the first type of case; the tumor and kidney are more prominent, a relatively common type. **Figure 14-2** shows the results of the second type of case. In

TABLE 1 | Fine segmentation based on U-Net model.

Input image size (pixel)	Last layer image size (pixel)	Total training time	Kidney Dice	Tumor Dice	Composite score
128 × 128	8 × 8	About 500 s	0.391	0.456	0.424
128 × 128	4 × 4	About 700 s	0.472	0.415	0.444
128 × 128	2 × 2	About 1,100 s	0.583	0.460	0.522
Average			0.482	0.444	0.463

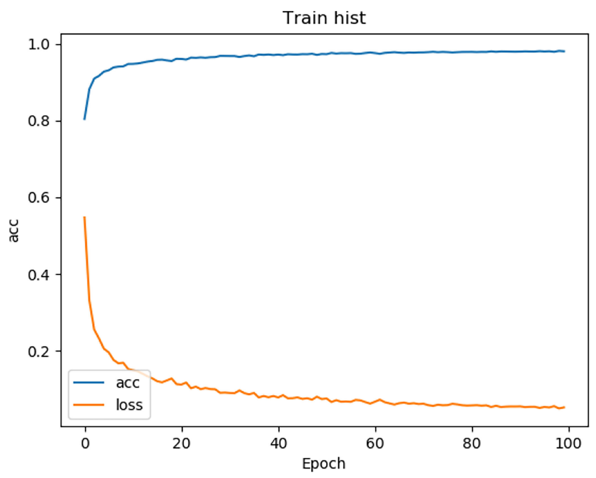


FIGURE 11 | Training result (R2AttU-Net).

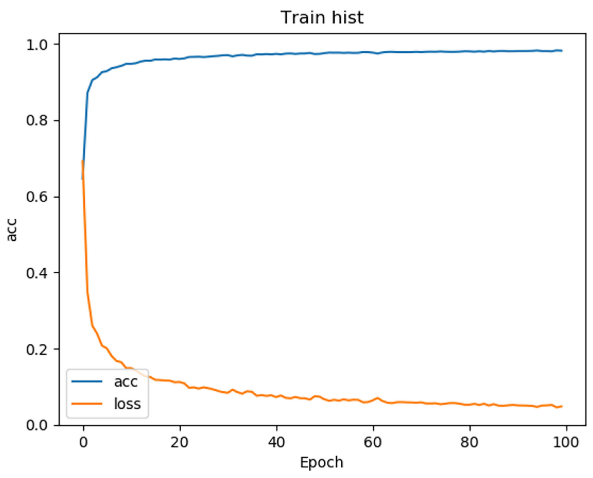


FIGURE 12 | Training result [FR2PAttU-Net (kernel = 3)].

this case, both the kidney and tumor area are small, and the tumor is blurred, making it difficult to distinguish with the human eye directly. Finally, **Figures 14-3**, **14-4** are the third types of cases in which both kidneys have tumors, our model detects two tumors separately, and two ROI regions are extracted from the image.

Recreating the anatomy of the patient in CT images is a significant problem (31). We can post-process the CT image of the patient after the kidney tumor segmentation is completed so that the doctor can observe the spatial structure of the kidney and tumor of the patient. **Figure 15** shows the post-processing process. In the fine segmentation stage, we use an image of 128 × 128 pixels, so the segmentation result is also 128 × 128 pixels. We constructed a marked ROI region for the segmentation results of kidney and tumor (ROI 1, ROI 2). The background pixels remained unchanged and converted the pixels of the kidney and tumor into pixels of the segmentation result. The ROI area is then matched to the CT image of the patient (512 × 512 pixels),

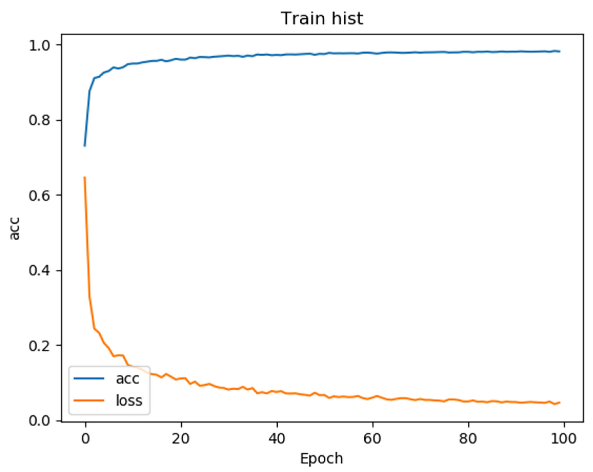


FIGURE 13 | Training result [FR2PAttU-Net (kernel = 3 and 5)].

TABLE 2 | Fine segmentation based on R2AttU-Net model.

Input image size (pixel)	Last layer image size (pixel)	Total training time	Kidney Dice	Tumor Dice	Composite score
128 × 128	8 × 8	About 1,500 s	0.906	0.836	0.871
128 × 128	4 × 4	About 2,000 s	0.925	0.858	0.892
128 × 128	2 × 2	About 3,700 s	0.921	0.867	0.894
Average			0.917	0.854	0.886

TABLE 3 | Fine segmentation based on FR2PAttU-Net model (parallel convolutions [convolution kernel = 3 × 3)].

Input image size (pixel)	Last layer image size (pixel)	Total training time	Kidney Dice	Tumor Dice	Composite score
128 × 128	8 × 8	About 2,200 s	0.948	0.906	0.927
128 × 128	4 × 4	About 3,000 s	0.929	0.902	0.916
128 × 128	2 × 2	About 6,300 s	0.951	0.915	0.933
Average			0.943	0.908	0.926

TABLE 4 | Fine segmentation based on FR2PAttU-Net model [parallel convolutions (convolution kernel = 3 × 3 and 5 × 5)].

Input image size (pixel)	Last layer image size (pixel)	Total training time	Kidney Dice	Tumor Dice	Composite score
128 × 128	8 × 8	About 2,700 s	0.948	0.914	0.931
128 × 128	4 × 4	About 4,400 s	0.951	0.913	0.932
128 × 128	2 × 2	About 11,400 s	0.946	0.905	0.926
Average			0.948	0.911	0.930

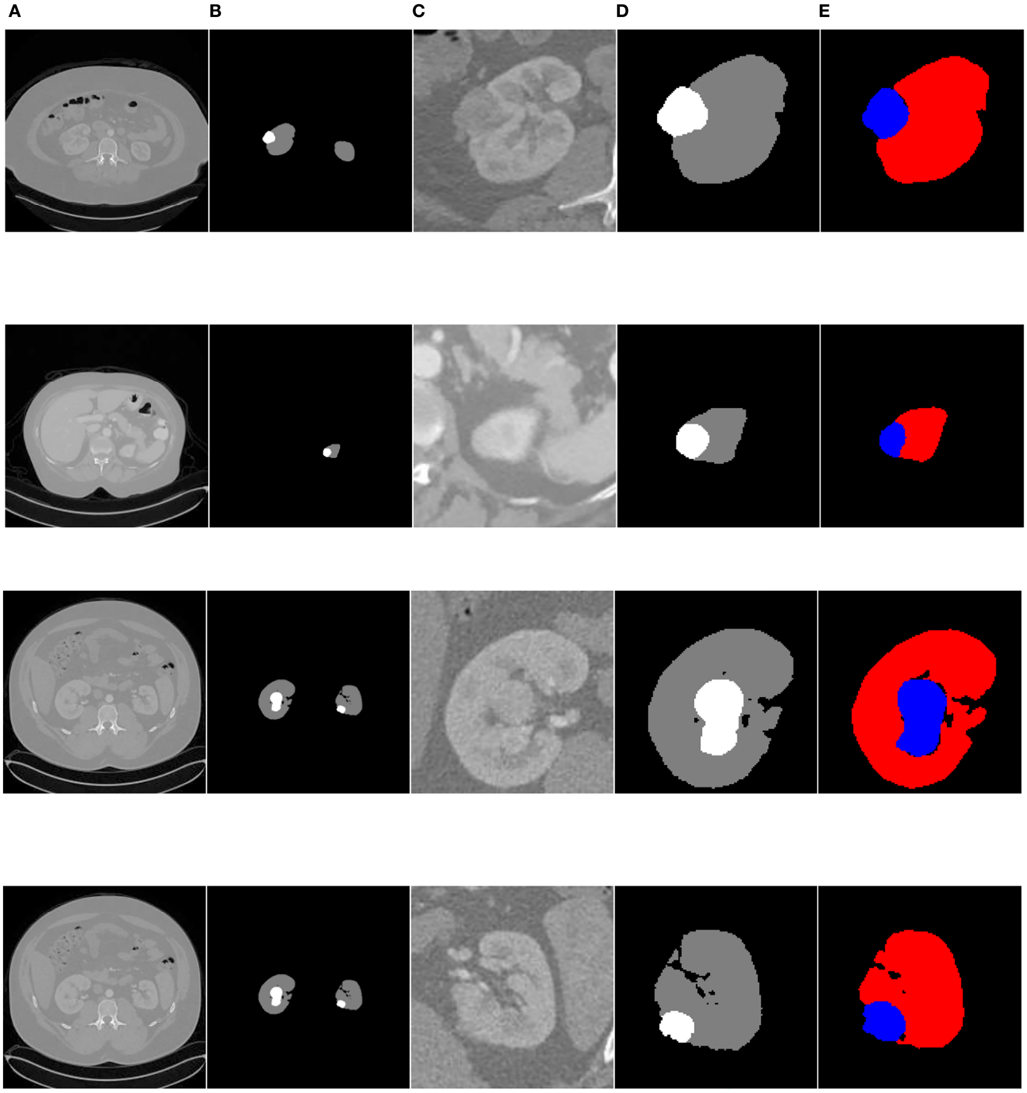


FIGURE 14 | Kidney tumor segmentation based on FR2PAttU-Net model. **(A)** Original image, **(B)** Label, **(C)** Result of coarse segmentation, **(D)** label of ROI, **(E)** Result of fine segmentation.

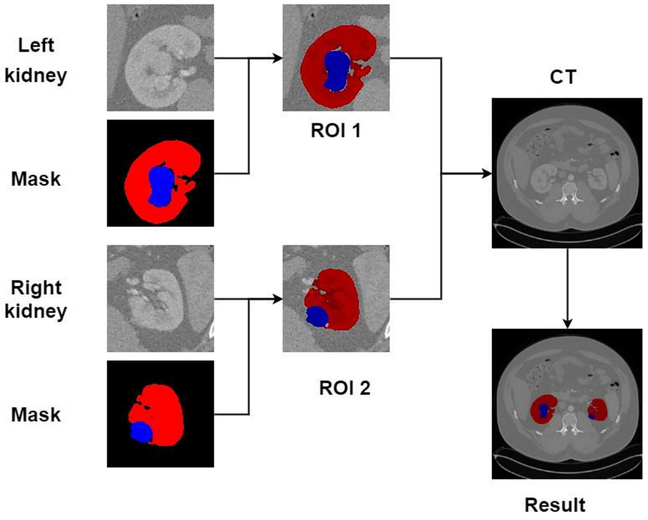


FIGURE 15 | Post-processing.

showing the specific location of the kidney and tumor of the patient, which is convenient for expert diagnosis and observation.

DISCUSSION AND CONCLUSIONS

Many deep learning methods have been used for kidney and tumor segmentation in the past few years. **Figure 14** can intuitively see that the FR2PAttU-Net model proposed in this paper is used for the segmentation effect of kidneys and tumors. **Table 5** shows the average Dice calculated by some algorithms or methods. Among them, the data used by FR2PAttU-Net, U-Net, ResU-Net, AttU-Net, R2U-Net, R2AttU-Net, and nnU-Net are precisely the same. It is the data introduced in *Data Preparation*, and the methods of data preprocessing and data enhancement are the same. The other models from References (17, 18) and (20–24) use the KiTS19 dataset, but the FR2PAttU-Net model uses fuzzy sets to enhance the image. Therefore, we directly quoted their results without additionally testing the performance of our data on their model. As a result, we get scored a 0.948 kidney Dice and a 0.911 tumor Dice resulting in a 0.930 composite score; in the case of this test, the effect is better than U-Net, ResU-Net, AttU-Net, R2U-Net, R2AttU-Net, nnU-Net. However, our kidney Dice is about 0.2 lower when compared to other algorithms. Still, tumor Dice is about 0.4 higher, which means that the proposed method can simultaneously pay attention to the more prominent feature (kidney) and more minor features (tumors). It proves that the parallel convolution method has a particular segmentation effect and research value in kidney and tumor segmentation.

In conclusion, this paper proposes a kidney tumor segmentation model based on FR2PAttU-Net, which can effectively segment kidney tumors. This method is a cascade deep learning model, adding residual-recurrent-parallel convolutional networks, attention gates, Leaky-ReLU, and a 20% batch normalization layer to the original U-shaped structure of the U-Net. We also use an Image enhancement algorithm with fuzzy sets to alter the input image pixels to

TABLE 5 | Segmentation results of several algorithms or methods.

References	Algorithms or methods	Kidney Dice	Tumor Dice	Composite score
This paper	FR2PAttU-Net	0.948	0.911	0.930
Reference (11)	U-Net	0.482	0.444	0.463
Reference (12)	ResU-Net	0.688	0.694	0.691
Reference (13)	AttU-Net	0.789	0.735	0.763
Reference (14)	R2U-Net	0.681	0.711	0.696
Reference (15)	R2AttU-Net	0.917	0.854	0.886
Reference (16)	nnU-Net	0.905	0.864	0.882
Reference (17)	AlexNet+ U-Net	0.9303	\	0.9303
Reference (18)	Hybrid V-Net	0.977	0.865	0.921
Reference (20)	Cascaded U-Net ensembles	0.973	0.825	0.899
Reference (21)	Cascaded volumetric convolutional network	0.974	0.831	0.902
Reference (22)	multi-resolution VB-nets	0.973	0.832	0.903
Reference (23)	Cascaded semantic segmentation	0.967	0.845	0.906
Reference (24)	3d U-net based on five-fold cross-validation	0.974	0.851	0.912

improve the robustness of the model. The FR2PAttU-Net model increases the width of the model and enhances the adaptability of the model to the features of different image scales, and obtains an excellent segmentation effect in the kidney CT image. In future work, we will collect more medical data for validating the reliability of the FR2PAttU-Net model.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Data curation, ZM and TM. Formal analysis, FH and YZ. Investigation, FL. Methodology, PS. Validation, ZC. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (81873913), the National Major Instrument Development Project (61627807), and the Science and Technology Major Project of Guangxi (2019AA12005).

REFERENCES

1. Huang WC, Donin NM, Levey AS, Campbell SC. Chronic Kidney Disease and Kidney Cancer Surgery: New Perspectives. *J Urol* (2020) 203(3):475–85. doi: 10.1097/JU.0000000000000326
2. Checcucci E, De Cillis S, Granato S, Chang P, Andrew Shea A, Okhunov Z, et al. Applications of Neural Networks in Urology: A Systematic Review. *Curr Opin Urol* (2020) 30(6):788–807. doi: 10.1097/MOU.0000000000000814
3. Checcucci E, De Cillis S, Granato S, Chang P, Afyouni AS, Okhunov Z. Uro-Technology and SoMe Working Group of the Young Academic Urologists Working Party of the European Association of Urology. *Artif Intell Neural Networks Urol: Curr Clin Appl Minerva Urol Nefrol* (2020) 72(1):49–57. doi: 10.23736/S0393-2249.19.03613-0
4. Lund CB, van der Velden BHM. Leveraging Clinical Characteristics for Improved Deep Learning-Based Kidney Tumor Segmentation on CT. *arXiv* (2021) 2109:5816. doi: 10.48550/arXiv.2109.05816
5. Lin DT, Lei CC, Hung SW. Computer-Aided Kidney Segmentation on Abdominal CT Images. *IEEE Trans Inf Technol Biomed* (2006) 10(1):59–65. doi: 10.1109/TITB.2005.855561
6. Thong W, Kadoury S, Piché N, Pal CJ. Convolutional Networks for Kidney Segmentation in Contrast-Enhanced CT Scans. *Comput Methods Biomech Biomed Eng: Imaging Visualization* (2018) 6(3):277–82. doi: 10.1080/21681163.2016.1148636
7. Zöllner FG, Kociński M, Hansen L, Golla AK, Trbalić AŠ, Lundervold A, et al. Kidney Segmentation in Kidney Magnetic Resonance Imaging-Current Status and Prospects. *IEEE Access* (2021) 9:71577–605. doi: 10.1109/ACCESS.2021.3078430
8. Li L, Ross P, Kruusmaa M, Zheng X. A Comparative Study of Ultrasound Image Segmentation Algorithms for Segmenting Kidney Tumors. *Proc 4th Int Symp Appl Sci Biomed Commun Technol* (2011), 1–5. doi: 10.1145/2093698.2093824
9. Kim T, Lee K, Ham S, Park B, Lee S, Hong D, et al. Active Learning for Accuracy Enhancement of Semantic Segmentation With CNN-Corrected Label Curations: Evaluation on Kidney Segmentation in Abdominal CT. *Sci Rep* (2020) 10(1):1–7. doi: 10.1038/s41598-019-57242-9
10. Costantini F, Kopan R. Patterning a Complex Organ: Branching Morphogenesis and Nephron Segmentation in Kidney Development. *Dev Cell* (2010) 18(5):698–712. doi: 10.1016/j.devcel.2010.04.008
11. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2015) p. 234–41.
12. Yang Y, Li Q, Guo Y, Liu Y, Li X, Guo J, et al. Lung Parenchyma Parameters Measure of Rats From Pulmonary Window Computed Tomography Images Based on ResU-Net Model for Medical Respiratory Researches[J]. *Math Biosci Eng* (2021) 18(4):4193–211. doi: 10.3934/mbe.2021210
13. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* (2018). 1804.03999. doi: 10.48550/arXiv.1804.03999
14. Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK. Recurrent Residual U-Net for Medical Image Segmentation[J]. *J Med Imaging* (2019) 6(1):014006. doi: 10.1117/1.JMI.6.1.014006
15. Wang Y, He Z, Xie P, Yang C, Zhang Y, Li F, et al. *Segment Medical Image Using U-Net Combining Recurrent Residuals and Attention[C]//International Conference on Medical Imaging and Computer-Aided Diagnosis*. Springer, Singapore, 2020: 77-86.
16. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. Nnu-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation. (2018) 9:1809.10486. doi: 10.48550/arXiv.1809.10486
17. da Cruz LB, Araujo JDL, Ferreira JL, Diniz JOB, Silva AC, de Almeida JDS, et al. Kidney Segmentation From Computed Tomography Images Using Deep Neural Network. *Comput Biol Med* (2020) 123:103906. doi: 10.1016/j.compbiomed.2020.103906
18. Turk F, Luy M, Barisci N. Kidney and Kidney Tumor Segmentation Using a Hybrid V-Net-Based Model. *Mathematics* (2020) 8(10):1772. doi: 10.3390/math8101772
19. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, et al. The State of the Art in Kidney and Kidney Tumor Segmentation in Contrast-Enhanced CT Imaging: Results of the KiTS19 Challenge[J]. *Med Image Anal* (2021) 67:101821. doi: 10.1016/j.media.2020.101821
20. Ma J. Solution to the Kidney Tumor Segmentation Challenge 2019. *Submissions to the Kidney and Kidney Tumor Segmentation Challenge 2019*. (2019).
21. Zhang Y, Wang Y, Hou F, Yang J, Xiong G, Tian J, et al. Cascaded Volumetric Convolutional Network for Kidney Tumor Segmentation From CT Volumes. *Electr Eng Syst Sci* (2020) 5:arXiv.1910.02235. doi: 10.48550/arXiv.1910.02235
22. Mu G, Lin Z, Han M, Yao G, Gao Y. Segmentation of Kidney Tumor by Multi-Resolution VB-Nets. *Submissions to the Kidney and Kidney Tumor Segmentation Challenge 2019*. (2019).
23. Hou X, Xie C, Li F, Nan Y. Cascaded Semantic Segmentation for Kidney and Tumor. *Submissions to the Kidney and Kidney Tumor Segmentation Challenge 2019*. (2019).
24. Isensee F, Maier-Hein KH. An Attempt at Beating the 3D U-Net. *arXiv* (2019) 10:1908.02182. doi: 10.48550/arXiv.1908.02182
25. Jebadass JR, Balasubramaniam P. Low Contrast Enhancement Technique for Color Images Using Interval-Valued Intuitionistic Fuzzy Sets With Contrast Limited Adaptive Histogram Equalization. *Soft Comput* (2022) 1:1–12. doi: 10.1007/s00500-021-06539-x
26. Han M, Liu H. Super-Resolution Restoration of Degraded Image Based on Fuzzy Enhancement. *Arabian J Geosci* (2021) 14(11):1–7. doi: 10.1007/s12517-021-07218-9
27. Mastromichalakis S. ALReLU: A Different Approach on Leaky ReLU Activation Function to Improve Neural Networks Performance. *arXiv* (2020) 2012.07564. doi: 10.48550/arXiv.2012.0756
28. *The KiTS19 Grand Challenge* (2020). Available at: https://kits19.grand-challenge.org/data/ (Accessed on 13 October 2020).
29. Shi B, Akbari P, Pourafkari M, Iliuta IA, Guiard E, Quist CF, et al. Prognostic Performance of Kidney Volume Measurement for Polycystic Kidney Disease: A Comparative Study of Ellipsoid vs. Manual Segmentation. *Sci Rep* (2019) 9(1):1–8. doi: 10.1038/s41598-019-47206-4

30. Tanabe Y, Ishida T, Eto H, Sera T, Emoto Y. Evaluation of the Correlation Between Prostatic Displacement and Rectal Deformation Using the Dice Similarity Coefficient of the Rectum. *Med Dosim* (2019) 44(4):e39–43. doi: 10.1016/j.meddos.2018.12.005

31. Porpiglia F, Bertolo R, Checcucci E, Amparore D, Autorino R, Dasgupta P, et al. Development and Validation of 3D Printed Virtual Models for Robot-Assisted Radical Prostatectomy and Partial Nephrectomy: Urologists' and Patients' Perception. *World J Urol* (2018) 36(2):201–7. doi: 10.1007/s00345-017-2126-1

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 30 March 2022
doi: 10.3389/fonc.2022.842759

Exploring Histological Similarities Across Cancers From a Deep Learning Perspective

Ashish Menon[1†], Piyush Singh[1†], P. K. Vinod[2] and C. V. Jawahar[1]*

[1] Center for Visual Information Technology, International Institute of Information Technology (IIIT) Hyderabad, Hyderabad, India, [2] Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology (IIIT) Hyderabad, Hyderabad, India

Histopathology image analysis is widely accepted as a gold standard for cancer diagnosis. The Cancer Genome Atlas (TCGA) contains large repositories of histopathology whole slide images spanning several organs and subtypes. However, not much work has gone into analyzing all the organs and subtypes and their similarities. Our work attempts to bridge this gap by training deep learning models to classify cancer vs. normal patches for 11 subtypes spanning seven organs (9,792 tissue slides) to achieve high classification performance. We used these models to investigate their performances in the test set of other organs (cross-organ inference). We found that every model had a good cross-organ inference accuracy when tested on breast, colorectal, and liver cancers. Further, high accuracy is observed between models trained on the cancer subtypes originating from the same organ (kidney and lung). We also validated these performances by showing the separability of cancer and normal samples in a high-dimensional feature space. We further hypothesized that the high cross-organ inferences are due to shared tumor morphologies among organs. We validated the hypothesis by showing the overlap in the Gradient-weighted Class Activation Mapping (GradCAM) visualizations and similarities in the distributions of nuclei features present within the high-attention regions.

Keywords: TCGA, cross-organ inference, tissue morphology, class activation map (CAM), histopathology, deep learning, cancer classification

OPEN ACCESS

Edited by:
Abhishek Mahajan,
Tata Memorial Hospital, India

Reviewed by:
Zhiqiang Liu,
Tianjin Medical University, China
Tao Huang,
Shanghai Institute of Nutrition and Health (CAS), China

****Correspondence:***
P. K. Vinod
vinod.pk@iiit.ac.in

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *24 December 2021*
Accepted: *22 February 2022*
Published: *30 March 2022*

Citation:
Menon A, Singh P, Vinod PK and Jawahar CV (2022) Exploring Histological Similarities Across Cancers From a Deep Learning Perspective. Front. Oncol. 12:842759. doi: 10.3389/fonc.2022.842759

1 INTRODUCTION

Cancers originating from different organs and cell types are known, with the most common ones being breast, lung, colorectal, prostate, and stomach. The most common causes of cancer deaths are lung, colorectal, and liver (1). Pan-cancer omics studies have revealed commonalities in driver mutations, altered pathways, and immune signatures (2, 3). Molecular profiling helps to cluster and distinguish different cancers and their subtypes by different computational methods (4–7). Given the diverse nature of different cancers and their origin, it will also be interesting to examine the morphological patterns that are unique and shared across different cancers from the histopathological standpoint. Histopathology continues to play a crucial role in cancer diagnostics. Digitization of tissue samples as whole slide images (WSIs) enables computer-based diagnosis and analysis. The deep learning approaches can be used to analyze the cancerous and non-cancerous patterns present in these tissues.

Deep learning has significantly improved the accuracy of a wide variety of computer vision tasks. The success of convolutional neural networks (CNNs) in the ImageNet Large Scale Visual Recognition Competition (8) resulted in a widespread adoption of CNNs for the task of image recognition, object detection, and image retrieval in several fields. Different studies show the effectiveness of CNNs and the utility of models with ImageNet pretrained weights in analyzing the tissue (9–14). Coudary et al. (12) extracted 512 × 512 non-overlapping patches of whole slide tissue images as input image patches for the WSI. The method rejected all the background and noisy patches with a mean intensity of half of the pixels greater than a set threshold. An ImageNet pretrained Inception-v3 (15) network was finetuned for the classification of cancerous and non-cancerous lung tissue slides. Tabibu et al. (11) extended the same idea to the renal cell carcinomas and performed cancer vs. normal classification and subtype classification by finetuning the entire ResNet-18,34 (16) networks and reported both slide-wise and patch-wise results. Wang et al. (13) adopted a threshold-based segmentation for background region detection by operating on the Hue Saturation Value (HSV) color space to get the required mask for patch filtering and identified the regions of metastatic breast cancer using ImageNet pretrained GoogLeNet (17). Xu et al. (10) performed classification and segmentation tasks on brain and colon pathological images using CNNs for feature extraction and training using a fully connected network (FCN). There are also few attempts to perform pan-cancer analysis using a deep learning approach. Fu et al. (18) have used features from models trained for cancer vs. normal classification task to predict genomic, molecular, and prognostic associations across organs. Cheerla et al. (19) have used multimodal learning to predict survival from genetic data as well as histopathology images across organs. Noorbakhsh et al. (20) have reported the correlation of organs based on the slide-wise area under the receiver operating characteristic curve (ROC-AUC). In this work, training is performed at the patch level, and inference is made at the slide level using a threshold for the fraction of patches in a slide predicted as cancerous. They used inception v3 (15) by using the CNN as a feature extractor and finetuning the last fully connected layer. They also performed hierarchical clustering of slide-wise ROC-AUC scores across organs and showed correlations of logits of the models of specific organs to suggest shared tumor morphology. We took this a step further to analyze cross-organ correlations quantitatively as well as qualitatively.

The contribution of this work is three-fold:

- Analyze each slide at the patch level and report high patch-level cancer vs. normal accuracies to set high benchmarks.
- Reveal tumor similarities between certain groups of organs/ subtypes using patch-level analysis of WSIs from a deep learning perspective.
- Demonstrate the consistencies of these correlations both qualitatively and quantitatively, which is the first of its kind to our knowledge.

We reported the self organ classification results with AUC, F1 score, and accuracies for 11 cancer subtypes and the best and worst cross-organ inference results for each of these trained models.

In the cross-organ inference, the trained models are used for inference on the images of the other organs. The t-distributed stochastic neighbor embedding (t-SNE) (21) plot of embeddings obtained from each trained model shows the separability of cancer and normal features across organs. The GradCAM visualization of each trained model tested on the patches of other organs supports the cross-organ performance between a specific pair of organs, indicating the presence of common morphological patterns. We showed that the distributions of the nucleus features present in the high-attention regions for pairs with good cross-organ performance are well aligned compared to those with poor cross-organ performance. A uniform workflow which performs satisfactorily across organs is established. This includes patch extraction from tissue-rich regions of WSI based on intensity values and connected components present in its binarized format, hyperparameter tuning (using Bayesian optimization) to decide on the model architecture.

2 METHOD

2.1 Dataset and Preprocessing

We used the publicly available data set of WSIs from TCGA project (22) across multiple organs. Experiments were performed using the formalin-fixed paraffin-embedded (FFPE) slides. As pointed out by (23), the FFPE sections reveal useful cellular details of the tissue. These slides can confirm the diagnosis, in contrast to the frozen slides that can affect the morphological features of the tissue. 9,792 whole slide images spanning seven organs, namely, breast, colorectal, kidney, liver, lung, prostate, and stomach, were used. Some of these organs have multiple subtypes: lung [lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC)], kidney [kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), and kidney chromophobe (KICH)], and colorectal [colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ)]. We also considered cancer images specific to breast [breast invasive carcinoma (BRCA)], stomach [stomach adenocarcinoma (STAD)], liver [liver hepatocellular carcinoma (LIHC)], and prostate [prostate adenocarcinoma (PRAD)]. The number of slides and images considered in this study are shown in **Figure 1**.

H&E-stained WSI contains several cells and comprises as many as tens of billions of pixels, which is computationally infeasible for training neural networks. Resizing the entire image to a smaller size would hamper the cellular-level details, resulting in lower classification performance (24). Therefore, the entire WSI is commonly divided into partial patches or tiles analyzed independently. We adopted the strategy mentioned in Coudary et al. (12), by extracting 512 × 512-sized patches with no overlap at a ×20 magnification. The patch-filtering method of (11) was used to filter out background and noisy patches. We also added another patch-filtering step to avoid patches with a fractal

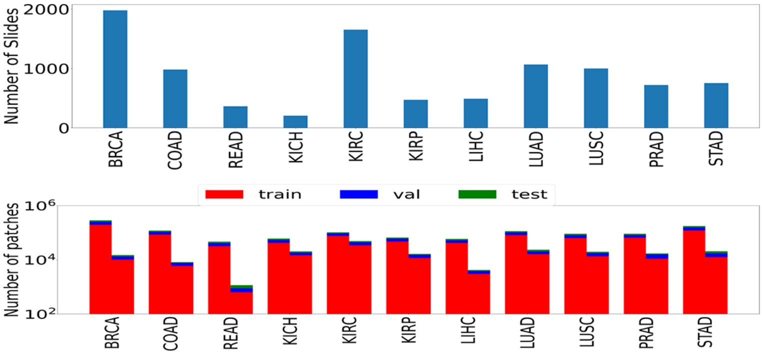


FIGURE 1 | The number of slides (top) and patches (bottom) used in the study. Numbers of patches belonging to both classes (left bar represents cancer samples and right bar represents normal samples) are shown in the form of two rectangular bar plots.

structure by considering only those patches with ten or more connected components present in its binarized format. Since patch-wise labels were not available for TCGA dataset, the slide label was assigned to patches as shown to be effective by (11, 12). A train-validation-test split of 70–20–10 was performed before training the models. Data augmentation techniques such as random horizontal flip and random crop were used to improve generalizability. The images were normalized using the mean and standard deviation across all the three (RGB) channels calculated on the training set.

2.2 Cancer vs. Normal Classification

We trained one model for each of the eleven subtypes (eleven models in total) using a ResNet-18 architecture pretrained on the ImageNet dataset. The ResNet style of architecture has performed well compared to other computer vision models on the ImageNet dataset (16). ResNet-18 was chosen over other models (ResNet-34,50,101) since 18 layers were found sufficient to yield superior performance in the classification tasks across most cancers, and a further increase in the number of layers led to a marginal increase in performance at the expense of a large increase in the number of trainable parameters. The schematic flow diagram is shown in **Figure 2** for the classification task. We replaced the last layer of ResNet-18 which provided the logits for the thousand classes of the ImageNet classification task with a fully connected network (FCN). The size of the last layer of this FCN was fixed at two since the task was a binary classification.

The entire network parameters were optimized to minimize the cross-entropy loss on the train data *via* backpropagation. The optimizer, learning rate, number of FCN layers, number of neurons in each layer, and dropout probabilities for each FCN layer were chosen by a hyperparameter search using Bayesian

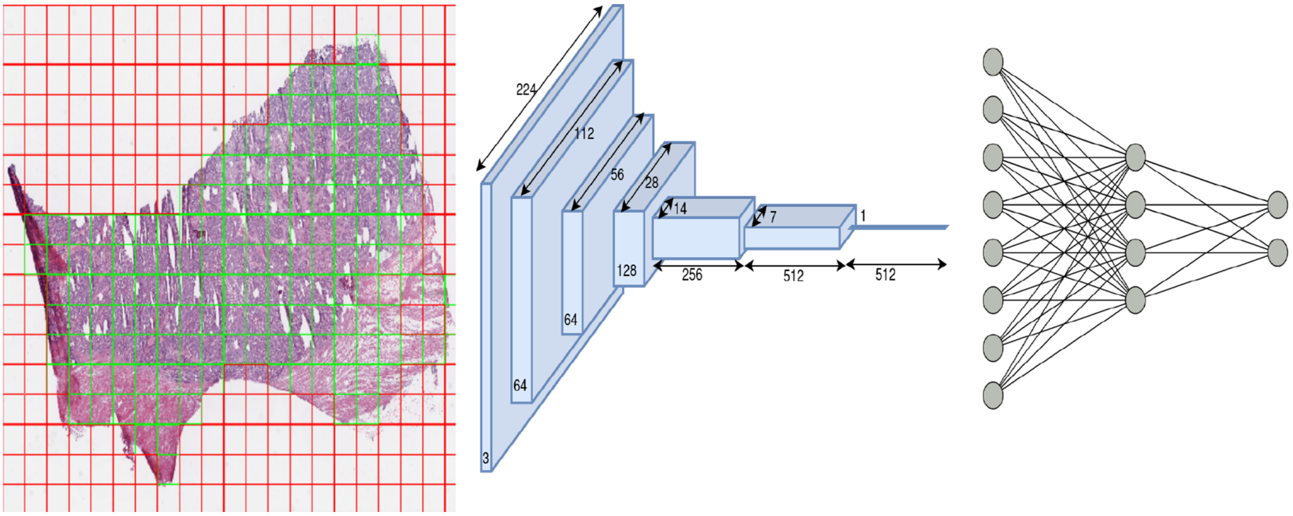


FIGURE 2 | Overview of architecture used in our work: patch extraction (left): red shows rejected background patches, and green shows patches used for the training model, ResNet-18 architecture (middle) and Fully connected network (right).

optimization. The batch size was set to 256. Owing to the class imbalance in the cancer and normal samples across organs, weighted cross entropy was used as the loss function. We also employed a stratified sampling technique to maintain the ratio of positives and negatives.

2.3 Hyperparameter Search

We used Optuna framework (25) for hyperparameter tuning with the search space of the optimizer sampled from a categorical distribution of optimizers (Adam, RMSProp, SGD), learning rate sampled from a log-uniform distribution of values ranging [1*e*−05, 1*e*−01], dropout sampled from a uniform distribution of values from [0.2, 0.5], number of layers of FCN uniformly sampled from values [1, 3], and number of neurons per layer uniformly sampled from values ranging [4, 128]. We ran 20 trials for hyperparameter search, and in each trial we trained the model for 20 epochs. Finally, the optimal hyperparameters that had the maximum validation accuracy across all trials were used to train the model for 50 epochs. We tested the usefulness of hyperparameter tuning on four organs and found a significant improvement in the performance (accuracy, AUC, F1 score). Hence, we adopted the same strategy for all the other organs during the training. The contour plot indicating the hyperparameter tuning is shown in **Supplementary Figure S1**.

2.4 GradCAM Analysis

We used the GradCAM (26) visualization technique to support the cross-organ inference results. We obtained a thresholded GradCAM heatmap and a bounding box over the high-attention region for each of the patches under study. Thresholding of the high-attention regions (green) of the heatmap was done by converting the image to the HSV color space, since the hue channel models the color type and is helpful in segmenting regions based on a specific color criteria. To obtain the bounding box containing the segmented region, we applied canny edge detection to the thresholded image. For each of the obtained contours, we applied closed-polygon approximation followed by finding a rectangular bounding box. We explored through these thresholded and bounding box outputs whether the regions of high saliency have overlap across models trained on different organs. We quantified the overlap by using IoU (intersection over union) of the bounding box representations, with IoU = 1 representing a perfect overlap and IoU = 0 representing no overlap. We also reported the Jaccard index to quantify the overlap using the thresholded pixel maps.

2.5 Nucleus Feature Extraction

Different studies have demonstrated the association of nucleus features to the clinical outcome and molecular data (11, 27–29). We hypothesized that the shared regions between cancers might show similar nucleus shapes and density features due to the similarity in the tumor microenvironment. We used the GradCAM high-attention regions to analyze the geometrical features of the nuclei such as eccentricity, convex area, region solidity, diameter, major axis, and minor axis and graphical features such as Voronoi diagram, Delaunay triangulation, minimum spanning tree, and nucleus density that characterize the arrangement of nuclei. We compared the distributions of these features to comment on the shared tumor morphology. The steps involved are shown in **Figure 3**.

- Region extraction: for the patches under study, we first extracted the high-attention regions corresponding to the model trained using that organ and the high-attention regions of the model trained on the other organ. We extracted three regions, the overlapped area of intersection and areas specific to each of the models. The overlap region was obtained by performing a logical AND operation between the thresholded GradCAM images. Specific regions were obtained by subtracting the overlapped regions from the thresholded GradCAM images.

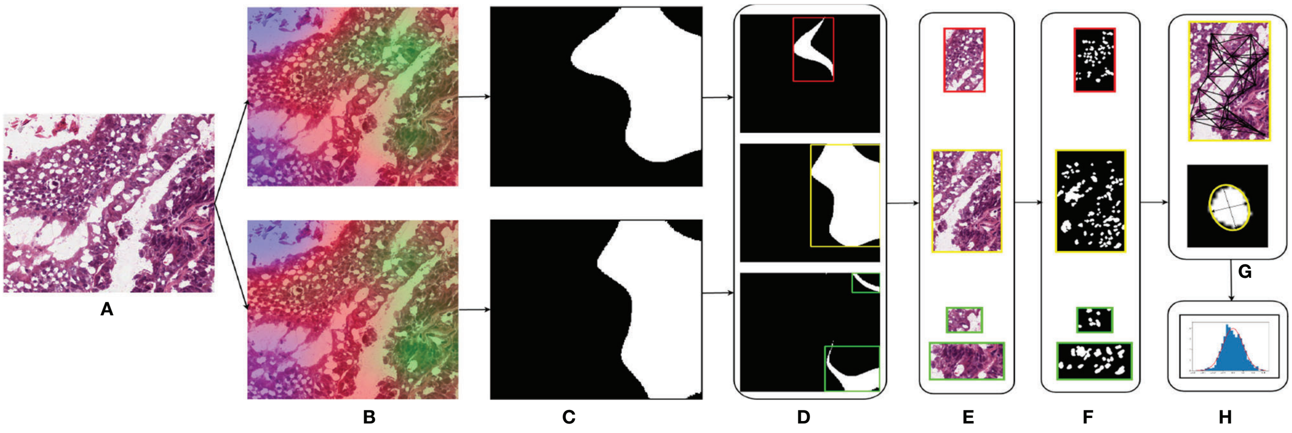


FIGURE 3 | Nucleus segmentation workflow involved in segmenting nuclei from the specific regions of a sample patch: **(A)** COAD sample patch, **(B)** GradCAM outputs of BRCA model (top) and COAD model (bottom), **(C)** thresholded GradCAM mask, **(D)** BRCA-specific mask (top), overlapping mask (middle), and COAD-specific mask (bottom), **(E)** masked regions of BRCA-specific (top), overlap (middle), and COAD-specific (bottom), **(F)** nucleus segmented regions of BRCA-specific (top), overlap (middle), and COAD-specific (bottom), **(G)** obtaining the nucleus shape and graphical features for each region, and **(H)** distributions of these features.

- Nucleus segmentation: for each of the extracted regions, we performed the nucleus segmentation using a hierarchical multilevel thresholding approach (30).
- Nucleus features: we extracted geometrical shape features from the nucleus segmented images using the connected component analysis (11). Inter-nucleus architecture-based features were obtained by using graph-based techniques (31).

3 RESULTS AND DISCUSSION

3.1 Quantitative Analysis

We performed two sets of experiments for the overall analysis. The first experiment was to come up with a trained model for the cancer vs. normal classification task in each of the mentioned organs/subtypes. A high classification performance was observed for most models (**Figure 4**). The second experiment was the cross-organ inference by testing each of these trained models on the held-out test of all the other organs. We report similarities between specific organ pairs based on performance (accuracy > 0.9) (**Figure 5**). Best and worst performances (AUC, F1) for the cross-organ inference are indicated in **Table 1**. The ROC curve for the cross-organ inference is shown in **Figure S2**.

3.2 Cross-Organ Similarities

We found that most models show a good cross-organ inference accuracy when tested on BRCA, LIHC, COAD, and READ (**Figure 5**), which suggests that these cancers may have shared tumor morphologies. Colorectal subtypes (READ and COAD) show similarities with each other along with BRCA and LIHC. These observations on COAD, READ, and BRCA are consistent with the clustering of pan-gynecological and pan-gastrointestinal observed by (20). In contrast, most of the models perform poorly when tested on the kidney (KIRC, KIRP, and KICH) and lung subtypes (LUAD and LUSC). This suggests that kidney and lung cancer subtypes have morphology features localized relative to the organ of origin. The unique characteristics of kidney cancers are also seen with respect to their gene expression pattern as observed in our previous work (32). Interestingly, within cancer subtypes, we also observed that the performance of KICH and KIRP models on KIRC as a test set does not yield comparable performance. This suggests that KIRC has more subtype-specific features that are not present in other subtypes. Although READ and STAD are gastrointestinal cancers, the cross-organ inference is not high using the READ model. We observed that the cross-organ performance is not uniform within adenocarcinomas (LUAD, COAD, PRAD, READ, and STAD).

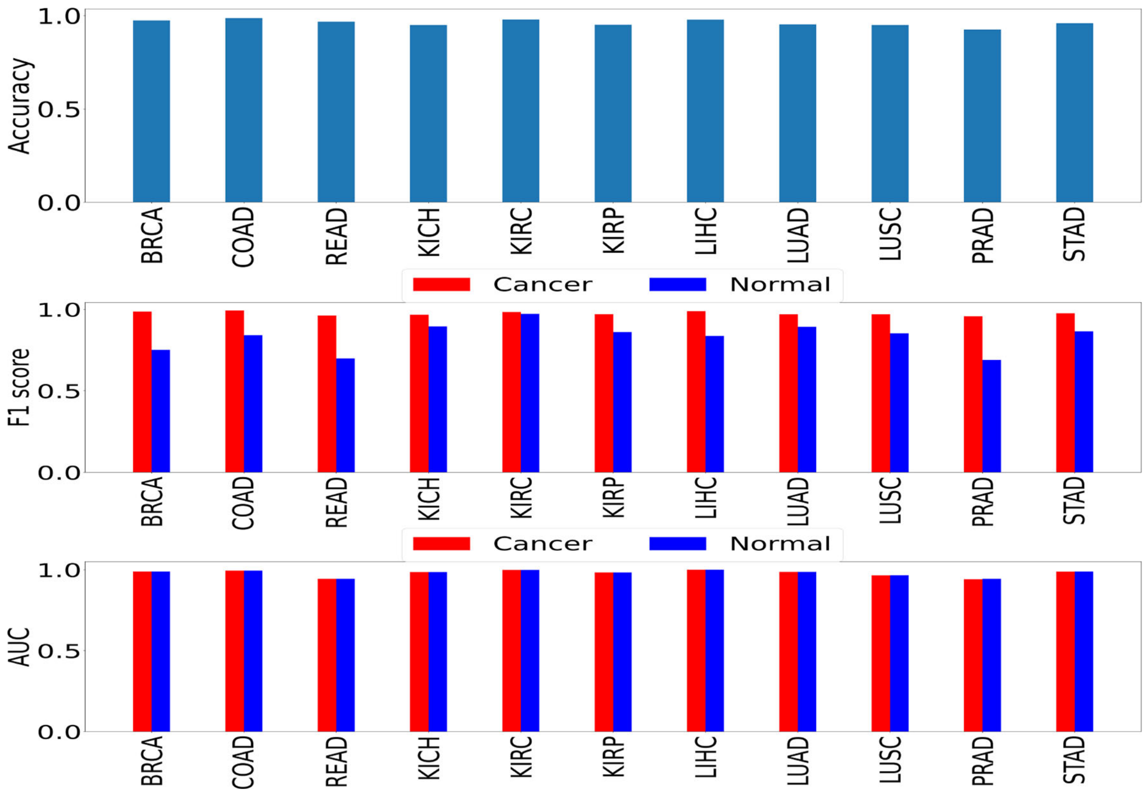
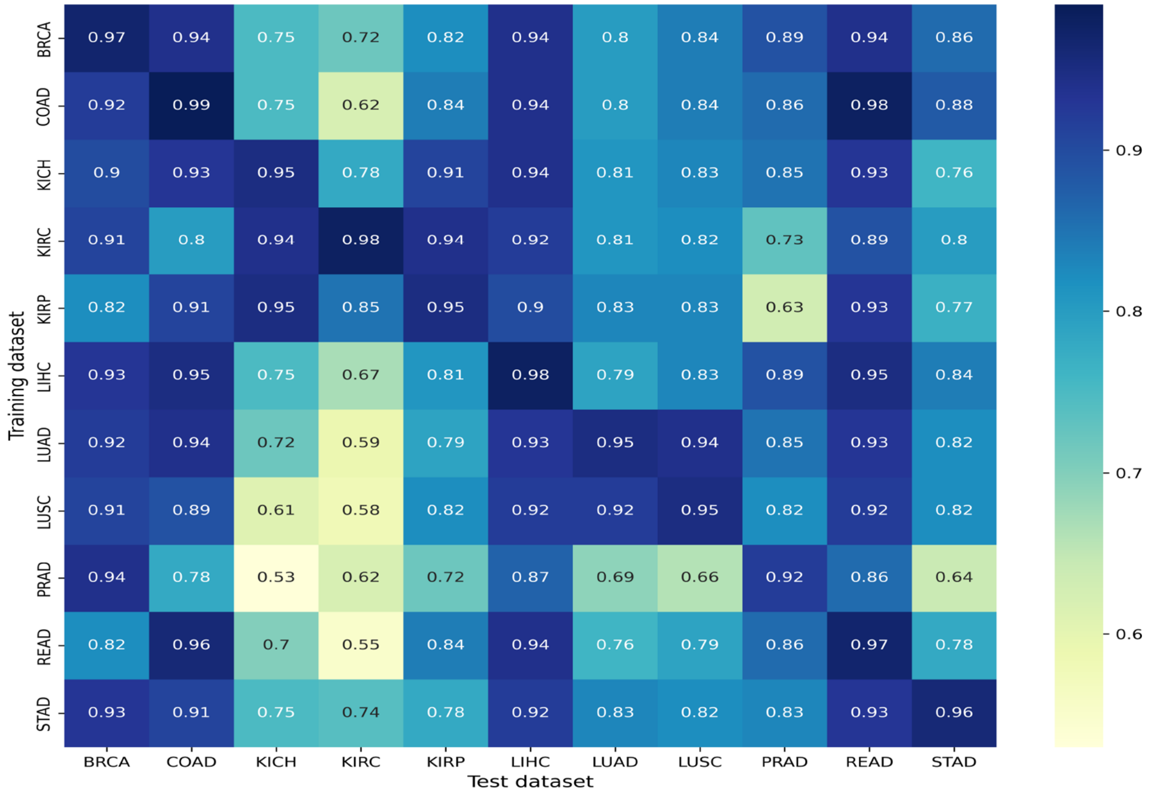


FIGURE 4 | Self-organ inference showing the performance obtained using models trained on each cancer and tested on a held-out test set of the same cancer.



Training dataset \ Test dataset	BRCA	COAD	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	STAD
BRCA	0.97	0.94	0.75	0.72	0.82	0.94	0.8	0.84	0.89	0.94	0.86
COAD	0.92	0.99	0.75	0.62	0.84	0.94	0.8	0.84	0.86	0.98	0.88
KICH	0.9	0.93	0.95	0.78	0.91	0.94	0.81	0.83	0.85	0.93	0.76
KIRC	0.91	0.8	0.94	0.98	0.94	0.92	0.81	0.82	0.73	0.89	0.8
KIRP	0.82	0.91	0.95	0.85	0.95	0.9	0.83	0.83	0.63	0.93	0.77
LIHC	0.93	0.95	0.75	0.67	0.81	0.98	0.79	0.83	0.89	0.95	0.84
LUAD	0.92	0.94	0.72	0.59	0.79	0.93	0.95	0.94	0.85	0.93	0.82
LUSC	0.91	0.89	0.61	0.58	0.82	0.92	0.92	0.95	0.82	0.92	0.82
PRAD	0.94	0.78	0.53	0.62	0.72	0.87	0.69	0.66	0.92	0.86	0.64
READ	0.82	0.96	0.7	0.55	0.84	0.94	0.76	0.79	0.86	0.97	0.78
STAD	0.93	0.91	0.75	0.74	0.78	0.92	0.83	0.82	0.83	0.93	0.96

FIGURE 5 | Cross-organ inference results: accuracies obtained using models trained on the organs along the rows and tested on the organs along the column are shown.

The t-SNE embedding was obtained for different model-organ pairs. **Figure 6** shows t-SNE plots for KICH, LUSC, PRAD, and READ. The t-SNE plots of other model-organ pairs are shown in the supplementary section (**Figure S3**).

The embeddings show that the models are able to exhibit separability in feature space between cancer and normal patches for the subtype that it was trained on as well as for subtypes/organs with cross inference accuracy >90%. However, the t-SNE embeddings also indicate that few of the normal and cancer samples are at close proximities after projection to the 2D space. This could possibly be attributed to the models not being fully accurate, the 2D projection error, or the assumption that all patches in a cancer slide are cancerous.

3.3 Cross-Organ GradCAM Visualization

A further qualitative analysis was done comparing the GradCAM outputs of the model-organ pairs, with cross-organ inference accuracy >90% as well as cross-organ inference accuracy < 80%. **Figure 7** shows the quantitative results of the degree of overlap between GradCAM outputs using the IoU and Jaccard index.

TABLE 1 | Cross-organ inference indicating the quantitative results of best and worst inferences of individually trained models when tested on other unseen organs.

Model	F1 score				AUC			
	Best		Worst		Best		Worst	
BRCA	READ	0.9443	KICH	0.6600	READ	0.9837	KIRC	0.7815
COAD	READ	0.9799	KIRC	0.5287	READ	0.9981	KIRC	0.6246
KICH	COAD	0.9294	STAD	0.7519	KIRP	0.9783	STAD	0.8163
KIRC	KIRP	0.9423	PRAD	0.7678	KICH	0.9881	PRAD	0.8069
KIRP	KICH	0.9490	PRAD	0.6893	READ	0.9840	PRAD	0.6692
LIHC	READ	0.9442	KIRC	0.6157	READ	0.9893	KICH	0.7203
LUAD	LUSC	0.9381	KIRC	0.5675	LUSC	0.9831	KIRC	0.6256
LUSC	BRCA	0.9251	KIRC	0.5769	LUAD	0.9683	KIRC	0.5998
PRAD	BRCA	0.9422	KICH	0.5632	LIHC	0.9453	KICH	0.5481
READ	COAD	0.9680	KIRC	0.4987	LIHC	0.9507	KIRC	0.5246
STAD	READ	0.9410	KICH	0.6921	BRCA	0.9822	KICH	0.8062

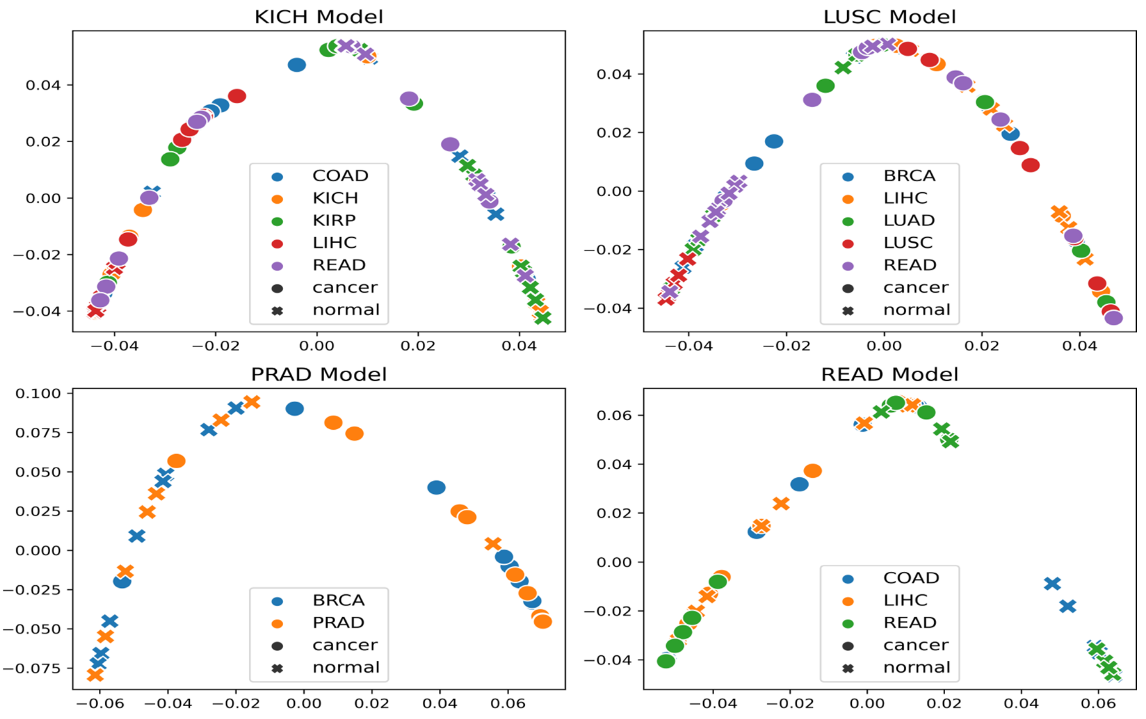


FIGURE 6 | t-SNE embeddings of the trained models (mentioned in the title of each figure) helping to visualize the separability of cancer and normal embeddings of organs unseen by the trained models.

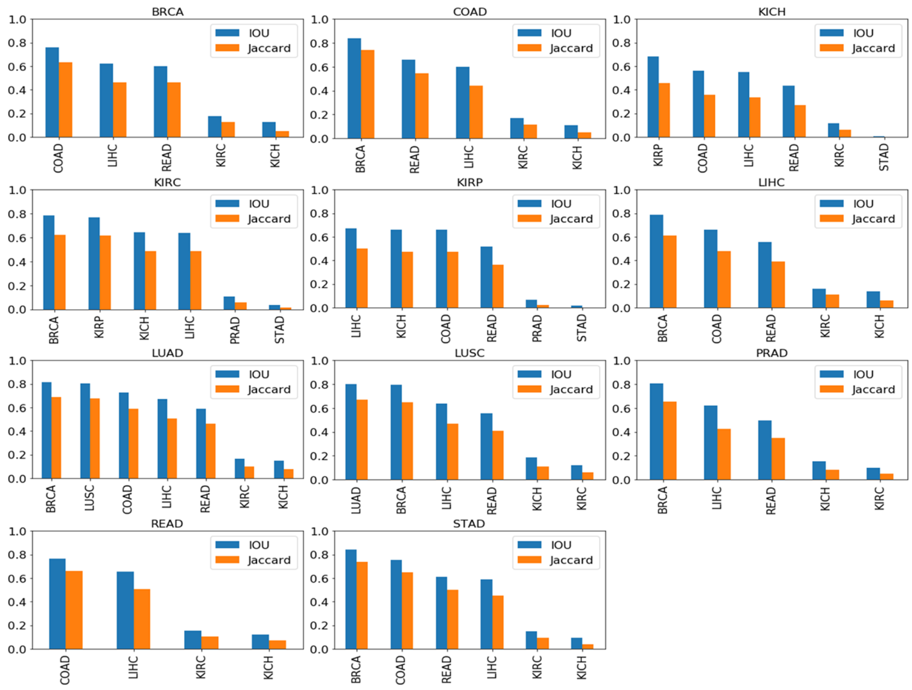


FIGURE 7 | Cross-organ GradCAM results showing the IoU and Jaccard index of high-attention regions. The model used for visualization is indicated on the title of each plot, and the subtypes used are indicated on the x-axis.

Figure 8 shows the visualization using the BRCA model on COAD, LIHC, and READ subtypes. The visualization for other cross-organ inferences are provided in the supplementary section (**Figures S4**, **S5**). The visualization outputs in green indicate regions with high attention, those in red indicate regions with moderate attention, and those in blue indicate no attention during the classification task. Ground-truth visualizations for the patch of an organ are obtained by using the model trained on the same organ. We compared the degree of overlap of the visualization outputs to comment on the shared tumor morphology. We observed a positive correlation between the observed cross-organ inference accuracy, i.e., the IoU and the Jaccard index are high for model-organ pairs with high cross-organ inference accuracy and low for model-organ pairs with low cross-organ inference accuracy. For example, the BRCA model has the highest cross-organ accuracy, highest IoU, and Jaccard index on COAD. The same trend is observed in the models of other organs.

3.4 Cross-Organ Similarities Seen in the Distribution of Nucleus Features

To further strengthen the hypothesis about cross-organ similarities, we observed the distribution of shape features of the nuclei present in the high-attention regions. We considered

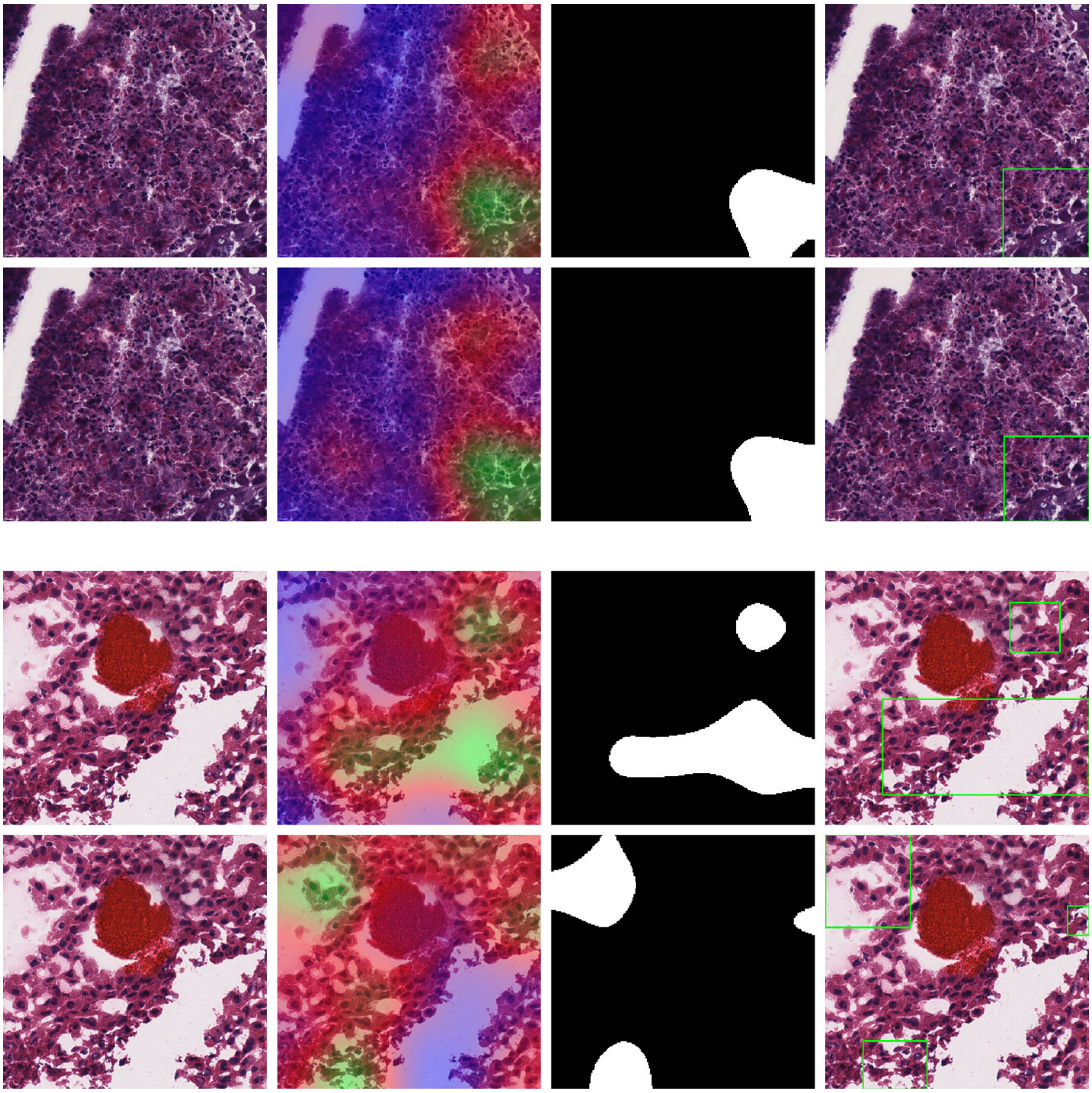


FIGURE 8 | Cross-organ GradCAM visualization of the BRCA model on COAD and KICH cancer patches. Columns show the input patch, GradCAM output, GradCAM thresholded, and GradCAM with bounding box, respectively. Top 2 rows show COAD input patches and visualization using the BRCA model (1st row) and COAD model (2nd row). Bottom 2 rows show KICH input patches and visualization using the BRCA model (3rd row) and KICH model (4th row).

two groups that showed good (BRCA and COAD) and another that showed poor (BRCA and KICH) performances in cross-organ inferences to characterize the nucleus morphological characteristics. We considered the high-probability patches [*P* (*cancer*) > 0.98] of COAD and KICH for the analysis. The distributions of some of the geometrical features of nuclei (main region extent and solidity) present in the regions focused by BRCA and COAD models on COAD patches are similar and correlated in contrast to the distributions seen with BRCA and KICH model on KICH patches (**Figure 9**).

We found that eight nucleus shape features and three inter-nucleus density features are significantly (p-value greater than 0.05) associated with the similarities observed between tumor morphologies (**Table 2**). Some of the significant nucleus shape features include total area (p-value = 0.0736), main extent (p-value = 0.1002), main region solidity (p-value = 0.0583), and some of the significant nucleus density features include neighbor count within a radius of 10, 20, and 30 pixels (p-value = 0.5974, 0.6044, 0.1945). We observe from the cross-organ performance table and the cross-organ GradCAM results that the BRCA model performs well on COAD patches and poorly on KICH patches and a similar behavior is seen in the distribution of nucleus geometrical features observed between the pairs of two groups (BRCA-COAD and BRCA-KICH).

4 CONCLUSION

In this work, we explored tumor features and morphology across multiple organs from a deep learning perspective. This has not been extensively studied compared to the pan-cancer studies based on molecular profiling. We report similarities based on very high performance obtained with models trained on one cancer and tested directly on another. This level of performance can be achieved only if the learnt features are general or common between cancers. Our observations span not only cancers originating from the same organ but also different organs, which are interesting. We observed that good cross-organ performance is also reflected in the separability of normal and cancerous patches in feature space when visualized using the t-SNE plot.

We also explored GradCAM techniques to establish that the models with high cross inference accuracy had a significant overlap in their attention regions. This suggests that the deep learning model is able to pick up shared morphological features that span across organs during classification. We further showed similarity at the nucleus level by analyzing the distribution of geometrical and graphical features of nuclei present in the overlapping and non-overlapping regions. Overall, our study presents the proof-of-principle experiment that deep learning

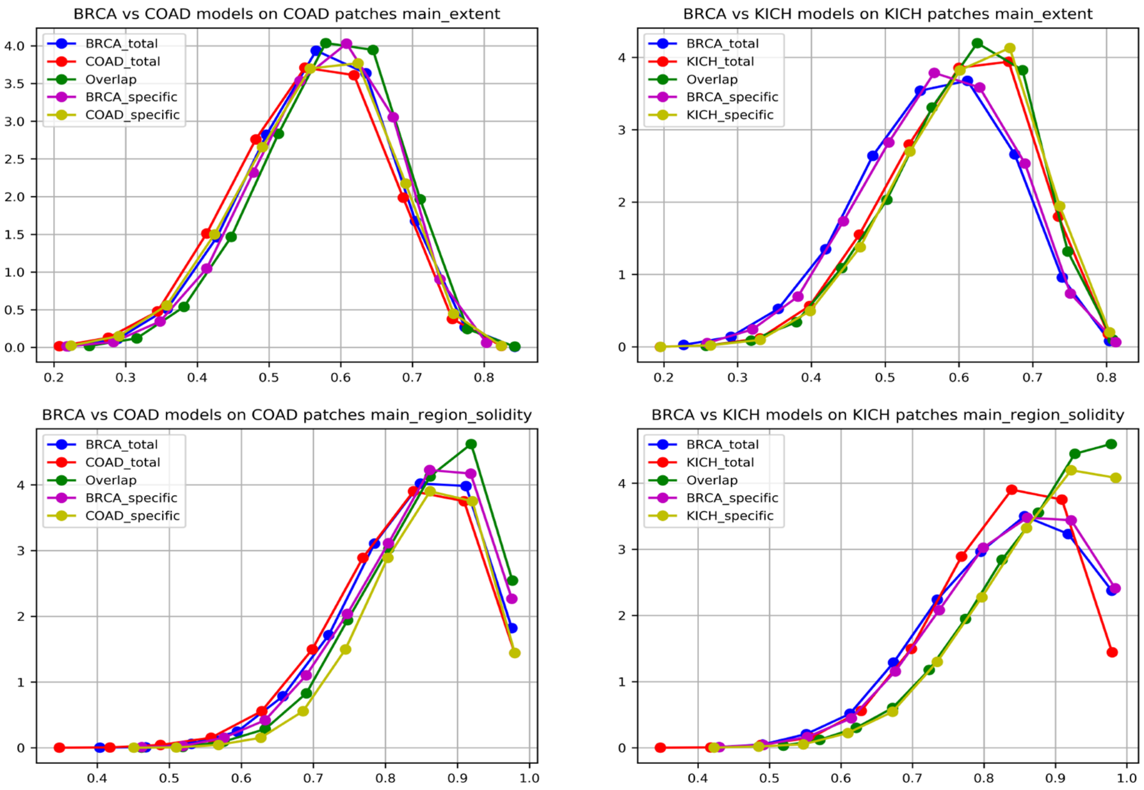


FIGURE 9 | Graph showing nuclei shape distribution of BRCA and COAD models inferred on COAD patches (left) and BRCA and KICH models inferred on KICH patches (right). The x-axis represents value of the feature, and the y-axis represents the PDF. In each subplot, "Total" is the overall high-attention region of the corresponding model, "overlap" is the common region of high attention for the two models, and "specific" is the "total" region excluding the "overlap".

TABLE 2 | Nucleus feature statistical analysis: the table showing the p-values obtained after performing the t-test on two pairs of groups (BRCA-COAD) and (BRCA-KICH).

Feature type	Features	p-value (BRCA and COAD)	p-value (BRCA and KICH)
Nucleus shape features	Total area	0.0736	1.0711E-132
	Total convex area	0.0823	4.6482E-147
	Total perimeter	0.0004	3.3526E-240
	Total filled area	0.0738	9.6456E-133
	Total major axis	0.0004	3.3976E-228
	Total minor axis	0.0002	1.8502E-194
	Total peri by area	0.0768	1.9164E-235
	Main region area	0.0171	2.0913E-246
	Main region convex area	0.0182	1.1829E-228
	Main region eccentricity	0.1387	4.3698E-18
	Main extent	0.1002	5.0677E-37
	Main region solidity	0.0583	2.4308E-34
	Main region perimeter	0.0007	0
	Main region angle	0.091	0.4033
	Main region peri by area	0.0142	1.4962E-203
	Main region major axis	0.0004	0
	Main region minor axis	0.0014	0
	Total diameter	0.0003	7.524E-225
Inter-nucleus density features	Neighbor count within a 10-pixel radius	0.5974	0.0009
	Neighbor count within a 20-pixel radius	0.6044	6.6978E-07
	Neighbor count within a 30-pixel radius	0.1945	1.0807E-19

A higher p-value indicates similarity, and a lower p-value indicates differences. The high p-values of the BRCA-COAD group go in agreement with the observed distribution plot.

and computational approaches can be adopted to explore the shared morphology across different cancers. There is a need for further characterization at the experimental level, which will be taken up as future work. We made publicly available the model checkpoints, the source code, and the best model architectures for most common cancers using TCGA data. All the resources can be accessed from the project page at https://bhasha.iiit.ac.in/tcga_cross_organ_project.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://portal.gdc.cancer.gov/.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Conceptualization, AM, PS, PV, and CJ. Methodology, AM and PS. Software, AM and PS. Validation, PV and CJ. Formal analysis, AM. Investigation, PS. Resources, AM and PS. Data curation, PS. Writing—original draft preparation, AM and PS. Writing—review and editing, PV and CJ. Visualization, AM, PS. Supervision, PV and CJ. Project administration, CJ. Funding acquisition, PV and CJ. All authors contributed to the article and approved the submitted version.

FUNDING

We thank IHub-Data, International Institute of Information and Technology, Hyderabad, for the financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.842759/full#supplementary-material

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J Clin* (2021) 71:209–49. doi: 10.3322/caac.21660
2. Chen F, Wendl MC, Wyczalkowski MA, Bailey MH, Li Y, Ding L. Moving Pan-Cancer Studies From Basic Research Toward the Clinic. *Nat Cancer* (2021) 29:879–90. doi: 10.1038/s43018-021-00250-4
3. Li D, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* (2018) 173:305–30.e10. doi: 10.1016/j.cell.2018. 03.033
4. Li J, Xu Q, Wu M, Huang T, Wang Y. Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection. *Front Bioeng Biotechnol* (2020) 8:766. doi: 10.3389/fbioe.2020.00766
5. Way GP, Sanchez-Vega F, La KC, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-Cancer Ras PathwayActivation in The Cancer Genome Atlas. *Cell Rep* (2018) 23:172–80.e3. doi: 10.1016/j.celrep.2018.03.046

6. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness FeaturesAssociated With Oncogenic Dedifferentiation. *Cell* (2018) 173:338–54.e15. doi: 10.1016/j.cell.2018.03.034
7. Arshi A, Olshen AB, Seshan VE, Shen R. Pan-Cancer Identification of Clinically Relevant Genomic Subtypes Using Outcome-Weighted Integrative Clustering. *Genome Med* (2020) 12:1–13. doi: 10.1186/s13073-020-00804-8
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet Classification With Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* (2012), 1097–105. doi: 10.1145/3065386
9. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. *Proc IEEE Conf Comput Vision Pattern Recog* (2016), 2424–33. doi: 10.1109/CVPR.2016.266
10. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for Segmenting and Classifying Epithelial and Stromal Regions in Histopathological Images. *Neurocomputing* (2016) 191:214–23. doi: 10.1016/j.neucom.2016.01.034
11. Sairam T, Vinod PK, Jawahar CV. Pan-Renal Cell Carcinoma Classification and Survival Prediction From Histopathology Images Using Deep Learning. *Sci Rep* (2019) 9.1:1–9. doi: 10.1038/s41598-019-46718-3
12. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and Mutation Prediction From non–Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat Med* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
13. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. *ArXiv* (2016), abs/1606.05718. doi: 10.48550/arXiv.1606.05718
14. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, et al. Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv preprint arXiv* (2017), 1703.02442. doi: 10.48550/arXiv.1703.02442
15. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. *Proc IEEE Conf Comput Vision Pattern Recog* (2016), 2818–26. doi: 10.1109/CVPR.2016.308
16. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *Proc IEEE Conf Comput Vision Pattern recognition* (2016), 770–8. doi: 10.1109/CVPR.2016.90
17. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper With Convolutions. *Proc IEEE Conf Comput Vision Pattern Recognition* (2015), 1–9. doi: 10.1109/CVPR.2015.7298594
18. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis. *Nat Cancer* (2020) 1.8:800–10. doi: 10.1038/s43018-020-0085-8
19. Cheerla A, Gevaert O. Deep Learning With Multimodal Representation for Pancancer Prognosis Prediction. *Bioinformatics* (2019) 35:i446–54:14. doi: 10.1093/bioinformatics/btz342
20. Noorbakhsh J, Farahmand S, Namburi S, Caruana D, Rimm D, Soltanieh-ha M, et al. Deep Learning-Based Cross-Classifications Reveal Conserved Spatial Behaviors Within Tumor Histological Images. *bioRxiv* (2020) 11(1):1–14. doi: 10.1101/715656
21. van der Maaten L, Hinton GE. Visualizing Data Using T-SNE. *J Mach Learn Res* (2008) 9:2579–605.
22. Tomczak K, Czerwin´ska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp Oncol* (2015) 19 (1A):A68–77. doi: 10.5114/wo.2014.47136
23. Ad Cooper L, Demicco EG, Saltz JH, Powell RT, Rao A, Lazar AJ. PanCancer Insights From The Cancer Genome Atlas: The Pathologist's Perspective. *J Pathol* (2018) 244(5):512–24. doi: 10.1002/path.5028
24. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. *Comput Struct Biotechnol J* (2018) 16:34–42. doi: 10.1016/j.csbj.2018.01.001
25. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-Generation Hyperparameter Optimization Framework. *Proc 25rd ACM SIGKDD Int Conf Knowl Discov Data Min* (2019), 2623–31. doi: 10.1145/3292500.3330701
26. Selvaraju RR, Das A , Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Visual Explanations From Deep Networks *via* Gradient-Based Localization. *Int J Comput Vision* (2019) 128:336–59. doi: 10.1007/s11263-019-01228-7
27. Zhan X, Cheng J, Huang Z, Han Z, Helm B, Liu X, et al. Correlation Analysis of Histopathology and Proteogenomics Data for Breast Cancer*. *Mol Cell Proteomics* (2019) 18:S37–51. doi: 10.1074/mcp.RA118.001232
28. Jun C, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. *Cancer Res* (2017) 77 21:e91–e100. doi: 10.1158/0008-5472.CAN-17-0313
29. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener BN. Histopathological Image Analysis: A Review. *IEEE Rev Biomed Eng* (2009) 2:147–71. doi: 10.1109/RBME.2009.2034865
30. Phoulady HA, Goldgof DB, Hall LO, Mouton PR. Nucleus Segmentation in Histology Images With Hierarchical Multilevel Thresholding. *inSPIE Med Imaging* (2016) 9791:280–5. doi: 10.1117/12.2216632
31. Doyle S, Agner SC, Madabhushi A, Feldman MD, Tomaszeweski JE. Automated Grading of Breast Cancer Histopathology Using Spectral Clustering With Textural and Architectural Image Features. *2008 5th IEEE Int Symposium Biomed Imaging: From Nano to Macro* (2008) 496–9. doi: 10.1109/ISBI.2008.4541041
32. Pandey N, Lanke V, Vinod PK. Network-Based Metabolic Characterization of Renal Cell Carcinoma. *Sci Rep* (2020) 10:1–13. doi: 10.1038/s41598-020-62853-8

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 19 April 2022
doi: 10.3389/fonc.2022.881910

Effects of Biofilm Nano-Composite Drugs OMVs-MSN-5-FU on Cervical Lymph Node Metastases From Oral Squamous Cell Carcinoma

Jian Huang, *Zhiyuan Wu* and *Junwu Xu**

Department of Oral and Maxillofacial Surgery, Fujian Provincial Hospital, Fuzhou, China

OPEN ACCESS

Edited by:
Fu Wang,
Xi'an Jiaotong University, China

Reviewed by:
Tingting Luo,
University of Electronic Science and Technology, China
Bin Peng,
Yunyang County People's Hospital, China
Mengjiao Song,
Dongying People's Hospital, China

****Correspondence:***
Junwu Xu
junwuxufujian@163.com

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *23 February 2022*
Accepted: *14 March 2022*
Published: *19 April 2022*

Citation:
Huang J, Wu Z and Xu J (2022) Effects of Biofilm Nano-Composite Drugs OMVs-MSN-5-FU on Cervical Lymph Node Metastases From Oral Squamous Cell Carcinoma. Front. Oncol. 12:881910. doi: 10.3389/fonc.2022.881910

This work was developed to the effects of biofilm composite nano-drug delivery system (OMVs-MSN-5-FU) on lymph node metastasis from oral squamous cell carcinoma. Mesoporous silica nanoparticles loaded with 5-FU (MSN-5-FU) were prepared first. Subsequently, the outer membrane vesicles (OMV) of Escherichia coli were collected to wrap MSN-5-FU, and then OMVs-MSN-5-FU was prepared. It was then immersed in artificial gastric juice and artificial intestinal juice to explore the drug release rate. Next, the effects of different concentrations of the nano-drug delivery systems on the proliferation activity of oral squamous carcinoma cell line KOSC-2 cl3-43 were analyzed. Tumor-bearing nude mice models were prepared by injecting human tongue squamous cell carcinoma cells Tca8113 into BALB/c-nu nude mice. They were injected with the OMVs-MSN-5-FU nano drug carrier system, and peri-carcinoma tissue and cervical lymph node tissue were harvested to observe morphological changes by Hematoxylin – eosin (HE) staining. The scanning electron microscope (SEM) results showed that all MSN, MSN-5-FU, OMV, and OMV-MSN-5-FU were spherical and uniformly distributed, with particle sizes of about 60nm, 80nm, 90nm, and 140nm, respectively. Among them, OMV had a directional core-shell structure. The cumulative drug release rates of artificial gastric juice in 48 hours were 61.2 ± 2.3% and 26.5 ± 3.1%, respectively. The 48 hours cumulative drug release rates of artificial intestinal juice were 70.5 ± 6.3% and 32.1 ± 3.8%, respectively. The cumulative release of MSN-5-FU was always higher than OMV-MSN-5-FU. The cumulative release of MSN-5-FU was always higher than OMV-MSN-5-FU. After injection of OMVS-MSN-5-FU, the number of cancer cells was significantly reduced and cervical lymph node metastasis was significantly controlled. HE staining results showed that OMVS-MSN-5-FU injection reduced the number of stained cells. Dense lymphocytes were clearly observed in the cortex of neck lymphocytes. The OMVs-MSN-5-FU drug delivery system can slow down the drug release rate, significantly inhibit the proliferation activity of oral squamous cancer cells, and control the metastasis of cancer cells to cervical lymph nodes.

Keywords: outer membrane vesicle, mesoporous silica nanoparticle, oral squamous carcinoma, lymph node metastasis, drug release rate

INTRODUCTION

Squamous cell carcinoma is a common malignant tumor of the head and neck, accounting for approximately 80% of all head and neck tumors, classified into oral cancer, laryngeal cancer, and nasopharyngeal cancer, etc. (1). Lymph node metastasis is an important cause of poor treatment effects and even death in cancer patients (2). Due to the special physiological anatomy of the oral maxillofacial region, most patients with squamous cancer are prone to neck lymph node metastasis, which seriously affects the prognosis (3). Data show that approximately 14% to 40% of patients with oral squamous cell carcinoma have cervical lymph node metastasis (4). Therefore, diagnosis of the cervical lymph node metastasis of oral squamous cell carcinoma is important (5). 5-Fluorouracil (5-FU) is a pyrimidine fluoride. It can inhibit the activity of thymine nucleotide synthase and is a commonly used anti-metabolism and anti-tumor drug. 5-FU has been widely used in the treatment of colorectal cancer, head and neck squamous cell carcinoma, and liver cancer (6–8). However, 5-FU can cause serious systemic side effects, such as mucositis and diarrhea, which is attributable to too little accumulation of 5-FU drugs in tumor tissue (9). Therefore, increasing the accumulation of 5-FU in the target area can enhance the efficacy of the drug and reduce the side effects.

Nowadays, nano-drug delivery systems play an important role in the pharmaceutical research and application, and nano-drug delivery systems prepared by encapsulating drugs in natural or synthetic polymer compounds can improve the therapeutic effects (10). By molding drugs into various nanostructures, they can be made more bioavailable and therapeutic. Polymer nanocarriers, solid lipid nanoparticles, nanostructured lipid carriers, nanoemulsions, nanodiamonds, vesicle-based drug carriers, metal-based nanoparticles, and nano-vaccines all have positive application effects as intelligent substitutes for drug delivery in the central nervous system (11). Nevertheless, the nano-preparation is only enriched in the liver or spleen, without a long-term circulation, and thus it can’t reach the targeted organs or tissue. To enhance the therapeutic efficiency of the nano-drug delivery system, to modify the cell membrane on the outer layer can significantly increase the drug loading. Additionally, the cell membrane has good biocompatibility, which also improves the stability of the nano-particles. Gram-negative bacteria outer membrane vesicles (OMV) are spherical biofilms, which are closed entities originating from endophytic cells (12). It can regulate the host’s immune response and participates a variety of biological and pathophysiological processes. Zhang et al. (13) improved the radiosensitivity of extranodal nasal NK/T cell lymphoma by combining radiotherapy with nano-drug delivery system, overcame the multi-drug resistance of chemotherapy drugs, and provided a new idea for the further development and optimization of treatment regimen for extranodal nasal NK/T cell lymphoma.

In this study, a biofilm composite nano drug delivery system (OMVs-MSN-5-FU) was prepared and immersed in artificial gastric juice and artificial intestinal juice to explore the drug release rate, and the effect of OMVS-MSN-5-FU on cervical lymph node metastasis was investigated. This study may provide a theoretical basis for the therapeutic effect of oral squamous cell carcinoma.

MATERIALS AND METHODS

Laboratory Reagents

5-FU (Beijing Bailingwei Technology Co., LTD.); Cetyl trimethyl ammonium bromide (Jining Sanshi Biotechnology Co., LTD.); Ethyl orthosilicate (Tianjin Kermel Chemical Reagent Co., LTD.); (Tianjin Guangfu Fine Chemical Research Institute); Dimethyl sulfoxide (Jiangsu Haolong Chemical Co., LTD.); and Phosphate buffer salt solution (PBS, Hyclone Corporation, USA) were utilized. All other reagents were domestic analytical pure reagents.

Preparation of Mesoporous Silica Carrier

The synthesis steps of mesoporous silica carrier were shown in **Figure 1**. 0.535g cetyltrimethylammonium bromide (CTAB) was dissolved in 240mL sterilized ultrapure water, and ultrasonic dispersion was performed for 15min. Then, 1.25mL of 2mol/L NaOH solution was added, followed by ultrasonic dispersion for 5min. The liquid was stirred continuously at 80°C for 30 minutes. Ethyl orthosilicate (TEOS) was added dropwise every 4s, totaling 5mL. After 2-hour reaction, the liquid was left at room temperature for 30min. The lower layer solution was taken for centrifugation at 10,000rpm for 3min to obtain the precipitate. The mesoporous silica nanoparticles (MSN) were then immersed in ethanol solution, followed by ultrasonic dispersion and centrifugation at 10,000 rpm for 3 minutes. The above steps were repeated three times. Next, the sample was transferred in a vacuum drying oven and dried overnight to obtain the purified MSN sample.

3-aminopropyltriethoxysilane (AMEO) was mixed with MSN at a ratio of 2:3, and an appropriate amount of toluene was added, followed by reflux under nitrogen protection for 12h at 110°C. Then, the solution was centrifuged at 10,000rpm for 3 mins, and the precipitate was washed with ethanol solution. The above steps were repeated three times. Finally, the surface amination treatment of MSN was carried out under vacuum drying conditions. The synthesis steps were shown in **Figure 2**.

Preparation of MSN-5-FU

MSN-5-FU nanoparticles were prepared by reverse phase microemulsification. 7.5 mL cyclohexane, 1.6 mL hexanol, and 1.8 mL Triton-100 (surfactant) were mixed evenly, and 5-Fu solution was added to form an inverting microemulsion. Magnetic stirring was performed at room temperature for 5 min. 150 μL ethyl orthosilicate and 100 μL 25% ammonia were added to the mixture, and the reaction was stirred continuously at room temperature for 24 h. 2 mL acetone was added for demulsification and centrifugation for 20 min. The products were collected and dispersed with ethanol and water respectively, followed by centrifugation to remove unreacted 5-Fu and solvent. The final nanoparticles were vacuum-dried.

Preparation of OMVs-MSN-5-FU

Luria-Bertani solid medium was used to cultivate Escherichia coli. After culturing at 37°C for 36 hours, a single colony was

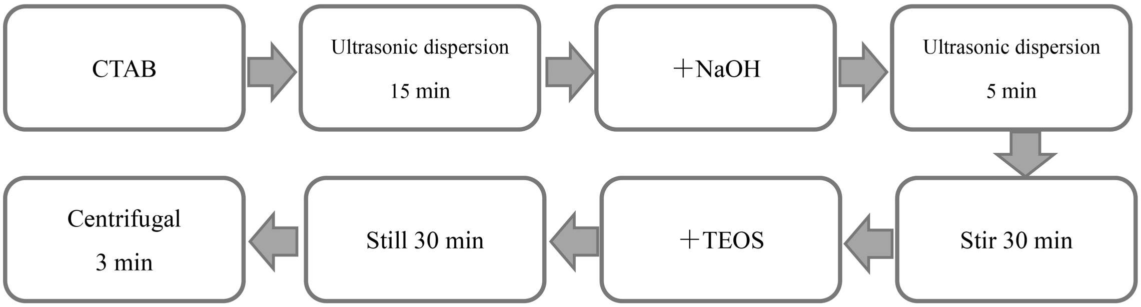


FIGURE 1 | Flow chart of preparation of mesoporous silica nanoparticles.

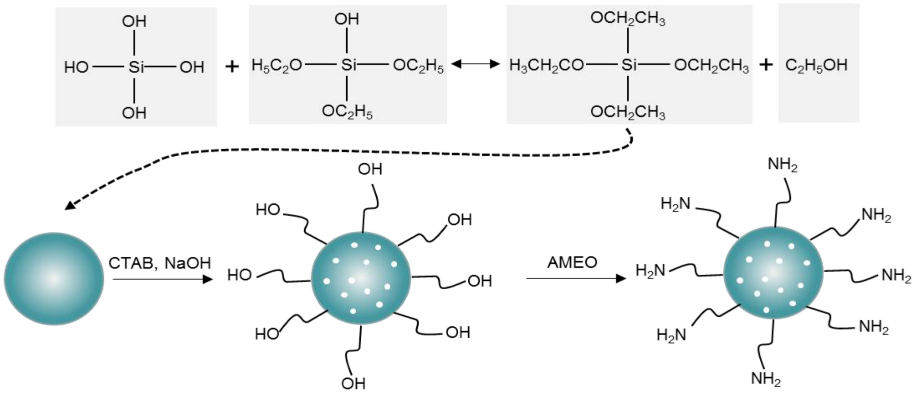


FIGURE 2 | MSN surface amine process.

inoculated in Luria-Bertani liquid medium and continued to be cultured for 24 hours. 2mL of bacterial solution was inoculated in blank Luria-Bertani liquid medium, set as the blank control group. The culture was terminated when the OD600 was 1. Then, centrifugation was performed at 12,000 rpm for 20 min at 4°C. The supernatant was then passed through the 0.45μm disposable filter. The filtrate was transferred to 20mL ultrafiltration centrifuge tube, followed by centrifugation at 8000rpm at 4°C for 15min. The centrifugal liquid was concentrated to 1.5mL, and PBS was used for resuspension to obtain OMVs.

High-pressure nitrogen was used to extrude the OMVs to uniform their particle sizes. The above steps were repeated 3 times. OMVs and MSN-5-FU were mixed at a ratio of 1:5, and then extruded 6 times. The extruded effluent was centrifuged at 8000 rpm at 4°C for 20 minutes, and the supernatant was discarded. The bottom precipitate was OMVs-MSN-5-FU.

Characterization Test

a. Transmission electron microscope (Talos F200X S/TEM, Beijing Opton Optical Technology Co., LTD.) and scanning electron microscope (LSM 900, Beijing Precise Instrument Co., LTD.) were used to observe the morphology of nanoparticles.

b. The Malvern Zeta particle size analyzer was used to measure the Zeta potential of nanoparticles, the measurement temperature was set to 25°C, and the equilibration time was 2 minutes.

In Vitro Release Test of OMVs-MSN-5-FU Drug Delivery System

The release amount of OMVs-MSN-5-FU drug delivery system was determined through the immersion test of artificial gastric juice and artificial intestinal juice. The first step was to prepare artificial gastric juice and artificial intestinal juice. Preparation of artificial gastric juice: 1mol/mL dilute hydrochloric acid with pH=1.5 was mixed with 0.001g/mL pepsin. The mixture was then passed through 0.2μm disposable sterile filter membrane for filter sterilization. Preparation of artificial intestinal juice: 6.8g KH2PO$_4$ was dissolved in 500mL sterile ultrapure water. With the pH of the solution set to 6.8, 0.01g/mL trypsin was then added. Next, the mixture was passed through a 0.2μm disposable sterile filter membrane for filter sterilization. Subsequently, 3mg of MSN-5-FU drug delivery system was added to 20mL of artificial gastric juice and 3mg OMVs-MSN-5-FU drug delivery system was added to another 20mL of artificial gastric juice, followed by shaking at 135rpm at 37°C. The artificial intestinal juice was treated in the same way. 2mL release solution was taken after 0.5h, 1.0h, 2.0h, 4.0h, 8.0h, 12.0h, 24.0h, 48.0h, and 72.0h of shaking, respectively, followed by centrifugation at 2000rpm for 5min at room temperature. Then, the supernatant was taken to determine the content of 5-FU.

In Vitro Cytotoxicity Test of OMVs-MSN-5-FU Drug Delivery System

KOSC-2 cl3-43 cells were inoculated in a 96-well plate. 0, 0.1, 0.25, 0.5, 1, 2.5, 5, and 10μmol/L 5-FU, MSN-5-FU, and OMVs-MSN-5-FU were inoculated, respectively. At 24h, 48h, and 72h of culturing, 10μL of MTT solution was added to each well. Then, 100μL of dimethylsulfoxide solution was added to each well, followed by shaking for 10 minutes. Finally, the absorbance was measured at a wavelength of 570nm. Cell inhibition rate was calculated. The calculation method of cell inhibition rate is shown in equation (1), where *RI* was the cell inhibition rate, *AS* was the absorbance value of the experimental well, *AK* was the absorbance value of the blank well, and *AY* was the absorbance value of the negative well.

$$RI = 1 - \frac{AS - AK}{AY - AK} \times 100\,\% \tag{1}$$

Preparation of Tumor-Bearing Animal Models

The clean-grade BALB/c-nu nude mice were used, aged about 5 weeks old, weighing 16-21g, regardless of the gender. They were provided by the XXX animal laboratory. Animal License No.:SCXK (Hebei)2019-0027. In XXX animal laboratory, the day and night cycle was 12 hours at 22~26°C and 45~50% humidity. After three days of adaptive breeding, they were used in formal experiments.

Human tongue squamous cell carcinoma cells Tca8113 were cultured first, and PBS was used to prepare a single cell suspension at a concentration of $1{\times}10^7$ during subculture. 10% chloral hydrate solution was intraperitoneally injected to anesthetize BALB/c-nu nude mice, and 0.2mL of Tca8113 cell

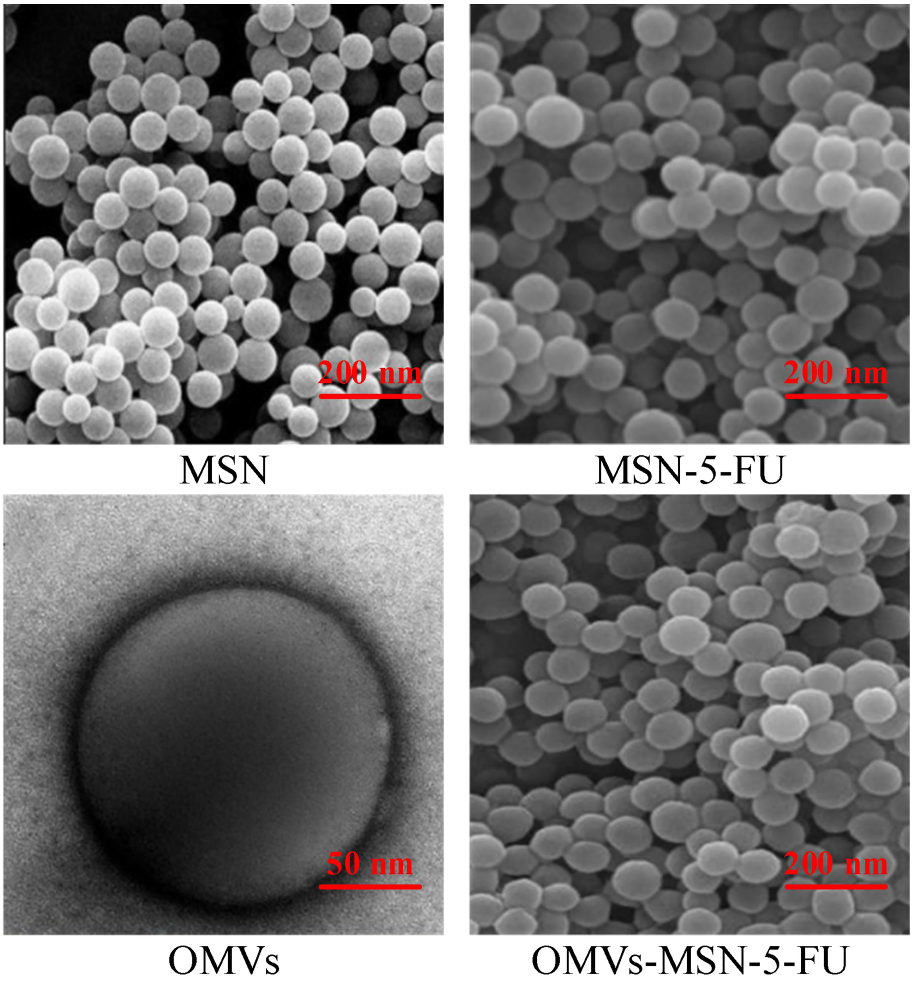


FIGURE 3 | Scanning electron microscope images.

suspension was injected **into the buccal mucosa**. The growth status of the tumor was observed regularly.

When the diameter of the tumor was approximately 0.5 cm, BALB/c-nu nude mice were anesthetized by intraperitoneal injection of 10% chloral hydrate solution, and 100μg/mL OMVs-MSN-5-FU suspension was injected into the buccal mucosa around the tumor in the oral cavity.

Detection of Naa10 and Lymphocyte Subtypes in Peripheral Blood

Blood was drawn from the tail vein of the tumor-bearing animal model, and enzyme-linked immunosorbent assay was used to detect the level of Naa10 in the peripheral blood. The serum sample was transferred in an enzyme-labeled plate, and incubated at 37°C for 2 hours. After the supernatant was discarded, 100μL of test reagent A was added to each well. After 1 hour, the supernatant was discarded and the cells were washed with sterile ultrapure water 3 times. Then, 100μL of test reagent B was added to each well. After 1 hour, 90μL of enzyme-labeled reagent was added to each well, and 50μL of stop solution was added after incubation for 20min. Finally, the absorbance was measured using a microplate reader at 450nm.

Flow cytometry was used to detect the changes of T lymphocyte subsets ($CD3^{+}$, CD^{4+} and CD^{8+}), B lymphocytes (CD^{19+}) and NK cells (CD^{56+}) in peripheral blood.

Observation of Tissue Sections

Three days after the injection of OMVs-MSN-5-FU suspension, the animal model was sacrificed by cervical dislocation, and the tissue around the carcinoma and ipsilateral cervical lymph nodes were harvested. The tissue was washed with PBS, and then put in the embedding box. Then, 70%, 90%, 95%, 95%, 100%, 100%, and 100% ethanol solution was used in turn to dehydrate the tissue, followed by immersion in xylene solution for 20 minutes. This step was repeated 3 times. The treated tissue was embedded in paraffin solution to made paraffin tissue sections with a thickness of 3μm. After being dried, the slide was immersed in xylene solution for 10 minutes, twice. Next, 100%, 100%, 95%, 95%, 80% ethanol solution and distilled water were used in turn

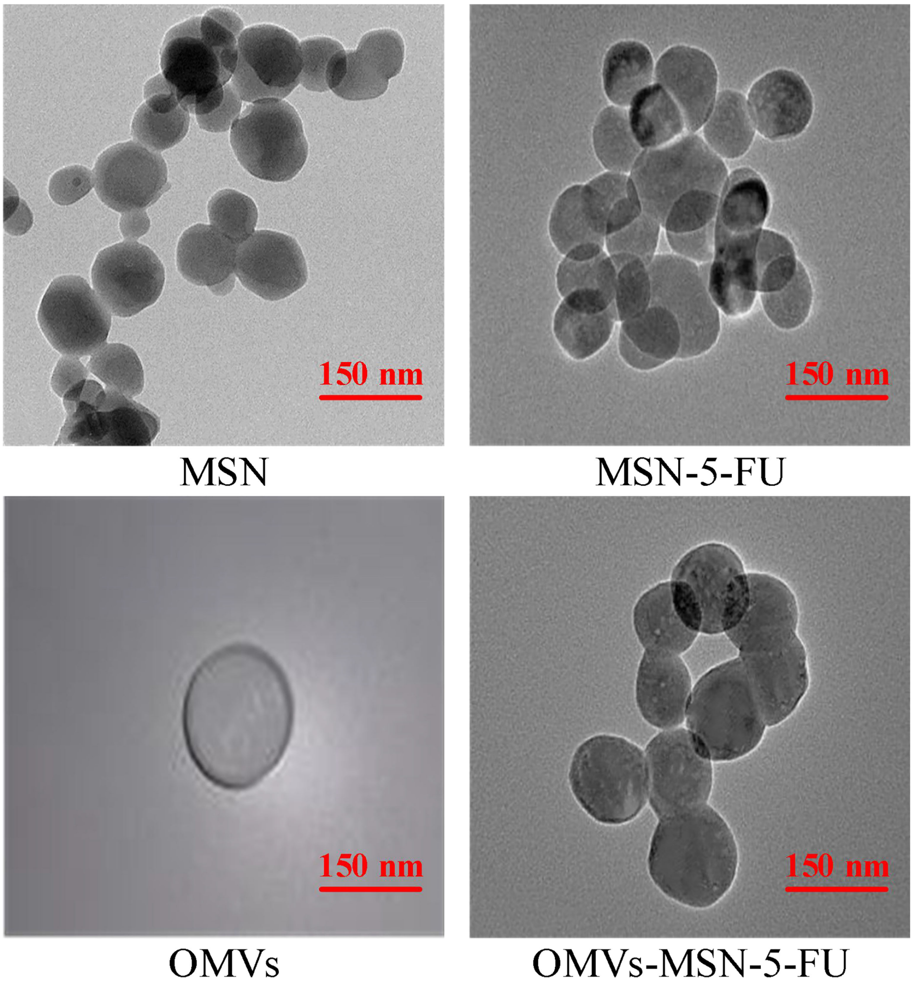


FIGURE 4 | Transmission electron microscope images.

for tissue rehydration. Subsequently, hematoxylin and eosin staining solutions were used to stain the tissue, followed by immersion in xylene. Finally, the slide was sealed using neutral gum, and visualized under an optical microscope.

Statistical Analysis

SPSS19.0 was used to process the data. The data were all expressed by the mean ± standard deviation (*x*(-) ± s).One-way analysis of variance was used for statistical analysis of the differences between multiple groups, and *P*<0.05 was the threshold for significance.

RESULTS

Transmission Electron Microscopy and Scanning Electron Microscopy Images of the Nano-Drug Delivery Systems

The morphology of MSN, MSN-5-FU, OMVs, and OMVs-MSN-5-FU was visualized under scanning electron microscope and transmission electron microscope. It was found that, all MSN, MSN-5-FU, OMVs, and OMVs-MSN-5-FU were spherical in shape, with uniform distribution. The particle diameters were approximately 60nm, 80nm, 90nm, and 140nm, respectively. Among them, OMVs had an oriented core-shell structure. The scanning electron microscope and transmission electron microscope results were shown in **Figures 3**, **4**.

Zeta Potential Measurement Results

The average Zeta potentials of MSN, MSN-5-FU, OMVs, and OMVs-MSN-5-FU were -20.6 ± 2.3mV, -28.7 ± 2.2mV, -18.2 ± 3.1mV, and -17.4 ± 1.7mV, respectively. The average zeta potential of OMVs was close to that of OMVs-MSN-5-FU. The results of Zeta potential detection were shown in **Figure 5**.

In Vitro Release of the Nano-Drug Delivery Systems

The *in vitro* release rate of MSN-5-FU and OMVs-MSN-5-FU drug delivery systems was analyzed. It was noted that,

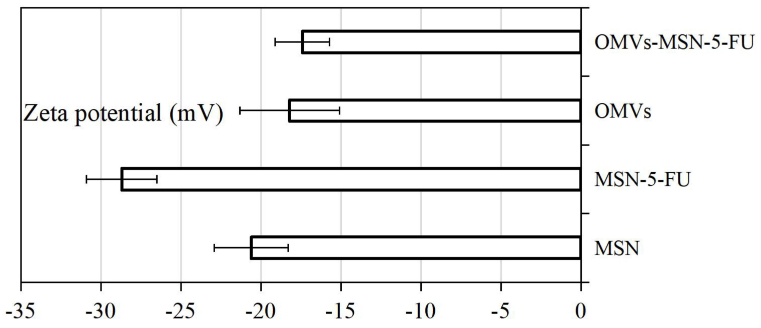


FIGURE 5 | Average Zeta potential plot.

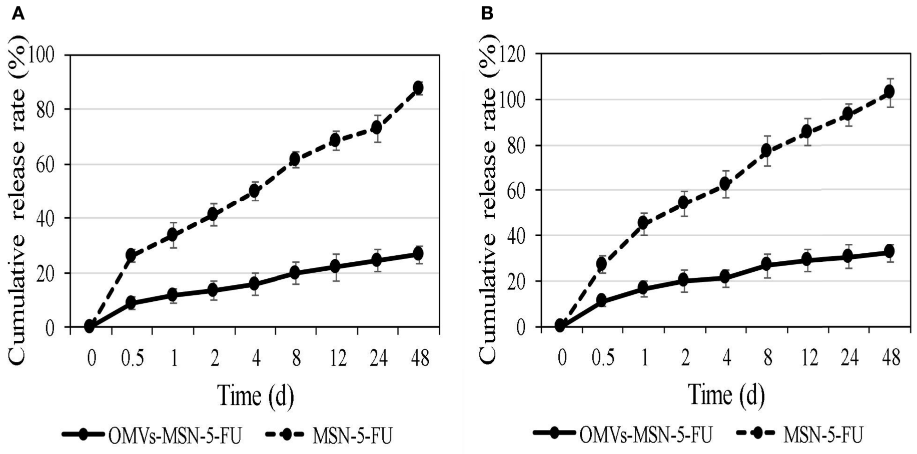


FIGURE 6 | *In vitro* cumulative release rate curve of the nano-drug delivery systems. **(A)** was the cumulative release curve in artificial gastric juice; **(B)** was the cumulative release curve in artificial intestinal juice.

the cumulative release rate of MSN-5-FU and OMVs-MSN-5-FU drug delivery system gradually increased in artificial gastric juice and artificial intestinal juice. In addition, the 48-hour cumulative drug release rate in artificial gastric juice was 61.2 ± 2.3% and 26.5 ± 3.1%, respectively, and the 48-hour cumulative drug release rate in artificial intestinal juice was 70.5 ± 6.3% and 32.1 ± 3.8%, respectively. The cumulative release of MSN-5-FU was always higher than that of OMVs-MSN-5-FU. The cumulative release rate results were shown in **Figure 6**.

Cytotoxicity of the Nano-Drug Delivery Systems

The results of cell viability detected by MTT showed that, with the increase of the concentration of 5-FU, MSN-5-FU and OMVs-MSN-5-FU drug delivery systems, the proliferation activity of KOSC-2 cl3-43 cells showed a gradually decreasing trend. At the same time, under the same dosage, the inhibitory rate of OMVs-MSN-5-FU drug delivery system on the proliferation activity of KOSC-2 cl3-43 cells was higher than that of 5-FU and MSN-5-FU. The specific results were shown in **Figure 7**.

Identification of Tumor-Bearing Animal Models

HE staining was used to analyze the pathological changes of tumor-bearing animal models. It was found that, the cancer tissue showed enlarged nuclei, darkened staining, and irregularly shaped cancer cells. The adjacent tissue mainly consisted of striated muscle, and there were oval nuclei on the edge. Observation of cervical lymph node slices revealed that, the lymphocytes in the cortex were very dense and were divided by cancer cells. The shape of the nucleus of lymphocytes was approximately round and there was no cytoplasm; while the

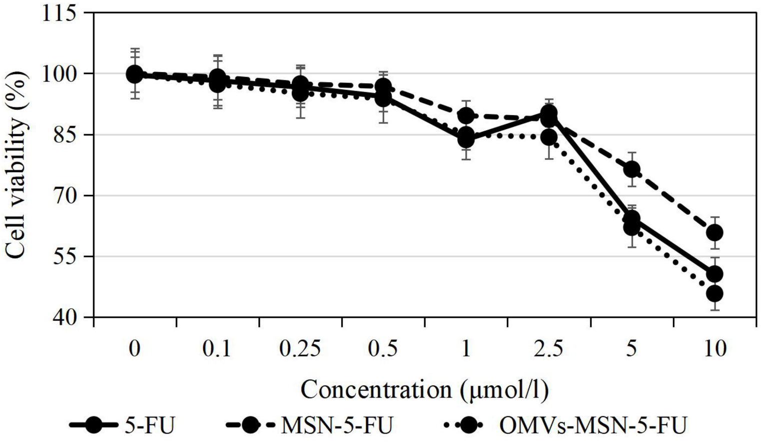


FIGURE 7 | The inhibitory effects of free drug and the nano-drug delivery systems on cell proliferation.

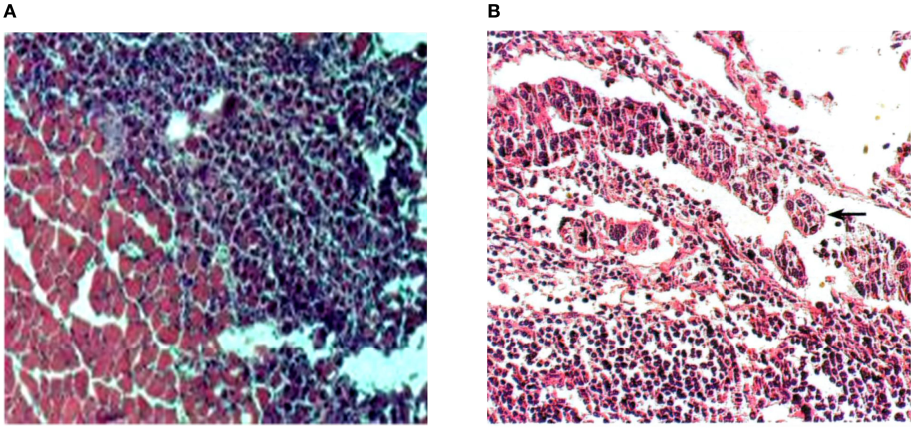


FIGURE 8 | HE staining results of tissue sections. **(A)** was HE staining of adjacent tissue sections (×400); **(B)** was HE staining of tissue sections of cervical lymph node metastases (×400).

nucleus of cancer cells was enlarged, and the cytoplasm was connected to each other into a sheet. It suggested that, the cancer cells have gradually metastasized to the lymph nodes in the neck, accompanied by cell proliferation. The specific staining results were shown in **Figure 8**.

Detection of Naa10 and Lymphocyte Subgroups in Peripheral Blood

The test results showed that, the model group showed increased levels of Naa10, $CD8^+$, and $CD56^+$ in peripheral blood, while decreased levels of $CD3^+$, $CD4^+$, $CD4^+/CD8^+$ and $CD19^+$ ($P<0.05$) versus the normal control group; and that compared with the model group, the OMVs-MSN-5-FU group showed decreased levels of Naa10, $CD8^+$, and $CD56^+$ in the peripheral blood, while increased levels of $CD3^+$, $CD4^+$, $CD^+/CD8^+$ and $CD19^+$ ($P<0.05$). However, there was no significant difference between the control group and the OMVs-MSN-5-FU group in the levels of Naa10 and lymphocyte subgroups ($P>0.05$). The detection results of Naa10 and lymphocyte subgroups in peripheral blood were shown in **Figure 9**.

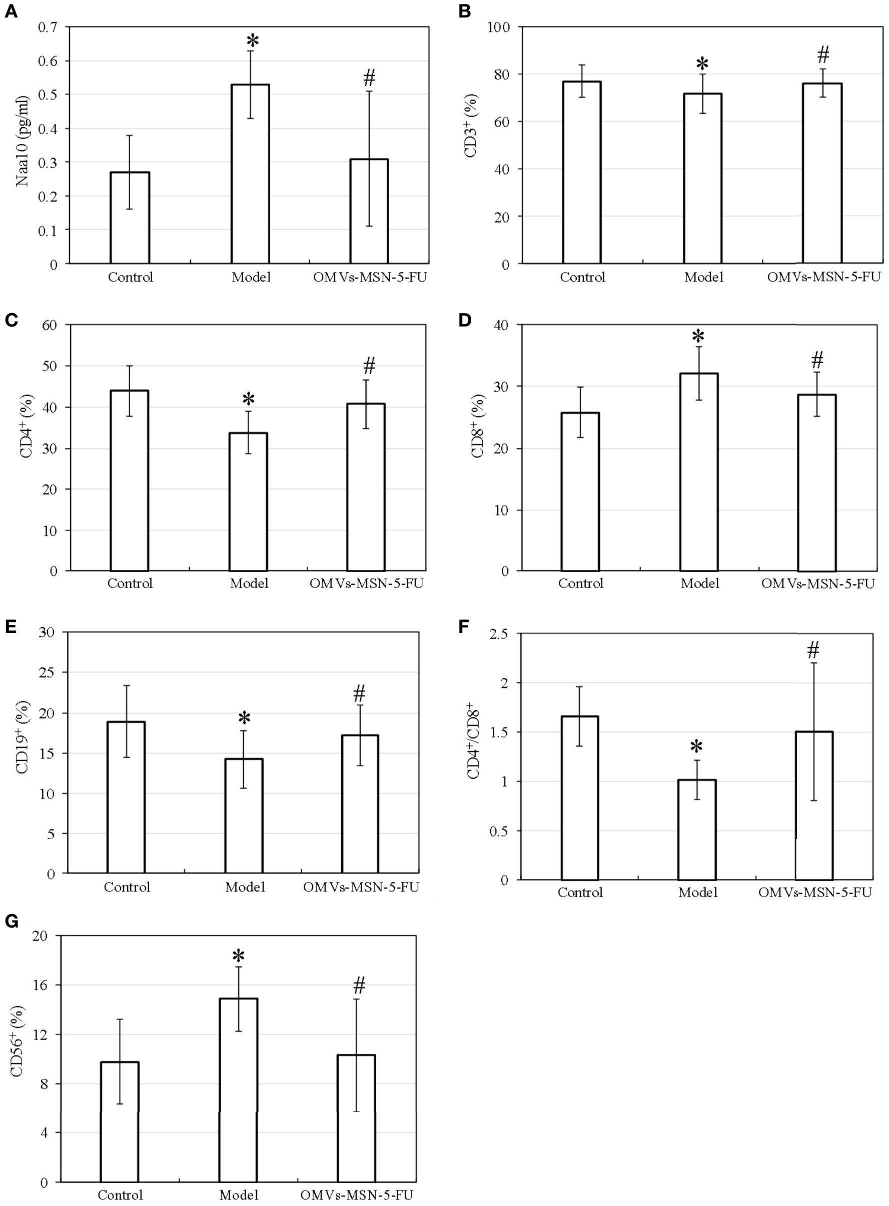


FIGURE 9 | Differences in the levels of Naa10 and lymphocyte subgroups in peripheral blood. **(A)** was Naa10; **(B)** was $CD3^+$ level; **(C)** was $CD4^+$; **(D)** was $CD8^+$; **(E)** was $CD19^+$; **(F)** was $CD4^+/CD8^+$ ratio; **(G)** was $CD56^+$; compared to the control group, $^*P<0.05$; compared to the model group, $^\#P<0.05$.

HE Staining of Peri-Carcinoma and Cervical Lymph Node Tissue

The HE staining results showed that, the aligned nuclei were noted in the striated muscle of the peri-carcinoma tissue, and the injection of OMVs-MSN-5-FU reduced the number of stained cells. In the cortex of the neck lymphocytes, dense lymphocytes were clearly observed. The specific staining results were shown in **Figure 10**.

DISCUSSION

Patients with oral squamous cell carcinoma have a low survival rate after treatment, and metastasis may lead to poor prognosis or even death (14). There are abundant lymphatic tissues in maxillofacial region, and facial movement promotes the metastasis of oral squamous cell carcinoma to cervical lymph nodes (15). Oral squamous carcinoma has a high probability of metastasis to cervical lymph nodes. If metastasis occurs, patient survival is greatly reduced. As a thymidylate synthase inhibitor, 5-FU is often injected intravenously for the treatment of cancer patients (16). 5-FU drugs show good therapeutic effects on digestive system tumors and breast tumors. It can also be used to treat ovarian cancer, bladder cancer, and head and neck cancer (17, 18). However, the drug has a great toxic effect on bone marrow and digestive tract, so it is important to improve the therapeutic effect of the drug and reduce the toxic side effects. William (2021) (19) found that 5-FU was beneficial to improve the survival rate of colorectal cancer patients, but severe systemic toxicity (including neutropenia) occurred in 30% of patients, and 0.5-1% of patients were fatal.

In the study, Escherichia coli biofilm was used to prepare a composite nano-drug carrier system containing 5-FU drugs. OMV helps bacteria adapt to the ecological niche, and enables them to compete with others, playing a protective role (20). In this study, Escherichia coli OMV was used to wrap the MSN-5-FU drug delivery system, which was then immersed in artificial gastric juice and artificial intestinal juice to analyze the drug release rate. The results showed that, compared with MSN-5-FU, the cumulative drug release rate of OMVs-MSN-5-FU was significantly reduced. This greatly prolonged the targeted action time of the drug and improved the therapeutic effects. Finally, after co-cultured with oral squamous cell carcinoma cell lines, it was found to significantly inhibit the proliferation activity of the cells. It suggested that OMVs-MSN-5-FU had significant inhibitory effects on the proliferation activity of oral squamous cell cancer cells, and then enhanced the therapeutic effects.

Tca8113 is a type of human tongue squamous carcinoma cell line with very stable genetic traits, and has been widely used in animal experiments (21). In view of oral squamous cell carcinoma prone to neck lymphocyte metastasis, Tca8113 cells were used to prepare a tumor-bearing mouse model, and the neck lymphocyte metastasis was analyzed by making sections. The results showed that, there was obvious edema in the peri-carcinoma tissue, and the increased internal pressure caused the anchor wire connecting the endothelial cells and surrounding tissues to be pulled, which increased the pores between the lymphatic endothelial cells. In order to evaluate the therapeutic effects of OMVs-MSN-5-FU on

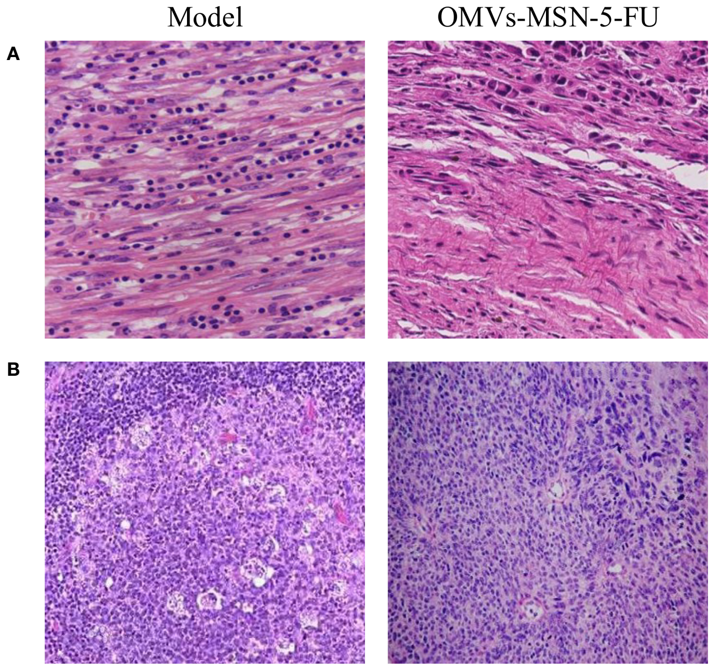


FIGURE 10 | HE staining of tissue sections after treatment. **(A)** was HE staining of tissue sections adjacent to the carcinoma (×400); **(B)** was HE staining of the tissue sections of cervical lymph node metastases (×200).

the animal model of oral squamous cell carcinoma, the levels of Naa10 and lymphocyte subgroups in the peripheral blood were factored into. Naa10 is the only subunit that can be catalyzed in the N-acetyltransferase A complex, and it plays an important role in the cell biology process (22). Studies have shown that, Naa10 participates in the autophagy, apoptosis, and proliferation of tumor cells (23). The results of this study showed that, the level of Naa10 in the model group was significantly increased, and the injection of OMVs-MSN-5-FU could reduce the level of Naa10. Cancer cells can cause the deterioration of the disease through processes such as immune escape (24). The subgroups of peripheral blood lymphocytes were then analyzed. The results showed that, the levels of $CD3^+$, $CD4^+$, $CD4^+/CD8^+$ and NK cells in the peripheral blood of the model group were significantly decreased, while the levels of CD8+ and B lymphocytes were significantly increased. This indicated that the lymphocyte subgroups of model group changed significantly. After injection of OMVs-MSN-5-FU, the levels of subgroups of peripheral blood lymphocytes in the animal model almost returned to normal. This indicated that OMVs-MSN-5-FU can regulate the balance between effector T cells and helper T cells, to maintain the stability of the body's environment, thereby improving the oral squamous cell carcinoma.

CONCLUSION

To investigate the effect of OMVs-MSN-5-FU compound drugs on lymph node metastasis of oral squamous cell carcinoma, the effects of different concentrations of nano drug delivery system on the proliferation activity of KOSC-2 cl3-43 oral squamous cell carcinoma cell line were analyzed. The results showed that OMVS-MSN-5-FU compound drugs could inhibit the proliferation activity of oral squamous cell carcinoma cells, regulate the peripheral blood subsets, and inhibit the metastasis of cancer cells to cervical lymph nodes. However, some limitations should be noted. This work only analyzed the effect of the drug on the animal model of oral squamous cell carcinoma, but did not explore its internal molecular mechanism. The molecular mechanism will be further explored in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

(I) Conception and design: JH; (II) Administrative support: JH and ZW; (III) Provision of study materials or patients: JH; (IV) Collection and assembly of data:JH and JX; (V) Data analysis and interpretation: JH, ZW,JX, (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

REFERENCES

1. Yokota T, Homma A, Kiyota N, Tahara M, Hanai N, Asakage T, et al. Immunotherapy for Squamous Cell Carcinoma of the Head and Neck. *Jpn J Clin Oncol* (2020) 50(10):1089–96. doi: 10.1093/jjco/hyaa139
2. Solomon B, Young RJ, Rischin D. Head and Neck Squamous Cell Carcinoma: Genomics and Emerging Biomarkers for Immunomodulatory Cancer Treatments. *Semin Cancer Biol* (2018) 52(Pt 2):228–40. doi: 10.1016/j.semcancer.2018.01.008
3. Forghani R, Chatterjee A, Reinhold C, Pérez-Lara A, Romero-Sanchez G, Ueno Y, et al. Head and Neck Squamous Cell Carcinoma: Prediction of Cervical Lymph Node Metastasis by Dual-Energy CT Texture Analysis With Machine Learning. *Eur Radiol* (2019) 29(11):6172–81. doi: 10.1007/s00330-019-06159-y
4. Gallegos-Hernández JF, Abrego-Vázquez JA, Olvera-Casas A, Minauro-Muñoz GG, Ortiz-Maldonado AL. Cervical Metastasis of Squamous Cell Carcinoma of the Head and Neck Therapeutic Options. *Cir Cir* (2019) 87 (2):141–5. doi: 10.24875/CIRU.18000412
5. Kann BH, Hicks DF, Payabvash S, Mahajan A, Du J, Gupta V, et al. Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *J Clin Oncol* (2020) 38(12):1304–11. doi: 10.1200/JCO.19.02031
6. André T, Boni C, Mounedji-Boudiaf L, Navarro M, Tabernero J, Hickish T, et al. Oxaliplatin, Fluorouracil, and Leucovorin as Adjuvant Treatment for Colon Cancer. *N Engl J Med* (2004) 350(23):2343–51. doi: 10.1056/NEJMoa032709
7. Vazquez T, Florez-White M. A Patient With Squamous Cell Carcinoma *in-Situ* Successfully Treated With Intralesional 5-Fluorouracil and Topical Trichloroacetic Acid. *J Dermatolog Treat* (2020) 31(2):180–2. doi: 10.1080/09546634.2019.1589642
8. Ding H, Wang Y, Zhang H. CCND1 Silencing Suppresses Liver Cancer Stem Cell Differentiation and Overcomes 5-Fluorouracil Resistance in Hepatocellular Carcinoma. *J Pharmacol Sci* (2020) 143(3):219–25. doi: 10.1016/j.jphs.2020.04.006
9. Bui AD, Grob SR, Tao JP. 5-Fluorouracil Management of Oculofacial Scars: A Systematic Literature Review. *Ophthalmic Plast Reconstr Surg* (2020) 36 (3):222–30. doi: 10.1097/IOP.0000000000001532
10. Zhou F, Teng F, Deng P, Meng N, Song Z, Feng R. Recent Progress of Nano-Drug Delivery System for Liver Cancer Treatment. *Anticancer Agents Med Chem* (2018) 17(14):1884–97. doi: 10.2174/1871520617666170713151149
11. Aggarwal N, Sachin, Nabi B, Aggarwal S, Baboota S, Ali J. Nano-Based Drug Delivery System: A Smart Alternative Towards Eradication of Viral Sanctuaries in Management of NeuroAIDS. *Drug Deliv Transl Res* (2022) 12(1):27–48. doi: 10.1007/s13346-021-00907-8
12. Jan AT. Outer Membrane Vesicles (OMVs) of Gram-Negative Bacteria: A Perspective Update. *Front Microbiol* (2017) 8:1053. doi: 10.3389/fmicb.2017.01053
13. Zhang X, Wu J, Lin D. Construction of Intelligent Nano-Drug Delivery System for Targeting Extranodal Nasal Natural Killer/Thymus Dependent Lymphocyte. *J BioMed Nanotechnol* (2021) 17(3):487–500. doi: 10.1166/jbn.2021.3048
14. Bilgic O, Duda L, Sánchez MD, Lewis JR. Feline Oral Squamous Cell Carcinoma: Clinical Manifestations and Literature Review. *J Vet Dent* (2015) 32(1):30–40. doi: 10.1177/089875641503200104
15. Liu Q, Li SK, Li H, Luo WM, Xu LQ. Expression Of Chondromodulin-1 And Vascular Endothelial Growth Factor-A In Esophageal Squamous Cell Carcinoma And Influence On Tumor Angiogenesis Authors. *Acta Med Mediterr* (2020) 36(5):3101–6. doi: 10.19193/0393-6384_2020_5_478
16. Hashimoto Y, Yoshida Y, Yamada T, Aisu N, Yoshimatsu G, Yoshimura F, et al. Current Status of Therapeutic Drug Monitoring of 5-Fluorouracil Prodrugs. *Anticancer Res* (2020) 40(8):4655–61. doi: 10.21873/anticanres.14464

17. Guler Y, Ovey IS. Synergic and Comparative Effect of 5-Fluorouracil and Leucoverin on Breast and Colon Cancer Cells Through TRPM2 Channels. *Bratisl Lek Listy* (2018) 119(11):692–700. doi: 10.4149/BLL_2018_124

18. Coen JJ, Zhang P, Saylor PJ, Lee CT, Wu CL, Parker W, et al. Bladder Preservation With Twice-A-Day Radiation Plus Fluorouracil/Cisplatin or Once Daily Radiation Plus Gemcitabine for Muscle-Invasive Bladder Cancer: NRG/RTOG 0712-A Randomized Phase II Trial. *J Clin Oncol* (2019) 37(1):44–51. doi: 10.1200/JCO.18.00537

19. Gmeiner WH. A Narrative Review of Genetic Factors Affecting Fluoropyrimidine Toxicity. *Precis Cancer Med* (2021) 4:38. doi: 10.21037/pcm-21-17

20. Mancini F, Rossi O, Necchi F, Micoli F. OMV Vaccines and the Role of TLR Agonists in Immune Response. *Int J Mol Sci* (2020) 21(12):4416. doi: 10.3390/ijms21124416

21. Wang TT, Chen JL, Chen G, Li ZQ. Low-Dose Bpa Promotes Proliferation, Invasion, and Migration on Tca8113 Cells Through Regulating Rip1. *Acta Med Mediterr* (2021) 37(4):2591–6. doi: 10.19193/0393-6384_2021_4_402

22. Kim SM, Ha E, Kim J, Cho C, Shin SJ, Seo JH. NAA10 as a New Prognostic Marker for Cancer Progression. *Int J Mol Sci* (2020) 21(21):8010. doi: 10.3390/ijms21218010

23. Kuhns KJ, Zhang G, Wang Z, Liu W. ARD1/NAA10 Acetylation in Prostate Cancer. *Exp Mol Med* (2018) 50(7):1–8. doi: 10.1038/s12276-018-0107-0

24. Chen DS, Mellman I. Elements of Cancer Immunity and the Cancer-Immune Set Point. *Nature* (2017) 541(7637):321–30. doi: 10.1038/nature21349

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fonc.2022.870457

Spherical Convolutional Neural Networks for Survival Rate Prediction in Cancer Patients

Fabian Sinzinger[1]*, *Mehdi Astaraki*[1,2], *Örjan Smedby*[1] *and Rodrigo Moreno*[1]

[1] Division of Biomedical Imaging, Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden, [2] Karolinska Institutet, Department of Oncology-Pathology, Karolinska Universitetssjukhuset, Stockholm, Sweden

Objective: Survival Rate Prediction (SRP) is a valuable tool to assist in the clinical diagnosis and treatment planning of lung cancer patients. In recent years, deep learning (DL) based methods have shown great potential in medical image processing in general and SRP in particular. This study proposes a fully-automated method for SRP from computed tomography (CT) images, which combines an automatic segmentation of the tumor and a DL-based method for extracting rotational-invariant features.

Methods: In the first stage, the tumor is segmented from the CT image of the lungs. Here, we use a deep-learning-based method that entails a variational autoencoder to provide more information to a U-Net segmentation model. Next, the 3D volumetric image of the tumor is projected onto 2D spherical maps. These spherical maps serve as inputs for a spherical convolutional neural network that approximates the log risk for a generalized Cox proportional hazard model.

Results: The proposed method is compared with 17 baseline methods that combine different feature sets and prediction models using three publicly-available datasets: Lung1 (n=422), Lung3 (n=89), and H&N1 (n=136). We observed comparable C-index scores compared to the best-performing baseline methods in a 5-fold cross-validation on Lung1 (0.59 ± 0.03 vs. 0.62 ± 0.04). In comparison, it slightly outperforms all methods in inter-data set evaluation (0.64 vs. 0.63). The best-performing method from the first experiment reduced its performance to 0.61 and 0.62 for Lung3 and H&N1, respectively.

Discussion: The experiments suggest that the performance of spherical features is comparable with previous approaches, but they generalize better when applied to unseen datasets. That might imply that orientation-independent shape features are relevant for SRP. The performance of the proposed method was very similar, using manual and automatic segmentation methods. This makes the proposed model useful in cases where expert annotations are not available or difficult to obtain.

Keywords: lung cancer, tumor segmentation, spherical convolutional neural network, survival rate prediction, deep learning, Cox Proportional Hazards, DeepSurv

OPEN ACCESS

Edited by:
Haibin Shi,
Soochow University, China

Reviewed by:
Wei Wei,
Xi'an Polytechnic University, China
Weiwei Zong,
Henry Ford Health System,
United States

****Correspondence:***
Fabian Sinzinger
fabiansi@kth.se

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *06 February 2022*
Accepted: *21 March 2022*
Published: *27 April 2022*

Citation:
Sinzinger F, Astaraki M, Smedby Ö and Moreno R (2022) Spherical Convolutional Neural Networks for Survival Rate Prediction in Cancer Patients. Front. Oncol. 12:870457. doi: 10.3389/fonc.2022.870457

INTRODUCTION

The objective of *Survival Rate Prediction* (SRP) is to estimate the time until a well-defined "terminal event", which occurs in some, but not necessarily all, cases. For cancer patients, the terminal event may be the death of the patient ("overall survival"), relapse, or progression of the disease ("relapse-free survival" or "progression-free survival", respectively). It has been shown that image-based characteristics of tumors such as shape, size and texture are associated with malignancy (1). A research avenue that has been explored in the last few years is whether those image-based tumor characteristics can also be used for predicting the survival of cancer patients (2). Survival rate prediction (SRP) from the shape, size, and texture of the tumor is challenging. First, it is not clear if imaging information alone is enough for SRP. Moreover, the prediction might be affected by different factors, including image acquisition parameters, inaccurate segmentation masks, the selected features used for the prediction, the prediction model itself, as well as the presence of right-censored data. Although clinical trials are often relied on clinical assessments like molecular profiling to conduct the survival analysis (3), such information is not always accessible.

While SRP could be framed as a regression-type problem, that is, to predict the time from the last observation to the terminal event, a practical difficulty is that part of the longitudinal data of patients is missing in training datasets for SRP in cancer. More specifically, these datasets usually contain right-censored data, which means that the start of the observation period is known for all data points, but the definitive end of the observation point might be missing for some cases. Consider a dataset where some patients were still alive when the study ended. In such an example, there would be a lower boundary of the survival times, namely, the last known date of record, which is lower than the definitive time of death for some cases. Other reasons for right-censorship in practice could be that patients dropped out of the study and did not have a time of death reported. However, it should be noted that exclusion of such cases is not recommended since that might bias the analysis towards the more lethal cases.

In the past, SRP was usually performed on small feature sets of descriptive statistics or clinical assessments (4). used ensemble data mining to train an outcome calculator on clinical data including features such as patient age at diagnosis, cancer grade, lymph node involvement, among many others. When working with imaging data, *radiomics* (5) provides a catalog of standard methods to extract such statistics automatically. These radiomics features can be used in CoxPH models (4) and other prediction methods such as decision trees, rule-based classification, or naive Bayes (6). More recently, *deep learning* (DL)-based methods have outperformed conventional algorithms in the field of image processing in general (7) and in image-based SRP in particular (8, 9). Among the first approaches using DL, Faraggi and Simon (10) proposed a feed-forward neural network for a non-linear risk-score approximation. A more recent example is DeepSurv, which provides a general framework for DL-based SRP (11).

Since the introduction of DL-based SRP, a vast body of work has been published where different DL algorithms have been applied to diverse modalities and features from various organs. Some examples include SPR for gastric cancer (12), cervical cancer (13), colorectal cancer (14), liver cancer (15), breast cancer (16) and oral cancer (17). In particular, this study is focused on non-small cell lung cancer (NSCLC) SRP. Previous studies from recent years have already shown the potential of DL models for survival analysis of lung cancer patients (18–21).

In some studies [e.g (22). and (23)], features from different modalities including imaging, radiomic features, clinical data, and molecular information, were combined as inputs to improve the performance of DL-based SRP models. While such multimodal prediction pipelines are theoretically superior to single modality-based predictions, the requirement for the respective data availability can be a disadvantage for the application in clinical practice. The financial cost of additional laboratory testing and expert clinical staging and tumor segmentation is another limiting factor of multimodal techniques. In addition, those approaches can only be applied to sites where the required data can be collected. Thus, it is clinically relevant to develop an SRP pipeline that requires only the CT scan of the lung region from the patient.

To our knowledge, previous studies have mainly used traditional convolutional neural networks (CNNs) for image-based SRP. One major issue of these types of neural networks is that their extracted features strongly depend on the spatial orientation of the tumor. That is, a rotated tumor can potentially get a different prediction by using traditional CNN. Instead, spherical CNNs (SphCNNs) are designed to be invariant against changes in orientation. Thus, SphCNNs are theoretically better suited for SRP. While traditional CNNs work with inputs structured in well-defined Cartesian grids, SphCNNs work with functions defined on the unit sphere. Thus, the use of SphCNNs for SRP requires a mapping from 3D CT images to functions on the unit sphere, which are intrinsically 2D. This dimensionality reduction has the additional effect that the derived DL models are less prone to overfitting in complex tasks with small datasets of 3D images (24). These reasons make it interesting to assess the ability of SphCNNs for SRP.

The aim of this study is to propose a fully-automatic solution for SRP of cancer patient data. First, we train a deep learning-based model that is able to segment tumors from CT images automatically. In a second step, we use spherical convolutional neural networks (SphCNNs) to perform deep feature extraction for SRP. To our knowledge, such spherical features SphCNNs have not been used in this context before. Thus, we also compare our SphCNN-based pipeline against more traditional methods using different prediction models for SRP on radiomic features or features extracted from fine-tuned DL-based pre-trained classifiers.

The remainder of this paper is structured in the following way. Section 2 establishes a general framework for SRP consisting of three stages: tumor segmentation, feature extraction, and survival prediction. Next, we describe how our proposed pipeline implements each of those stages. Moreover, the implemented baseline methods are described. Section 3 lists the experimental results comparing the proposed method with the baseline models. Section 4 discusses the findings from the experimental evaluations. Finally, section 5 reveals the main implications of the results and makes some conclusions of the study.

MATERIALS AND METHODS

In the context of this study, we model the SRP of cancer patients as a three-stage process, consisting of *segmentation*, *feature extraction*, and *survival prediction* (cf. **Figure 1**).

I. **Tumor Segmentation** describes the process of defining which of the voxels belong to the object of interest, that is, which parts of the CT image depict the cancerous mass. Therefore, a binary mask is generated either by manual annotation through a medical expert or an algorithmic segmentation method.
II. **Feature Extraction** is the transformation of high dimensional input data (in our case, segmented regions of the image) into fewer but more relevant features.
III. **Survival Prediction** takes the previously extracted features and determines the respective value of interest.

We will refer to these three stages when comparing different prediction pipelines in the experiments. The following subsections specify the methods we propose for each of the three SRP stages.

Tumor Segmentation

This study aims to introduce an end-to-end solution for SRP that does not rely on manual tumor segmentations. Therefore, we incorporate a fully automatic lung nodule segmentation model, concretely, the lung cancer detection and segmentation method we proposed in (25). This method decomposes the segmentation problem into three separate steps, as shown in **Figure 2**.

First, an in-painting network (26) is trained to fill randomly generated holes in Lung-CT images from healthy subjects. The resulting network can fill missing parts of an image with semantically meaningful patterns. By considering the annotated tumor regions of the unhealthy images as missing content, the in-painting network is used to generate healthy synthetic images from the unhealthy counterparts.

Second, the resulting healthy-unhealthy image pairs are used to train a normal appearance autoencoder (NAA). Here, the unhealthy images serve as an input and the healthy synthetic images as corresponding target images for the supervised training of the NAA model. Therefore, the trained NAA can generate tumor-free images from arbitrary unhealthy images without depending on manual annotation masks.

In the final stage, the original (unhealthy) image and the difference between the original image and the NAA-generated healthy outputs are fed to a standard U-Net segmentation model. The U-Net model benefits from this attention cue to learn the final segmentation mask by receiving the original and difference-image as separate channels. The method is described in more detail in (25). Performance metrics of this method for the datasets that are relevant for this study are presented in **Table 1**.

Feature Extraction

This section discusses how our pipeline extracts descriptive variables that are meaningful for the prediction task from the raw data, i.e., the lung-CT images.

SphCNNs (27, 28) extend the standard operations used by traditional Cartesian CNNs to work on signals defined on the sphere. The network topology of SphCNN consists of stacks of spherical filters that are applied on the spherical activation signals *via* spherical convolution (cf. **Figure 3**). The convolution operation is often carried out as a multiplication in the spherical harmonics domain. One characteristic property of SphCNN is that it can be used for solving problems where rotational equivariance (i.e., the output rotates when the input is rotated) or rotational invariance (i.e., the output is always the same even if the input is rotated) is required (28). As mentioned, SRP should be rotational invariant, which means that the prediction should be the same regardless of the orientation of the tumor in the lungs. Our implementation builds upon the code provided in (27).

In order to apply SphCNN on volumetric CT images, it is necessary to map the segmented tumor onto the unit sphere S^2.

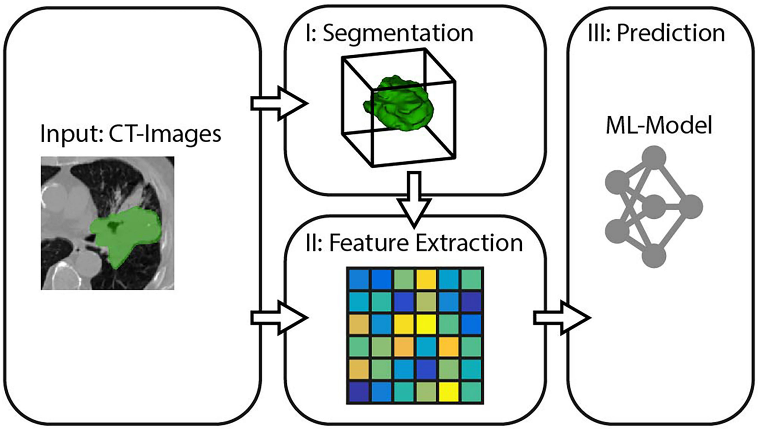


FIGURE 1 | General pipeline for survival rate prediction.

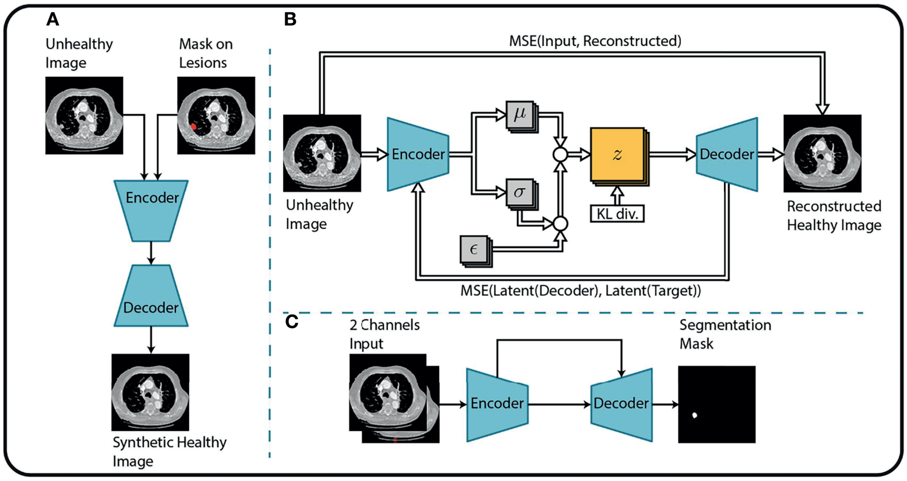


FIGURE 2 | Overview of the tumor segmentation method. The segmentation method incorporates **(A)** an image inpainting network, **(B)** a variational autoencoder, and **(C)** a U-Net for the final segmentation. Replicated from (25) with permission from Springer Nature Switzerland AG.

TABLE 1 | Performance metrics (mean ± std) of the automatic segmentation methods evaluated on two lung cancer datasets.

Dataset	Dice	Precision	Recall	Specificity
Lung1	0.77 ± 0.17	0.76 ± 0.20	0.82 ± 0.15	1.00 ± 0.00
Lung3	0.76 ± 0.18	0.74 ± 0.22	0.85 ± 0.16	1.00 ± 0.00

For the training dataset Lung1, the observed values are averaged over five evaluation folds. For the validation dataset Lung3, the values are averaged over all samples. Note that we rounded to two digits so 1.00 in the last column results from rounding a value close to one.

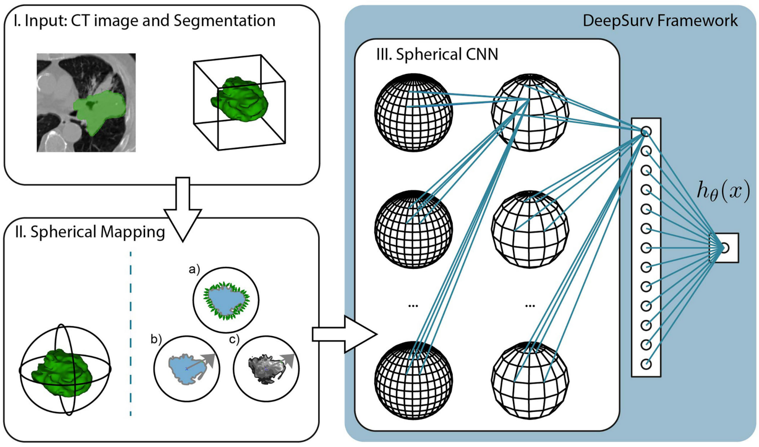


FIGURE 3 | Pipeline of the proposed method, which is divided into three steps. I. The input consists of the CT image and segmentation of the tumor mass. The experiments compare the predictive performance of provided manual segmentation masks with our automatic segmentation. II. The volumetric images are projected into the spherical domain to be usable with Spherical CNNs. In this study, we propose three spherical mappings; a) the extended Gaussian image (EGI), b) the depth-based projection of the mask (b), and c) the spherical intensity mapping of the masked image content. III. The Spherical CNN consists of a cascade of spherical kernel stacks followed by spherical pooling operations. The Spherical CNN is embedded in the DeepSurv framework that includes a fully connected layer that pass the activation signal to a single output node. This scalar output is the approximation of the log-risk function h_θ (*x*) in Cox proportional hazards model which is optimized through DeepSurv.

We propose three different mapping methods for this projection, a) the extended Gaussian image (EGI) of the tumor mask, which is the orientation distribution function of the normal vectors from the surface of the tumor (29), b) a depth-based projection (30) of the segmentation mask, and c) an intensity-based projection of the tumor. As for the EGI, it is generated from the normal vectors derived from the provided manual segmentations or the generated automatic segmentation masks (cf. **Figure 4A**). Regarding the depth-based projection, first, an enclosing sphere is centered at the tumor's center of mass. Next, a ray is cast from each sampling point on the surface of the sphere to the centroid. The distance to the first intersection point then decides the value of the spherical signal at that specific orientation (cf. **Figure 4B**). For an alternative mapping, we accumulate the intensity values within the tumor along every ray (cf. **Figure 4C**). These three functions on the sphere are used as input channels for the SphCNN. These three functions on the sphere are used as input channels for the SphCNN. Notably, we explore two configurations here; the first input configuration only uses the segmentations' depth-based projection (later referred to as SphCNN[1]). The second uses the EGI and the intensity-based projection from the image (SphCNN[2] in the following). This choice of input channels is motivated by the question of whether the image content carries additional predictive power to the use of the segmentation mask alone.

Prior to the comparative experimental evaluation presented in the results section, we empirically determined a suitable network topology for our purpose. Those tests uncovered that a deeper SphCNN was not beneficial over a more shallow architecture for the given problem. Therefore, the best-performing model consists of three layers. The first convolutional layer lifts the input signal from the sphere, S^2 onto the SO (3) manifold. Next, the spherical activation maps are fed to another convolutional layer [operating on SO (3)] and, finally, a dense layer that connects *via* linear activation function to the scalar output neuron. Interposed spherical pooling layers condense the spatial dimension of the activation maps. The last fully-connected layer encodes 40 features. The configuration of the empirically determined training parameters used in the experiments is provided in **APPENDIX A**.

Survival Prediction

In this paper, we aim to predict the relative risk of a patient and the chance of survival for different times. Every longitudinal entry in the clinical datasets records the observation time T and a binary event variable E, which indicates whether the event of death occurred at time T. T represents the actual survival time when E is equal to one (representing the state 'True'). However, if E is equal to zero (representing the state 'False'), the data entry is considered as *right-censored*, and T can only be seen as a lower bound for the actual unknown time of survival.

One possible approach to handle this type of data could be to disregard all data points with $E \neq 1$ and perform regression on the remaining data. However, as mentioned previously, this approach would bias the method towards the subjects with higher mortality. Instead, the problem of SRP under the presence of right-censored data is commonly modeled *via survivor-* and *hazard functions*.

The standard method of handling SRP on right-censored data is the *Cox's Proportional Hazard* model (CoxPH) (31). Cox (31) defined the survivor function as $F(t)=P(T \geq t)$, that is, the probability P of the actual death of the patient to be larger or equal to the time t. Cox also defined the hazard function λ(t) which models the age-specific failure rate as:

$$\lambda(t) = \lim_{\Delta t \to 0_+} \frac{1}{\Delta t} P(t \langle T \langle t + \Delta t | t \leq T). \tag{1}$$

He proposed the *Cox proportional hazards model* (CoxPH) to approximate the hazard function as:

$$\lambda(t|x) = \lambda_0(t) \cdot e^{h(x)} = \lambda_0(t) \cdot e^{\beta^T x}, \tag{2}$$

Where λ_0(t) is the (unknown) baseline hazard, β is the model parameter vector, $h(x)$ is the so-called *log-risk* function, and x are

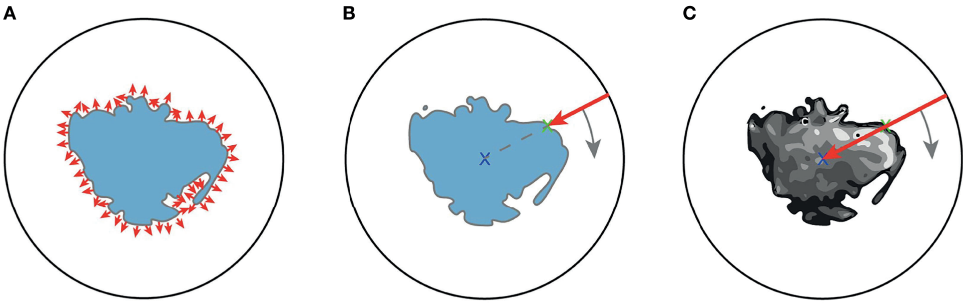


FIGURE 4 | Illustrative depiction of the three proposed spherical mappings. Note that volumes are here drawn as image slices, and therefore, the spheres are depicted as circles. **(A)** The extended Gaussian image (EGI) can be viewed as an accumulation of the gradient vectors (small red arrows) at the surface of the tumoral boundary. **(B)** Depth-based projection of the solid segmentation mask. A ray (red arrow) is cast from a projecting sphere to the surface of the segmentation. The distance from the sphere to the surface determines the value of the spherical signal at the respective position. **(C)** Intensity-based projection of the voxel image content. A ray (red arrow) is cast from the surrounding sphere through the segmented tumor image towards the centroid. The value of the spherical signal is the sum of all intensities of the voxels that the ray traversed.

covariates. Note that in our specific problem, the covariates are the features extracted from the sample, as discussed in the previous subsection.

One well-known restriction of the CoxPH is the assumption that $h(x)$ is linear, i.e. $h(x) = \beta^T x$, which can limit the capability of the function to model SRP. DeepSurv (11) tackled this problem by training a neural network to approximate $h(x)$ which is able to model non-linearities in the hazard function. Thus, the hazard function in DeepSurv becomes:

$$\lambda(t|x) = \lambda_0(t) \cdot e^{h_\theta(x)}, \tag{3}$$

where $h \approx h_\theta(x)$ with θ being the learned parameters of the neural network.

One advantage of DeepSurv is that it is more than a method, it is a generic pipeline that can easily be connected to a feature extraction neural network. In the original paper, DeepSurv used a set of fully-connected layers followed by a linear combination layer to estimate $h_\theta(x)$. Instead of fully-connected layers, we used the SphCNN described in the previous section while keeping the same loss function that aims to minimize the *average negative log partial likelihood* of $h(x)$, as described in (11).

Beyond DeepSurv, a family of techniques that aim to address the shortcomings of CoxPH are large-margin methods such as regression or ranking-based support vector machines (SVMs) (32, 33). Other techniques that have been applied successfully for SRP are ensemble models that use, e.g., gradient-boosting to learn a partial likelihood function (34). Notice that these methods can only be used when the feature extraction is independent of the survival prediction model, which is not our case. Thus, DeepSurv is a well-suited choice for combining feature extraction and prediction simultaneously and is therefore used in the proposed method.

Baseline Methods

In order to assess the relative performance of the proposed method, we compared it against multiple feature sets and prediction method combinations. As for the features, we computed radiomics features (RF) (5, 22) and deep learning-based 2D (slice-based) image features (DIF) (22). Instead of the pre-trained neural network used in (22), we used ResNet50 (35), which is very well-known for its good performance in transfer learning tasks. In particular, the DIF features were extracted from the 2D axial slice with the largest tumor area in the segmentation mask with the pre-trained ResNet50. The RF and DIF sets consist of ca. 1,500 and 1,000 features, respectively. Moreover, subsets of 32 features were extracted from RF and DIF after a feature selection procedure, which are referred to as RF32 and DIF32, respectively (more details are provided in **APPENDIX B**). For this, we used the library function from *scikit-learn* (36) to rank each regressor (i.e., each entry of the extracted feature vector) based on its cross-correlation with the target.

As survival prediction methods we used support vector machines with ranking (SVM-K) and regression (SVM-R) objective (32), CoxPH (31, 37), and the gradient boosting-based ensemble (EGB) model proposed in (34). Thus, we implemented the sixteen combinations of four features sets and four survival prediction methods. In addition, we implemented the method proposed by Aerts et al. (2) in which CoxPH is applied to the so-called radiomics signature that consists of four radiomic features. This method is referred to as RS-CoxPH in the experiments. Hyperparameters such as the learning rate or method-specific engineering values were empirically tuned in a set of preceding tests.

RESULTS

We performed intra- and inter-dataset experiments with different pipelines and dataset configurations to assess the model performance and robustness of the different methods as shown in **Figure 5**. In particular, we trained our models on the CT data from the publicly available Lung1 (n=422) (38), Lung3 (n=89) (39), and H&N1 (n=136) (40) datasets. While Lung1 and Lung3 contains data from Non-Small Cell Lung Cancer (NSCLC) patients acquired in different institutions, H&N1 depicts head and neck cancer. We used as ground truth prediction values the right-censored times of survival that were reported from the respective data providers.

We used the concordance index (C-Index) (41) as our main performance criterion, which is commonly used for problems with right-censored data like SRP. The C-Index measures how good the survival times of a set of patients are ranked and can be seen as a generalization of the area under the receiver operating characteristic (ROC) curve (AUROC) that can take into account right-censored data. We used the implementation of the C-Index from the python library *scikit-survival*[1].

Results for Lung1

Intra-dataset performance was assessed using 5-fold cross-validation on Lung1. We kept the fold splits consistent for all evaluated methods. In addition to the 17 baseline methods described in Subsection 2.4, we tested the proposed method with both manually annotated segmentation masks and automatic masks generated by the method described in Sect. 2.1.

Figure 6 shows the observed C-indices of the 5-fold cross-validation experiment. As shown, the combination of EGB and DIF32 obtained the best performance with a C-index of 0.62 ± 0.04 in this experiment, while the worst performance was measured on SVM-R with DLF32: 0.38 ± 0.04. In comparison, the proposed method achieved 0.58 ± 0.04 for the manual masks and 0.59 ± 0.03 for the automatic ones.

Inter-Dataset Evaluation

In order to assess the robustness of the methods, inter-dataset validation was carried out by training the methods (including the automatic segmentation) on Lung1 and validating on additional images from a different dataset. In particular, we used the models that were fitted to Lung1 for inference on two independent datasets: Lung3 and H&N1. As mentioned, Lung3 has the same type of patients (i.e. NSCLC-patients), while H&N1 contains images of patients with head and neck cancer.

[1] https://scikit-survival.readthedocs.io/en/stable/api/metrics.html

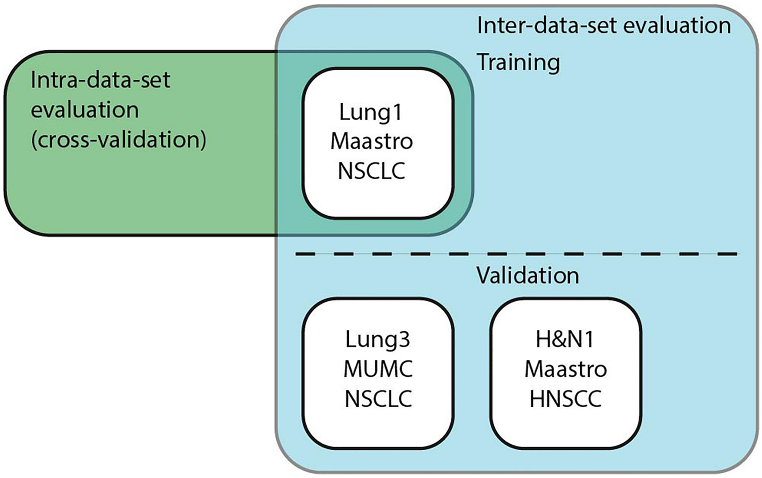


FIGURE 5 | Schematic overview of the experiments. **Intra-data-set** uses Lung1 for training and testing in a cross-validation setup. For **Inter-dataset** evaluation, methods were trained on Lung1 and evaluated on Lung3 or H&N1.

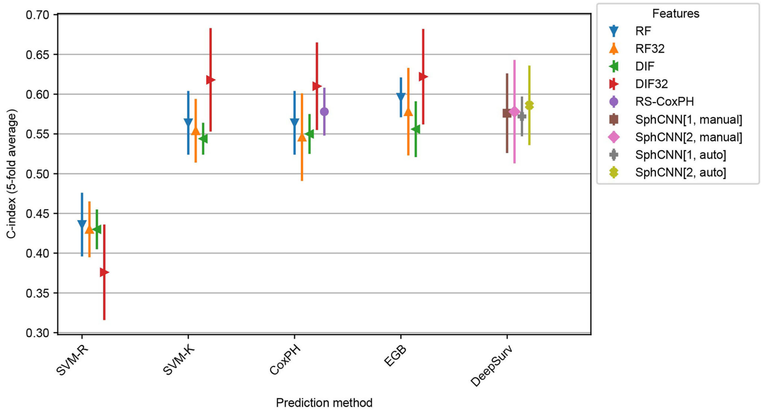


FIGURE 6 | Cross evaluation results for 17 baseline SRP methods and the proposed one. The models were trained on four splits from the Lung1 Data and evaluated on the remaining fifth split. We report the average across the splits (marker) as well as the observed minimum and maximum observed values (line). Compared prediction methods are Support Vector Machines with regression (SVM-R) and ranging (SVM-K) objective, Cox Proportional Hazards model (CoxPH), Ensemble Gradient Boosting (EGB) and DeepSurv, a deep-learning-based prediction framework. Baseline features are Radiomics Features (RF) and pre-trained deep 2D Image Features (DIF). Both feature sets were also used with feature selection (RF32 and DIF32 respectively). In addition, we also include the Radiomics Signature (RS-CoxPH) suggested by Aerts et al. (2) in out comparison. Our proposed method uses a Spherical Convolutional Neural Network (SphCNN) with manual (SphCNN [..., manual]) and automatic (SphCNN[..., auto]) tumor segmentation. The spherical input for the SphCNN is either extracted *via* depth-image projection from the segmentaion mask (SphCNN[1,...]) or composed of intensity projection and extended gaussian image from the CT-image (SphCNN[2,...]).

Results of these experiments are reported in **Figures 7** and **8**. The best-performing method was the proposed one (C-index 0.64 both for Lung3 and H&N1), followed by EGB with RF32 (C-index 0.63 both for Lung3 and H&N1). EGB with DLF32 - the best method in the previous experiment - decreased its performance in this test to 0.61 for Lung3 and 0.62 for H&N1, respectively.

Since our automatic segmentation method was trained and developed for lung cancer, it did not yield meaningful segmentation results for CT images from head and neck regions. Actually, it is well-known that automatic segmentation of head and neck cancer is a very difficult task (42). Thus, our methods were tested only with the manually annotated masks provided in the datasets.

As shown in the experiments, the proposed method performed better when all spherical mappings were used. As expected, the proposed method yields slightly better results with manual segmentations compared to the use of automatic segmentation.

Kaplan-Meier Analysis

Validation datasets from the respective experiment were stratified according to our best-performing method's assigned

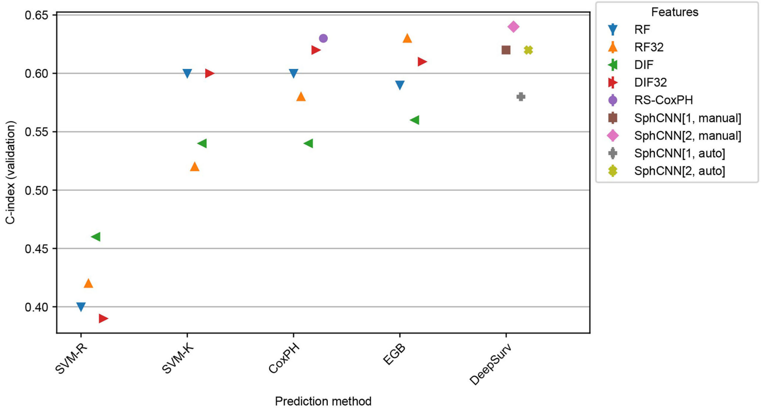


FIGURE 7 | Performance comparison for the models trained on Lung1 and evaluated on Lung3. We used the prediction methods and features as labelled in **Figure 6**.

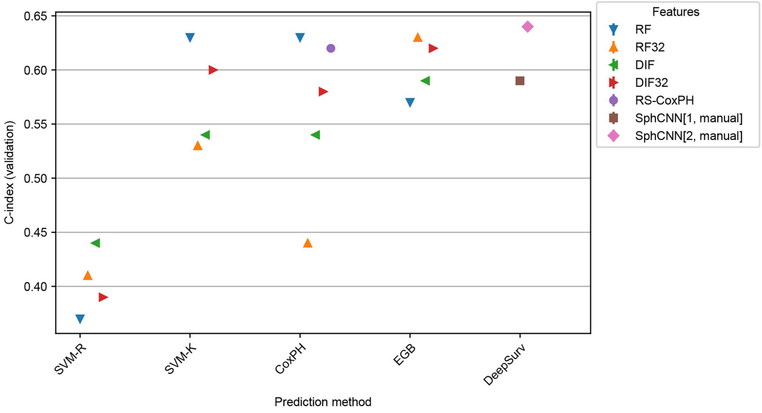


FIGURE 8 | Performance comparison for the models trained on Lung1 and evaluated on H&N1. We used the prediction methods and features as labelled in **Figure 6** with the exception that no automatic segmentation was evaluated here.

risk score (i.e., SphCNN[2]). The stratification into risk groups was done based on the median of the predicted risk scores. Therefore, half of the higher-risk samples were binned into one group and the other half into another. Then, non-parametric Kaplan-Meier (KM) estimations were evaluated on each group separately (cf. **Figure 9**).

As shown, the KM curves also reflect the relative performance assessment reported in the previous subsections. Concretely, the evaluation folds of the cross-validation experiments reported mixed observations regarding their performance. In contrast, our method showed promising stratification abilities when tested on the external datasets (Lung3, H&N1). There is also a potential trend observable regarding the type of cancer. Lung cancer data has good short-term separability but often fails for long-term prediction (over five years). In contrast, head and neck cancer data showed more confident separation (i.e., the survival curves become more distant from each other) for periods larger than ten years.

DISCUSSION

The primary goal of this study was to introduce a novel fully-automatic lung cancer SRP model based on CT-images. In addition, we performed a benchmark with various SRP models evaluated on publicly available data. We run our experiments in both inter and intra dataset evaluation schemes. This section discusses some findings and analytical aspects of the presented investigations.

Perhaps the most apparent observation is that the C-index values from SVM-R scored lower than all other methods across different features. This difference in prediction performance confirms the previous claim that regression methods are poorly fit for working with right-censored data. Historically, this misfit motivated the development of ranking and hazard-based SRP models. Thus, we will focus the discussion on the remaining prediction methods.

In contrast, we found that pre-trained ResNet50 with feature selection (DIF32) had the highest predictive power in the

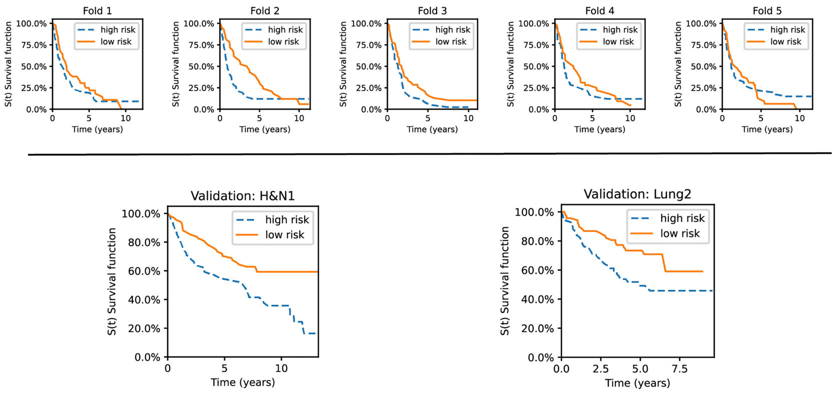


FIGURE 9 | Kaplan–Meier curves for the validation sets used in the experiments. The datasets are stratified into low- and high-risk groups based on the risk predictions of our best-performing method. Top row: survival curves from five individual folds from Lung1. Bottom: The model was trained on Lung1 and evaluated on the Lung3 (right) and H&N1 (left) datasets, respectively.

cross-validation experiment. Notice that the original set of features (DIF) did not perform well in the experiments. That means that the boost in performance is mainly due to the feature selection method. We like to emphasize that the reported use of DIF with feature selection entails the risk of overfitting the given training set. Since only the 32 highest-rated regressors were selected on the training data, the possibility arises that this selection might be biased towards the respective training dataset. By using additional validation datasets, we confirmed the predictive power of this model but, in that case, we did not observe any advantage of this method over the other tested methods. In comparison, the proposed method used 40 features that, according to the results, are enough to encode the most relevant information for SRP. Thus, the results support that a small set of spherical features or DIF are beneficial for SRP.

Regarding radiomic features, feature selection resulted in worse performance except for EGB in the inter-dataset evaluation. In specific, it should be noticed that the employed cross-correlation based feature selection method aim to hold only those subset of the features that are more linearly correlated with respect to the class labels statistically. However, this linear statistical association does not necessarily represent their more prognostic values of the feature subsets. Accordingly, although the selected subset of radiomic features is more correlated with the target values, their prediction power is not as high as the whole radiomic feature set. In addition, the observation that such a feature selection method leads to improving the performance of DL-based features but not the radiomic features can be explained by the fact that DL-based features were extracted from a single 2D slice, i.e., the central tumoral slice, while the radiomics descriptors were extracted from the tumor volumes. Therefore, the less complicated attributes of the tumors in 2D slices which were captured by the DL model, are more prone to show stronger association with the target labels compared against the 3D radiomics descriptors that were extracted from the irregular tumor volumes with a large variety of texture, intensity, and morphological characteristics.

Furthermore, we noted that except for SVM-R, all evaluated prediction methods resulted in C-indices at a comparable level. In our experimental setup, the individual remaining prediction methods do not seem to have an advantage over other prediction methods. Besides, the differences in the feature extraction methods were mostly consistent for different prediction models and datasets. Therefore, we observe that the choice of extracted features is much more important for designing a successful SRP pipeline than the selection of the prediction model, as long as they are designed to handle right-censored data.

Our proposed method automatically extracts morphological features in the spherical domain. Intuitively, our spherical mapping methods can be understood as a compact representation of tumor surface texture, size, shape, and internal structure. Using such spherical signals combined with a rotation-invariant SphCNNs, we obtained C-indices comparable to conventional methods on the cross-validation experiment. Moreover, the proposed method slightly surpassed the others when referring to the inter-dataset evaluations. Our results suggest that the proposed SphCNN-based SRP is robust when applied to new, unknown datasets. The observed statistics also indicate a similar accuracy on both the lung cancer data and the head and neck images. This finding hints that the morphological features that the SphCNN internalized during training might have prognostic relevance for tumors in general. However, since the differences between the proposed method and the best-performing baselines were small, we can only argue that the proposed method has overall competitive performance.

In clinical settings, it may be difficult or unfeasible to have high-quality annotated segmentation masks of the tumors. For that reason, it is relevant to have a fully automated solution that includes an automatic segmentation tool. Since manual annotations performed by experts have higher quality than segmentations from automatic tools, we expected a reduction in the performance of the proposed SRP method when used on automatically segmented tumors. From the results, such a reduction was slightly negative for the intra-dataset experiment

(0.58 ± 0.04 vs. 0.59 ± 0.03) and very small for the inter-dataset evaluation (0.64 vs. 0.62). This means that our proposed pipeline can yield similar results when it is run autonomously without a manual - and potentially expensive - human intervention.

To gain insights if the segmentation mask or the segmented image channels are beneficial for SRP, we tested our method in two different configurations, SphCNN[1] and SphCNN[2]. While the former represents a higher compressed version of the signal, the latter is assumed to preserve more structural information. Our reported results support the assumption that SphCNN[2] is slightly superior in this context.

As mentioned, DeepSurv is a general pipeline that can potentially be combined with any feature set, including RF and DIF. We did not include these combinations in the experiments since that would require a fine-tuning of the architecture of the neural network for every specific feature set, which is out of the scope of this study.

Since the implementation of methods by different research groups can yield different results, we decided to implement 17 baseline methods in order to have a more fair comparison. The performance of all tested methods was below 0.65, which is consistent with previous studies [e.g (2, 21).,]. That means that, although CT images convey important information for SRP, they should be complemented with other types of information to improve the predictions to a level that can be used in clinics.

Limitations of the Study

The main limitation of the study is the number of available images. It is well-known that DL-based methods require large datasets that are relatively scarce in cancer research at present. Thus, the main findings of this study require further validation with larger datasets. That could help to rule out the possibility that the differences in performance are related to the specific characteristics of the datasets. In this study, we avoided overfitting by using two strategies: a) dimensionality reduction by mapping the 3D data into 2D spherical mappings and b) the architecture of the proposed SphCNN is relatively small and has just 40 features in the penultimate activation. While dimensionality reduction will always be beneficial and needed, using larger datasets would enable us to evaluate larger SphCNN architectures with more parameters bigger feature vectors and overall higher capacity.

Another potential downside of the proposed solution is the representation of the spherical images and activation functions. The spherical signals are represented as a regular 2D grid in the implemented pipeline. While this common practice allows easy integration into the SphCNN framework, it might introduce distortions in the image due to the lack of equidistant sampling on the sphere. Concretely, regions close to the poles are oversampled compared to the equator. An alternative approach that samples the sphere more uniformly is described in (43). It is unclear at this point if and how this change of sampling can affect the predictions; therefore, more research would be required.

CONCLUSION

This work introduced a new method for image-based lung cancer SRP. For automatic, relevant feature extraction, we mapped the tumor extracted with a DL-based method into a spherical domain and used SphCNN for prediction. The experimental evaluation confirmed the competitive predictive power of our model when compared to state-of-the-art approaches on the Lung1 data. A slight advantage over the other techniques was observed when tested on data from additional datasets (Lung3, H&N1). The results support that SphCNNs are helpful for attaining rotational invariance in SRP problems.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

FS: Conceptualization, Methodology, Implementation, Formal analysis, Writing - original draft, Visualizations. RM: Conceptualization, Methodology, Writing - review and editing, Supervision, Project administration, MA: Methodology, Implementation, Writing - review and editing, ÖS: Writing - review and editing, Supervision, Resources, Funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This study has partially been funded by the Swedish Research Council, [grant no. VR-2018-04375] and the Swedish Childhood Cancer Foundation [grant no. MT2016-0016]. The funding agencies had no role in research or preparation of this study.

ACKNOWLEDGMENTS

We thank NVIDIA Corp. for the donation of the Titan Xp GPU that was used for the computations of this study.

REFERENCES

1. Astaraki M, Zakko Y, Toma Dasu I, Smedby Ö, Wang C. Benign-Malignant Pulmonary Nodule Classification in Low-Dose CT With Convolutional Features. *Physica Med* (2021) 83:146–53. doi: 10.1016/J.EJMP.2021.03.013
2. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. *Nat Commun* (2014) 5:1–9. doi: 10.1038/ncomms5006
3. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving Breast Cancer Survival Analysis Through Competition-Based

Multidimensional Modeling. *PloS Comput Biol* (2013) 9:e1003047. doi: 10.1371/JOURNAL.PCBI.1003047

4. Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. (2011). A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: Association for Computing Machinery, San Diego California. doi: 10.1145/2003351.2003356
5. van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* (2017) 77:e104–7. doi: 10.1158/0008-5472.CAN-17-0339
6. Hawkins SH, Korecki JN, Balagurunathan Y, Gu Y, Kumar V, Basu S, et al. Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features. *IEEE Access* (2014) 2:1418–26. doi: 10.1109/ACCESS.2014.2373335
7. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
8. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting Cancer Outcomes With Radiomics and Artificial Intelligence in Radiology. *Nat Rev Clin Oncol* (2021) 19:132–46. doi: 10.1038/s41571-021-00560-7
9. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection. *Genome Med* (2021) 13:1–17. doi: 10.1186/S13073-021-00968-X
10. Faraggi D, Simon R. A Neural Network Model for Survival Data. *Stat Med* (1995) 14:73–82. doi: 10.1002/SIM.4780140108
11. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Med Res Method* (2018) 18:1–12. doi: 10.1186/s12874-018-0482-1
12. Zhang L, Dong D, Zhang W, Hao X, Fang M, Wang S, et al. A Deep Learning Risk Prediction Model for Overall Survival in Patients With Gastric Cancer: A Multicenter Study. *Radiotheraph Oncol* (2020) 150:73–80. doi: 10.1016/j.radonc.2020.06.010
13. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, et al. Survival Outcome Prediction in Cervical Cancer: Cox Models vs Deep-Learning Model. *Am J Obstetrics Gynecol* (2019) 220:1–381. doi: 10.1016/J.AJOG.2018.12.030
14. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep Learning for Prediction of Colorectal Cancer Outcome: A Discovery and Validation Study. *Lancet* (2020) 395:350–60. doi: 10.1016/S0140-6736(19)32998-8
15. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* (2018) 24:1248–59. doi: 10.1158/1078-0432.CCR-17-0853
16. Gupta S. Gupta MK. A Comparative Analysis of Deep Learning Approaches for Predicting Breast Cancer Survivability. *Arch Comput Methods Eng* (2021) 1:1–17. doi: 10.1007/S11831-021-09679-3/TABLES/11
17. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep Learning-Based Survival Prediction of Oral Cancer Patients. *Sci Rep* (2019) 9:1–10. doi: 10.1038/s41598-019-43372-7
18. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep Learning for Lung Cancer Prognostication: A Retrospective Multi-Cohort Radiomics Study. *PloS Med* (2018) 15:1–25. doi: 10.1371/JOURNAL.PMED.1002711
19. Hosny A, Aerts HJ, Mak RH. Handcrafted Versus Deep Learning Radiomics for Prediction of Cancer Therapy Response. *Lancet Digital Health* (2019) 1: e106–7. doi: 10.1016/S2589-7500(19)30062-7
20. Kim H, Mo Goo J, Hee Lee K, Kim YT, Park CM. Preoperative Ct-Based Deep Learning Model for Predicting Disease-Free Survival in Patients With Lung Adenocarcinomas. *Radiology* (2020) 296:216–24. doi: 10.1148/RADIOL.2020192764/ASSET/IMAGES/LARGE/RADIOL.2020192764.FIG5D.JPEG
21. Kadoya N, Tanaka S, Kajikawa T, Tanabe S, Abe K, Nakajima Y, et al. Homology-Based Radiomic Features for Prediction of the Prognosis of Lung Cancer Based on CT-Based Radiomics. *Med Phys* (2020) 47:2197–205. doi: 10.1002/mp.14104
22. Paul R, Hawkins SH, Hall LO, Goldgof DB, Gillies RJ. Combining Deep Neural Network and Traditional Image Features to Improve Survival Prediction Accuracy for Lung Cancer Patients From Diagnostic CT, in: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE (2016). 2570–5. doi: 10.1109/SMC.2016.7844626
23. Vale-Silva LA, Rohr K. Long-Term Cancer Survival Prediction Using Multimodal Deep Learning. *Sci Rep* (2021) 11:13505. doi: 10.1038/s41598-021-92799-4
24. Sinzinger F, van Kerkvoorde J, Pahr DH, Moreno R. Predicting the Trabecular Bone Apparent Stiffness Tensor With Spherical Convolutional Neural Networks. *Bone Rep* (2022) 16:101179. doi: 10.1016/j.bonr.2022.101179
25. Astaraki M, Toma-Dasu I, Smedby Ã, Wang C. Normal Appearance Autoencoder for Lung Cancer Detection and Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science. Springer, Cham (2019) 11769:249–56. doi: 10.1007/978-3-030-32226-7_28
26. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B. Image Inpainting for Irregular Holes Using Partial Convolutions. Proceedings of the European conference on computer vision (ECCV) (2018) 11215:89–105. doi: 10.1007/978-3-030-01252-6_6
27. Cohen TS, Geiger M, Koehler J, Welling M. (2018). Spherical CNNs, in: *6th International Conference on Learning Representations (ICLR)*.
28. Esteves C, Allen-Blanchette C, Makadia A, Daniilidis K. Learning SO(3) Equivariant Representations With Spherical CNNs. *arXiv* Cham: Springer International Publishing (2018) 11217:54–70. doi: 10.1007/978-3-030-01261-8
29. Horn B. Extended Gaussian Images, in: *Proceedings of the IEEE*, (1984) 72:1671–86. doi: 10.1109/PROC.1984.13073
30. Cao Z, Huang Q, Karthik R. (2017). 3d Object Classification via Spherical Projections, in: *2017 International Conference on 3D Vision (3DV)*, pp. 566–74. IEEE. doi: 10.1109/3DV.2017.00070
31. Cox DR. Regression Models and Life-Tables. *J R Stat Society: Ser B (Methodological)* (1972) 34:187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x
32. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support Vector Methods for Survival Analysis: A Comparison Between Ranking and Regression Approaches. *Artif Intell Med* (2011) 53:107–18. doi: 10.1016/j.artmed.2011.06.006
33. Pölsterl S, Navab N, Katouzian A. Fast Training of Support Vector Machines for Survival Analysis. (2015), 243–59. doi: 10.1007/978-3-319-23525-7_15
34. Friedman JH. Stochastic Gradient Boosting. *Comput Stat Data Anal* (2002) 38:367–78. doi: 10.1016/S0167-9473(01)00065-2
35. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York: Las Vegas, NV, USA (2015). pp. 770–8. doi: 10.1109/CVPR.2016.90
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
37. Raykar VC, Steck H, Krishnapuram B, Dehing-Oberije C, Lambin P. On Ranking in Survival Analysis: Bounds on the Concordance Index. *Adv Neural Inf Process Syst* (2007) 20:1209–16.
38. Aerts HJWL, Wee L, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, et al. Data From NSCLC-Radiomics [Data Set]. (2019). doi: 10.7937/K9/TCIA.2015.PF0M9REI
39. Aerts H, Rios Velazquez E, Leijenaar R, Parmar C, Grossmann P, Carvalho S, et al. Data From NSCLC-Radiomics-Genomics [Data Set]. (2015). doi: 10.7937/K9/TCIA.2015.L4FRET6Z
40. Wee L, Dekker A. Data From Head-Neck-Radiomics-HN1 [Data Set]. (2019). doi: 10.7937/tcia.2019.8kap372n
41. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures With Censored Survival Data. *Stat Med* (2011) 30:1105–17. doi: 10.1002/sim.4154
42. Andrearczyk V, Oreiller V, Jreige M, Vallières M, Castelli J, Elhalawani H, et al. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. In: *Lecture Notes in Computer Science*, vol. 12603. . Deutschland GmbH: Springer Science and Business Media (2020). p. 1–21. doi: 10.1007/978-3-030-67194-5_1
43. Coors B, Condurache AP, Geiger A. SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images. *Proc ECCV* (2018) 11213:518–32. doi: 10.1007/978-3-030-01240-3_32

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A. TRAINING CONFIGURATION OF THE PROPOSED METHOD

The training configuration of the SphCNNs used in this study is reported in **Table 2**.

APPENDIX B. Feature Selection for Baseline Methods

Motivated by the study of Aerts et al. (2), we decided to add a basic feature selection approach to our baseline feature extraction methods. The basic idea of feature selection in this context is to reduce the size of the feature vector that is presented to the prediction model for better convergence. Therefore, the cross-correlation of the features with the ground truth was measured on the respective training set. Next, only the features with the top 32 correlation scores were selected for prediction. The number 32 was selected to have a similar order of magnitude to the deep spherical features (i.e., the number of neurons in the penultimate layer of the SphCNN). Other feature vector sizes might lead to different results.

TABLE 2 | Configuration of the SphCNN prediction models.

Batch-size:	32
Number of epochs:	1000
Learning-rate:	Epoch 0-500: 1e-3 - 1e-5 (decreasing), Epoch 500-1000: 1e-5 (constant).
Optimizer:	ADAM
Dropout:	Yes, Rate 0.01
Batch-normalization:	Yes
Normalize inputs:	Yes

B.1. Feature Selection for Radiomics Features

In (2), features were selected based on their score within one of four categories (tumor intensity, tumor shape, tumor texture, and wavelet). In contrast, we score the cross-correlation as mentioned above on the complete set of approx 1500 radiomics features. A posthoc inspection of the selected radiomics revealed that the biggest group of selected features (18 out of 32) were first order-based statistics (mainly energy and total energy from different wavelet levels). The next biggest group were Gray Level Run Length Matrix components (8 out of 32), followed by Gray Level Size Zone Matrix entries (6 out of 32). For a detailed explanation of the different types of radiomics features, we refer to the documentation of *pyradiomics*.[2]

B.2. Feature Selection for Deep ResNet50 Features

Feature extraction *via* a pre-trained ResNet50 model transforms the CT-images slices into feature vectors of length 1000. The auxiliary hypothesis is that the ResNet50 model is equipped with 2D filter stacks that extract meaningful information for general image processing tasks. Since some of the tasks that the model was previously trained for might be unrelated to the prediction problem at hand, we test a possible reduction of the included features. For consistency with the radiomics baseline, the previously described feature selection method is used to reduce the size of the prediction input to 32. Unlike the features selected from radiomics, deep ResNet50 features are extracted by the pre-trained model and therefore do not carry interpretable labels.

[2] https://pyradiomics.readthedocs.io/en/latest/

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/fonc.2022.849447

Deep Learning Model for Classifying Metastatic Epidural Spinal Cord Compression on MRI

James Thomas Patrick Decourcy Hallinan[1,2*†], *Lei Zhu*[3†], *Wenqiao Zhang*[4], *Desmond Shi Wei Lim*[1,2], *Sangeetha Baskar*[1], *Xi Zhen Low*[1,2], *Kuan Yuen Yeong*[5], *Ee Chin Teo*[1], *Nesaretnam Barr Kumarakulasinghe*[6], *Qai Ven Yap*[7], *Yiong Huak Chan*[7], *Shuxun Lin*[8], *Jiong Hao Tan*[9], *Naresh Kumar*[9], *Balamurugan A. Vellayappan*[10], *Beng Chin Ooi*[4], *Swee Tian Quek*[1,2] and *Andrew Makmur*[1,2]

[1] Department of Diagnostic Imaging, National University Hospital, Singapore, Singapore, [2] Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, [3] NUS Graduate School, Integrative Sciences and Engineering Programme, National University of Singapore, Singapore, Singapore, [4] Department of Computer Science, School of Computing, National University of Singapore, Singapore, Singapore, [5] Department of Radiology, Ng Teng Fong General Hospital, Singapore, Singapore, [6] National University Cancer Institute, NUH Medical Centre (NUHMC), Singapore, Singapore, [7] Biostatistics Unit, Yong Loo Lin School of Medicine, Singapore, Singapore, [8] Division of Spine Surgery, Department of Orthopaedic Surgery, Ng Teng Fong General Hospital, Singapore, Singapore, [9] University Spine Centre, Department of Orthopaedic Surgery, National University Health System, Singapore, Singapore, [10] Department of Radiation Oncology, National University Cancer Institute Singapore, National University Hospital, Singapore, Singapore

Background: Metastatic epidural spinal cord compression (MESCC) is a devastating complication of advanced cancer. A deep learning (DL) model for automated MESCC classification on MRI could aid earlier diagnosis and referral.

Purpose: To develop a DL model for automated classification of MESCC on MRI.

Materials and Methods: Patients with known MESCC diagnosed on MRI between September 2007 and September 2017 were eligible. MRI studies with instrumentation, suboptimal image quality, and non-thoracic regions were excluded. Axial T2-weighted images were utilized. The internal dataset split was 82% and 18% for training/validation and test sets, respectively. External testing was also performed. Internal training/validation data were labeled using the Bilsky MESCC classification by a musculoskeletal radiologist (10-year experience) and a neuroradiologist (5-year experience). These labels were used to train a DL model utilizing a prototypical convolutional neural network. Internal and external test sets were labeled by the musculoskeletal radiologist as the reference standard. For assessment of DL model performance and interobserver variability, test sets were labeled independently by the neuroradiologist (5-year experience), a spine surgeon (5-year experience), and a radiation oncologist (11-year experience). Inter-rater agreement (Gwet's kappa) and sensitivity/specificity were calculated.

Results: Overall, 215 MRI spine studies were analyzed [164 patients, mean age = 62 ± 12 (SD)] with 177 (82%) for training/validation and 38 (18%) for internal testing. For internal testing, the DL model and specialists all showed almost perfect agreement (kappas = 0.92–0.98, $p < 0.001$) for dichotomous Bilsky classification (low versus high grade) compared to the reference standard. Similar performance was seen for external testing on

Edited by:
*Abhishek Mahajan,
Clatterbridge Cancer Centre NHS
Foundation Trust, United Kingdom*

Reviewed by:
*Jefferson W. Chen,
University of California, Irvine,
United States
Weiwei Zong,
Henry Ford Health System,
United States
Chirag Kamal Ahuja,
Post Graduate Institute of Medical
Education and Research (PGIMER),
India
Abhishek Chatterjee,
Tata Memorial Hospital, India*

***Correspondence:**
*James Thomas Patrick Decourcy Hallinan
james_hallinan@nuhs.edu.sg*

†These authors have contributed equally to this work and share first authorship

Specialty section:
*This article was submitted to
Cancer Imaging and
Image-directed Interventions,
a section of the journal
Frontiers in Oncology*

Received: 06 January 2022
Accepted: 18 March 2022
Published: 04 May 2022

Citation:
Hallinan JTPD, Zhu L, Zhang W, Lim DSW, Baskar S, Low XZ, Yeong KY, Teo EC, Kumarakulasinghe NB, Yap QV, Chan YH, Lin S, Tan JH, Kumar N, Vellayappan BA, Ooi BC, Quek ST and Makmur A (2022) Deep Learning Model for Classifying Metastatic Epidural Spinal Cord Compression on MRI. Front. Oncol. 12:849447. doi: 10.3389/fonc.2022.849447

a set of 32 MRI spines with the DL model and specialists all showing almost perfect agreement (kappas = 0.94–0.95, $p < 0.001$) compared to the reference standard.

Conclusion: A DL model showed comparable agreement to a subspecialist radiologist and clinical specialists for the classification of malignant epidural spinal cord compression and could optimize earlier diagnosis and surgical referral.

Keywords: deep learning model, metastatic epidural spinal cord compression, MRI, Bilsky classification, spinal metastasis classification, spinal metastatic disease, epidural spinal cord compression

INTRODUCTION

Spinal metastases are common and seen in up to 40% of cancer patients. Up to 20% of these patients develop complications including spinal cord compression, which can lead to permanent neurological dysfunction if treatment is delayed. With the development of more effective systemic therapy (such as targeted and immunotherapy), the survival of patients with metastatic cancer has increased, and consequently, the incidence of spinal metastases is expected to rise (1–3).

Suspicion for spinal metastases begins in the clinic, as greater than 85% of patients present with back pain. Imaging is then required to confirm the presence of spinal metastases and the associated complications. MRI is the most accurate modality due to improved soft-tissue resolution, which allows assessment of the extent of metastatic bony involvement, compression fractures, and the presence of metastatic epidural spinal cord compression (MESCC) (4).

The degree of MESCC is assessed on axial T2-weighted (T2W) MR images using a six-point grading scale developed by the Spine Oncology Study Group (SOSG), commonly referred to as the Bilsky grading scale (5). Low-grade disease (Bilsky 0, 1a, and 1b) can be considered for initial radiotherapy (including stereotactic body radiotherapy (SBRT)/stereotactic radiosurgery), whereas higher-grade disease (Bilsky 1c, 2, and 3) should be considered for surgical decompression followed by radiotherapy (6). MESCC requires urgent treatment to prevent permanent neurological injury, but significant delays in management have been reported. A study by van Tol et al. (2021) showed median delays of 21.5, 7, and 8 days for the diagnosis, referral, and treatment of MESCC, respectively (7).

A deep learning (DL) model to automatically detect and classify low- versus high-grade Bilsky MESCC on MRI could alert the radiologist and clinical teams, ensuring prompt reporting and appropriate referral. This is important to prevent poor functional outcomes and increased requirements of healthcare resources (8). Automated tools for detecting urgent findings on MRI are important due to increasing demand for the modality, while faced with a shortage of radiologists (9). In the United Kingdom, 3.4 million MRI studies are reported every year, and patients can wait over 30 days for a report (10, 11). Even for emergent indications including suspected MESCC where reporting should be performed within hours, more than a third of reports were provided greater than 48 h later at one healthcare trust (10, 12).

Prior DL in spine MRI has shown promise, especially with the use of convolutional neural networks (CNNs), which can automatically learn representative features from images to perform classification tasks. Most recently, several teams have developed DL models for the automated classification of degenerative narrowing in the lumbar spine (13, 14) or adjacent segment disease along the cervical spine (15). DL for spinal metastases on advanced imaging, including MRI, is still in the preliminary phase. A study by Wang et al. (2017) showed the feasibility of automated spinal metastatic disease detection on MRI using a small set of 26 patients (16). The group achieved a true positive rate of 90% with a false-positive rate of up to 0.4 per case. DL for the detection of spinal metastases on CT has also shown promise for quantifying metastatic bone disease burden (17). Currently, to our knowledge, no DL model has been developed to assess MESCC on MRI.

The aim of this study was to train a DL model for the automated Bilsky classification of MESCC using axial T2W MRI. This could aid earlier diagnosis of MESCC and identify suitable candidates for radiotherapy versus emergent surgical decompression. Once trained, the performance of the DL model was compared with that of a radiation oncologist, spine oncology surgeon, and subspecialty radiologist, on an internal test set. The DL model performance and generalizability were also assessed on an external test set.

MATERIALS AND METHODS

This study was approved by our institutional review board and compliant with the Health Insurance Portability and Accountability Act (HIPAA). A waiver of consent was granted due to the retrospective nature of the study and the minimal risk involved.

Dataset Preparation

Retrospective, manual extraction, and anonymization of MRI spines from patients with known vertebral metastatic disease and thoracic MESCC were done over a 10-year period from September 2007 to September 2017 at the National University Hospital, Singapore. Adult patients (≥18 years) were included with a selection of studies across different MRI scanners (GE and Siemens 1.5- and 3.0-T platforms). A heterogeneous training dataset obtained using a range of MRI platforms and T2W parameters was used to prevent overfitting and provide a more generalizable DL algorithm. MRI spines with instrumentation, suboptimal image quality (e.g., motion and cerebrospinal fluid flow artifacts), and non-thoracic spine regions were excluded. Axial T2W DICOM images were utilized. **Supplementary Table 1** provides details on the MRI scanners and T2W sequence parameters.

The dataset at the National University Hospital, Singapore, was assigned as the internal dataset and was randomly split into 82% and 18% for the training/validation and test sets, respectively. This is an acceptable split for DL datasets (18).

A dataset of MRI spine studies from patients with known metastatic disease and MESCC was also obtained for external testing from Ng Teng Fong General Hospital (Siemens 1.5-T MRI platform). The inclusion and exclusion criteria were identical to the internal dataset. The MRI spines were obtained over a 5-year period from September 2015 to September 2020, encompassing anonymized axial T2W DICOM images. No further training was performed on this dataset.

Dataset Labelling

Internal training data were manually labeled by two board-certified radiologists with sub-specialization in musculoskeletal radiology (JH; 10-year experience) and neuroradiology (AM; 5-year experience). Each radiologist labeled at least 100 MRI thoracic spine studies independently. With the use of an open-source annotation software (LabelImg, https://github.com/tzutalin/labelImg), bounding boxes were drawn to segment the region of interest (ROI) around the spinal canal along the thoracic spine (C7–T1 through to the conus at T12–L3). A bounding box was placed on each axial T2W image.

When drawing each bounding box, the annotating radiologist classified the MESCC using the Bilsky classification (4). This grading scheme consists of six classifications with grades 0, 1a, and 1b amenable to radiotherapy and grades 1c, 2, and 3 more likely to require surgical decompression. A visual scale was provided to all annotating readers (**Figure 1**). Degenerative changes (disk bulges and ligamentum flavum redundancy) leading to moderate-to-severe spinal canal stenosis were labeled by the annotating radiologists and excluded from further analysis (19, 20).

The internal and external test sets were labeled using the same visual scale by the musculoskeletal radiologist (JH) with 10-year experience and served as the reference standard. For comparison with the DL model and to assess interobserver variability, the internal and external test sets were also labeled independently by a subspecialist neuroradiologist (AM; 5-year experience), a spine oncology surgeon (JT; 5-year experience), and a radiation oncologist (BV; 11-year experience). The specialist readers were blinded to the reference standard.

Deep Learning Model Development

A convolutional prototypical network is a newly proposed neural network architecture for robust image classification with cluster assumption (21). Specifically, it is assumed that there exists an embedding space in which data points cluster around a single prototype representation for each class. Different types of loss functions are proposed for the training of the network with the general stochastic gradient descent method (22). Several studies have demonstrated the robustness of convolutional prototypical networks towards data scarcity and class imbalance problems, which have also led to more compact and discriminative features in the embedding space (21, 22).

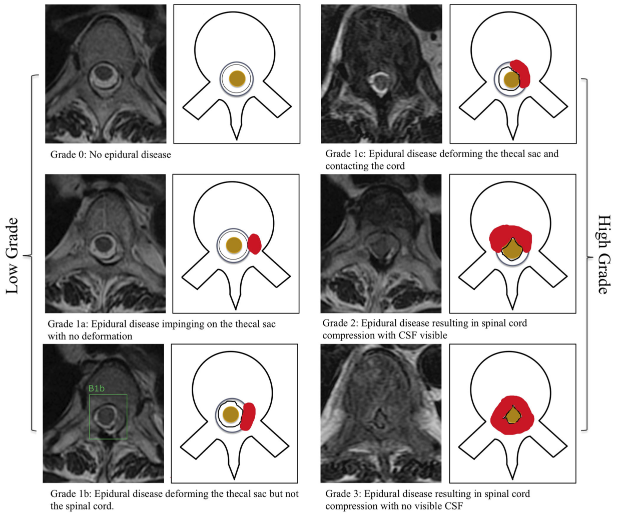

Low Grade
High Grade
Grade 0: No epidural disease
Grade 1c: Epidural disease deforming the thecal sac and contacting the cord
Grade 1a: Epidural disease impinging on the thecal sac with no deformation
Grade 2: Epidural disease resulting in spinal cord compression with CSF visible
B1b
Grade 1b: Epidural disease deforming the thecal sac but not the spinal cord.
Grade 3: Epidural disease resulting in spinal cord compression with no visible CSF

FIGURE 1 | Bilsky classification of metastatic epidural spinal cord compression on MRI of the thoracic spine. Axial T2-weighted (repetition time ms/echo time ms, 5,300/100) images were used. Training of the deep learning model was performed by a radiologist by placing a bounding box around the region of interest at each T2-weighted image. A bounding box example is included for a low-grade Bilsky 1b lesion (1b). CSF, cerebrospinal fluid.

In this paper, a convolutional prototypical network was trained with ResNet50 as its backbone to project ROI images into a high-dimensional embedding space (23). The Apache SINGA (24) platform was adopted for efficient training of the deep network, and MLCask (25), an efficient data analytics pipeline management system, was adopted to facilitate managing different versions of the developed pipelines. We used the output from the global average layer of ResNet50 as the feature representation for each image in the embedding space. A class prototype was assigned for each Bilsky score in the embedding space. The prediction probability of a data point was calculated for each class *via* a SoftMax over the negative distance to the class prototypes. The network was trained with a cross-entropy loss on the prediction probability using a standard SGD optimizer, and a compact regularization was introduced to further minimize the distance between the data points and their corresponding class prototypes. Simultaneously, the virtual adversarial loss was introduced to ensure our model makes consistent predictions around the neighborhood of each data point with adversarial local perturbation (26). An ablation study was also conducted to demonstrate the effectiveness of the virtual adversarial loss and the compact regularization loss. The ablation study details are included as the Supplementary Material, and **Supplementary Table 2** shows the ablation study results. The Supplementary Material including **Supplementary Table 3** has also been provided to compare our developed model with both the standard ResNet50 and the plain convolutional prototypical network.

For inference, the first step is to extract the ROI of an input image. The extracted ROI is then projected into the embedding space. Finally, the label of the input image is predicted as the label of its nearest class prototype in the embedding space (**Supplementary Figure 1**). The DL model (SpineAI@NUHS-NUS) code is at https://github.com/NUHS-NUS-SpineAI/SpineAI-Bilsky-Grading. **Supplementary Figure 2** shows a flow chart of the developed DL model in a clinical setting.

Statistical Analysis

All analyses were performed using Stata version 16 (StataCorp, College Station, TX, USA) with statistical significance set at 2-sided p < 0.05. Postulating that a kappa of 0.9 is to be anticipated, at least 138 samples (MRI studies) were required to provide a 95% CI width of 0.1. Over the 10-year study period, 174 subjects with 239 MRI studies were collected, which was sufficient for the analysis. Descriptive statistics for continuous variables were presented as mean ± SD (range) and n (%) for categorical variables. Inter-rater agreement using dichotomous (low-grade versus high-grade) Bilsky classification was assessed using Gwet's kappa to account for the paradox effect of a high percentage of normal classification (27). Sensitivity and specificity were also presented for dichotomous Bilsky gradings only. Sensitivity is the percentage of high-grade Bilsky classifications that are correctly identified by the DL model and specialist readers, whereas specificity is the percentage of low-grade Bilsky classifications that are correctly identified by the DL model and specialist readers.

Levels of agreement were defined for Gwet's kappa: <0 = poor, 0–0.2 = slight, 0.21–0.4 = fair, 0.41–0.6 = moderate, 0.61–0.8 = substantial, and 0.81–1 = almost-perfect agreement (28). Also, 95% CIs were calculated.

RESULTS

Patient Characteristics in Datasets

Data collection over the 10-year study period identified 174 patients with 239 MRI spines for analysis. Of these, 24 MRI spines from 10 patients were excluded due to instrumentation (4 MRI spines), suboptimal image quality (2 MRI spines), or non-thoracic spine MRI (18 MRI spines). A total of 164 patients encompassing 215 MRI thoracic spines were evaluated. Overall, the mean age of all 164 patients was 62 ± 12 (SD) (range: 18–93 years). The patient group was predominantly male (91/164 patients, 55.4%), with breast and lung being the most common primary cancers (63/164 patients, 38.4%). There was a wide range of sites of MESCC along the thoracic region, with a predominance of disease in the semirigid thoracic region between T3 and T10 (73/164 patients, 44.5%). The patient demographics, cancer subtypes, and MESCC distribution along the thoracic region for the training and test sets are displayed in **Table 1**.

The internal dataset of 215 MRI spines was randomly split into 177 (82%) studies for training/validation and 38 (18%) studies for internal testing. A flow chart of the internal dataset study design is provided in **Figure 2**.

For the external dataset, 32 patients with 32 MRI spines covering the thoracic region were available for external testing. Overall, the mean age of the 32 patients was 60 ± 13 (SD) (range: 19–85 years). Similar to the internal dataset, the patient group had a predominance of men (20/32 patients, 62.5%), with the lung being the most common primary cancer (13/32 patients, 40.6%).

Reference Standard

The number of ROIs and the corresponding Bilsky classifications in the internal training and test sets, and external test sets are highlighted in **Table 2**. In the internal training/validation set, high-grade Bilsky classification (1c/2/3) accounted for 462/5,863 ROIs (7.9%) with a predominance of low-grade Bilsky classification (0/1a/1b) at the remaining 5,401/5,863 ROIs (92.1%). In the internal test set, high-grade Bilsky classification (1c/2/3) accounted for 84/1,066 ROIs (7.9%) with a predominance of low-grade Bilsky classification (0/1a/1b) at the remaining 982/1,066 ROIs (92.1%). In comparison, for the external test set, there was a greater proportion of high-grade Bilsky classification (169/754 ROIs, 22.4%) and a reduced predominance of low-grade Bilsky classification (585/754 ROIs, 77.6%). The greater proportion of higher-grade Bilsky classification in the external test set was likely due to more targeted axial T2W images at the sites of MESCC.

Internal Test Set Region of Interest Classification

For the internal dataset, there was almost perfect agreement between the reference standard for dichotomous Bilsky classification and the DL model and all specialist readers, with kappas ranging from 0.92 to 0.98, all p < 0.001 (**Table 3**). A kappa of 0.98 (95% CI = 0.97–0.99, p < 0.001) for the spine surgeon was the highest, with similar kappas of 0.97 (95% CI = 0.96–0.98, p < 0.001) and 0.96 (95% CI = 0.95–0.98, p < 0.001) for the radiation oncologist and neuroradiologist, respectively. DL model kappa of 0.92 (95% CI = 0.91–0.94, p < 0.001) was slightly lower compared to that of the specialist readers.

The sensitivity for the DL model (97.6%, 95% CI = 91.7%–99.7%) was the highest for the internal dataset, and this was significantly higher compared to both the neuroradiologist (84.5%, 95% CI = 75.0%–91.5%) and spine surgeon (79.8%, 95% CI = 69.6%–87.7%), p = 0.003 and p < 0.001, respectively (**Table 4** and confusion matrix in **Supplementary Table 4**). High specificities (range = 93.6%–99.5%) were seen for the DL model and specialists. The spine surgeon had a specificity of 99.5% (95% CI = 98.8%–99.8%), which was significantly higher than the DL model, neuroradiologist, and radiation oncologist, with specificities of 93.6% (95% CI = 91.9%–95.0%), 98.1% (95% CI 97.0%–98.8%), and 97.9% (95% CI = 96.7%–98.7%), p < 0.001, p = 0.004, and p = 0.002, respectively.

External Test Set Region of Interest Classification

For the external dataset, the DL model and all the specialist readers also had almost perfect agreement (kappas 0.94–0.95, all p < 0.001) compared to the reference standard for dichotomous Bilsky classification (**Table 3**). The neuroradiologist kappa of 0.95 (95% CI = 0.93–0.97, p < 0.001) was only slightly higher compared to the rest, with similar kappas of 0.94 (95% CI = 0.92–

TABLE 1 | Patient demographics and clinical characteristics for the internal and external test sets.

Characteristics	Internal training set (n = 129)	Internal test set (n = 35)	External test set (n = 32)
Age (years)*	61 ± 13 (18–93)	61 ± 12 (39–87)	60 ± 13 (19–85)
Women	55 (42.6)	18 (51.4)	12 (37.5)
Men	74 (57.4)	17 (48.6)	20 (62.5)
Ethnicity			
Chinese	93 (72.1)	28 (80)	23 (71.9)
Malay	21 (16.3)	3 (8.6)	7 (21.9)
Indian	7 (5.4)	2 (5.7)	0 (0)
Others	8 (6.2)	2 (5.7)	2 (6.2)
Cancer subtype			
Breast	23 (17.8)	8 (22.9)	3 (9.4)
Lung	21 (16.3)	11 (31.4)	13 (40.6)
Prostate	19 (14.7)	5 (14.3)	4 (12.5)
Colon	15 (11.6)	3 (8.6)	3 (9.4)
Renal cell carcinoma	10 (7.8)	2 (5.7)	1 (3.1)
Nasopharyngeal carcinoma	9 (7)	3 (8.6)	1 (3.1)
Others	32 (24.8)	3 (8.6)	7 (21.9)
No. of MRI thoracic spines	177/215 (82.3)	38/215 (17.6)	32
MESCC location			
Diffuse thoracic#	30 (23.3)	8 (22.9)	3 (9.4)
C7–T2	13 (10.1)	3 (8.6)	6 (18.8)
T3–T10	55 (42.6)	18 (51.4)	15 (46.9)
T11–L3	31 (24.0)	6 (17.1)	8 (25)

MESCC, malignant epidural spinal cord compression.

**Values are mean ± SD (range) for numerical variables and n (%) for categorical variables.*

#Two or more sites of thoracic epidural disease.

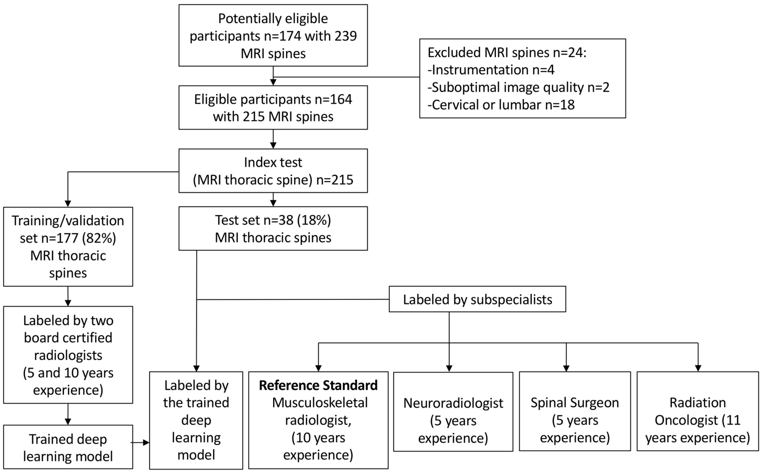

Potentially eligible participants n=174 with 239 MRI spines
Excluded MRI spines n=24:
-Instrumentation n=4
-Suboptimal image quality n=2
-Cervical or lumbar n=18
Eligible participants n=164 with 215 MRI spines
Index test (MRI thoracic spine) n=215
Training/validation set n=177 (82%) MRI thoracic spines
Test set n=38 (18%) MRI thoracic spines
Labeled by subspecialists
Labeled by two board certified radiologists (5 and 10 years experience)
Trained deep learning model
Labeled by the trained deep learning model
Reference Standard Musculoskeletal radiologist, (10 years experience)
Neuroradiologist (5 years experience)
Spinal Surgeon (5 years experience)
Radiation Oncologist (11 years experience)

FIGURE 2 | Flow chart of the study design for the internal training/validation and test sets. The deep learning model performance was compared with an expert musculoskeletal radiologist (reference standard) and three specialist readers.

0.96, p < 0.001), 0.94 (95% CI = 0.92–0.96, p < 0.001), and 0.94 (95% CI = 0.91–0.96, p < 0.001) for the DL model, radiation oncologist, and spine surgeon, respectively.

The sensitivity for the DL model on the external dataset was 89.9% (95% CI = 84.4%–94.0%), and this was not significantly different from the other readers, including the neuroradiologist with the highest sensitivity of 92.9% (95% CI = 87.9%–96.2%), all p > 0.05 (**Table 4** and confusion matrix in **Supplementary Table 5**). The neuroradiologist had no significantly higher sensitivity compared to the other readers, all p > 0.05. The spine surgeon had a specificity of 99.3% (95% CI = 98.3%–99.8%), which was significantly higher than the specificity of the neuroradiologist at 97.9% (95% CI 96.4%–98.9%), p = 0.042.

DISCUSSION

MRI is an essential tool in the assessment of MESCC, which is a potentially devastating complication of advanced cancer. Bilsky

TABLE 2 | Reference standards for the internal (training and test) and external (test) sets showing the number of Bilsky MESCC grades.

Bilsky MESCC grade	Internal training/validation set	Internal test set	External test set
0	4,508 (76.9)	849 (79.6)	454 (60.2)
1a	424 (7.2)	82 (7.7)	48 (6.4)
1b	469 (8.0)	51 (4.8)	83 (11)
1c	216 (3.7)	35 (3.3)	51 (6.7)
2	105 (1.8)	26 (2.4)	39 (5.2)
3	141 (2.4)	23 (2.2)	79 (10.5)
Total	5,863	1,066	754

Values are n (%). A region of interest (bounding box) for Bilsky grade was drawn at each axial T2-weighted image.
MESCC, malignant epidural spinal cord compression.

TABLE 3 | Internal and external test set classifications using dichotomous Bilsky gradings (low versus high grade) on MRI.

Reader	Internal test set		External test set	
	Kappa (95% CI)	p-Value	Kappa (95% CI)	p-Value
DL model	0.92 (0.91–0.94)	<0.001	0.94 (0.92–0.96)	<0.001
Neuroradiologist	0.96 (0.95–0.98)	<0.001	0.95 (0.93–0.97)	<0.001
Radiation oncologist	0.97 (0.96–0.98)	<0.001	0.94 (0.92–0.96)	<0.001
Spine surgeon	0.98 (0.97–0.99)	<0.001	0.94 (0.91–0.96)	<0.001

Gwet's kappa was used.
DL, deep learning model.

TABLE 4 | Internal and external test set sensitivity and specificity for the deep learning model and specialist readers using dichotomous Bilsky gradings (low versus high grade) on MRI.

Reader	Internal test set		External test set	
	Sens (95% CI)	Spec (95% CI)	Sens (95% CI)	Spec (95% CI)
DL model	97.6 (91.7–99.7)	93.6 (91.9–95.0)	89.9 (84.4–94.0)	98.1 (96.7–99.1)
Neuroradiologist	84.5 (75.0–91.5)	98.1 (97.0–98.8)	92.9 (87.9–96.2)	97.9 (96.4–98.9)
Radiation oncologist	94.0 (86.7–98.0)	97.9 (96.7–98.7)	88.8 (83.0–93.1)	98.5 (97.1–99.3)
Spine surgeon	79.8 (69.6–87.7)	99.5 (98.8–99.8)	83.4 (77.0–88.7)	99.3 (98.3–99.8)

DL, deep learning model; Sens, sensitivity; Spec, specificity.

et al. (2010) developed an MRI classification for MESCC that aimed to improve communication between specialists and aid decision making for initial radiotherapy versus expedited surgical decompression. In our study, we trained a DL model for automated Bilsky MESCC classification on thoracic spine MRI using manual radiologist labels. On an internal test set, the DL model showed almost-perfect agreement ($\kappa = 0.92$, $p < 0.001$) for dichotomous Bilsky classification (low grade versus high grade), similar to specialist readers (κ = 0.96–0.98, all $p < 0.001$), which included a radiation oncologist, a neuroradiologist, and a spine surgeon. In a further step, external testing of the DL model was performed on a dataset from a different institution to assess generalizability. For the external dataset, the DL model and all the specialist readers also had almost perfect agreement (kappas 0.94–0.95, all $p < 0.001$) for dichotomous Bilsky classification.

DL is already being used in spine diseases to aid in the diagnosis of spinal stenosis on MRI spines, surgical planning, and prediction of outcomes in patients with spinal metastases (8, 29). DL in spinal oncology imaging is limited with most researchers focusing on the detection of metastases (30), or automated spinal cord segmentation as an organ at risk for radiotherapy planning (31). Average Dice similarity coefficients for spinal cord segmentation are as high as 0.9 for automated lung cancer radiotherapy planning using DL on CT studies (32, 33). Automated detection of spinal cord compression on MRI has currently only been assessed in the cervical spine. Merali et al. (2021) developed a DL model for degenerative cervical spinal cord compression on MRI using 201 patients from a surgical database (34). Their DL model had an overall AUC of 0.94 with a sensitivity of 0.88 and specificity of 0.89.

To our knowledge, no team has currently looked at the automated prediction of metastatic epidural spinal cord compression on MRI, which is a medical emergency. The current National Institute for Health and Care Excellence (NICE) guidelines state that metastatic epidural spinal cord compression should be treated as soon as possible, ideally within 24 h, to prevent irreversible neurological dysfunction (35). Our MRI Bilsky grading prediction model could improve the imaging and clinical workflow of patients with spinal metastases. MRI studies with MRI studies with high-grade Bilsky disease could be triaged for urgent radiologist review, with the radiology reporting augmented by an automated selection of key images at the sites of the highest-grade Bilsky lesions and spinal cord compression. These key images could also be circulated to an on-demand spine oncology multidisciplinary team (spine surgeons, oncologists, and radiation oncologists) for more streamlined decision making and appropriate referral. It should be emphasized that the treatment of MESCC is not just dependent on imaging but is also heavily weighted on clinical presentation, e.g., myelopathy, weakness, and loss of bowel and bladder function. Individuals can present with high-grade Bilsky scores and not be suitable surgical candidates. Further work using our Bilsky prediction model could involve combining imaging data with clinical information (e.g., age, cancer subtype, and degree of neurological impairment) to improve the selection of patients for more aggressive management including surgery and/or SBRT (21, 36). Our DL model is focused on Bilsky classification and currently does not have the ability to segment or outline tumors. DL auto-segmentation of tumors in MR images could optimize and reduce the time taken for radiotherapy planning (32). Future research will focus on developing a DL model for this application, which will be especially useful for SBRT.

Our study has limitations. First, we utilized axial T2W images along the thoracic region, which was recommended as the most accurate method for MESCC classification on MRI in the study by Bilsky et al. (2010) (4). In further studies, we could enhance the model performance for the detection and classification of MESCC by combining multiple MRI sequences, including sagittal T2W and gadolinium-enhanced T1-weighted axial and sagittal image sets. Second, we chose to use dichotomous Bilsky classification (low grade vs. high grade) with the inclusion of Bilsky 1c under high-grade disease. This is controversial, as patients with Bilsky 1c are unlikely to have neurological deficits requiring urgent surgical treatment. However, for the purpose of treatment triaging, we decided to be more conservative and classify 1c under high grade. Third, the reference standard was a single expert musculoskeletal radiologist who reviewed the test set independently from the other three specialist readers. No consensus labeling was performed for the readers, as this may have been biased toward the expert. Fourth, the test sets were only assessed by specialist readers to ensure the most rigorous comparison with the DL model. Assessment by less experienced readers (e.g., radiology or surgical trainees) was not analyzed but could be performed through further studies that include the use of semi-supervised reporting augmentation by the DL model. Finally, labeling of images for model development was a labor-intensive manual process (highly supervised). This was believed to be the most accurate method for training the model but potentially limited the number of MRI studies that could be used for training. Alternatively, future larger datasets could utilize semi-supervised learning, which can leverage unlabeled data to boost the DL model performance and reduce the data annotation burden (37–39). Future work could also utilize additional external datasets to ensure the DL model is not overfitted to our institution data and is generalizable to new, unseen data.

In conclusion, we demonstrated that our DL model is reliable and may be used to automatically assess the Bilsky classification of metastatic epidural spinal cord compression on thoracic spine MRI. In clinical practice, the early diagnosis of spinal cord compression is important to prevent permanent neurological dysfunction (40). The DL model could be used to triage MRI scans for urgent reporting, augment non-sub-specialized radiologists when they report out of hours, and improve the communication and referral pathways between specialties including oncology, radiation oncology, and surgery. Finally, the proposed framework, which makes use of Apache SINGA (24) for distributed training, has been integrated into our MLCask (25) system for handling healthcare images and analytics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conception, methodology, data curation, supervision, visualization, and writing: JH, LZ, WZ, DL, SB, XL, ET, NBK, QY, YC, JT, NK, BV, BO, SQ, and AM. Investigation and project administration: JH, LZ, WZ, DL, SB, XL, ET, NBK, QY, YC, SL, JT, NK, BV, BO, and AM. Resources and software: JH, LZ, WZ, DL, KY, QY, YC, SL, NK, BV, BO, SQ, and AM. Formal analysis and validation: JH, LZ, WZ, DL, KY, NBK, QY, YC, SL, JT, NK, BV, BO, and SQ.

FUNDING

1) Direct funding from MOH/NMRC: this research is supported by the Singapore Ministry of Health's National Medical Research Council under its NMRC Clinician-scientist individual research grant, new investigator grant (CS-IRG NIG); Grant Title: Deep learning pathway for the management of spine metastases (CNIG20nov-0011, MOH-000725). 2) NCIS Centre Grant Seed Funding Program (December 2020 Grant Call); Grant Title: Artificial intelligence for the management of vertebral metastases.

ACKNOWLEDGMENTS

We acknowledge Mr. Yang Kaiyuan's contribution to the development of the deep learning model.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.849447/full#supplementary-material

Supplementary Figure 1 | Deep learning model development pipeline. Given the *x* as the input data of medical images, our goal is to classify these images into the corresponding Bilsky class. We first extract the region of interest (ROI) of x and feed them to the feature extractor and perturbation generator. The produced r and r_{adv} are the representation and virtual adversarial perturbation of data, respectively. We assign prototypes for each Bilsky class in the embedding space and calculate prediction probability for both the original and perturbated data points *via* a SoftMax over the negative of distance to the class prototypes. Correspondingly, $\bar{y}_1$ and $\hat{y}_1$ are the original prediction and perturbated predictions. Finally, the deep learning network is trained by minimizing the virtual adversarial loss on consistency regularization and the cross-entropy loss on the prediction probability. Note, in the embedding space, the orange-colored points are prototypes for each Bilsky class, data points of other colors represent images with different Bilsky classes. The grey-colored points are original data before perturbation.

Supplementary Figure 2 | Flow chart of deep learning model deployment for clinical usage. We embed the developed deep learning model in the above pipeline for deployment. Input MRI images from patient studies will go through ROI detection with the clinicians, then the developed model is used to make predictions for the studies and report the prediction results back to the clinicians

Supplementary Table 1 | MRI Platform and parameters for MRI spine axial T2-weighted Imaging. TE, echo time; TR, repetition time; GE, General Electric, *MRI scanner at the external center (Ng Teng Fong General Hospital, Singapore). All four other scanners were situated at the National University Hospital, Singapore. All studies were performed in the supine position with a torso coil.

Supplementary Table 2 | Ablation study on the developed model.

Supplementary Table 3 | Comparison study on the developed model.

Supplementary Table 4 | Confusion matrix of the deep learning model on the internal test set.

Supplementary Table 5 | Confusion matrix of the deep learning model on the external test set.

REFERENCES

1. Chiu RG, Mehta AI. Spinal Metastases. *JAMA* (2020) 323:2438. doi: 10.1001/jama.2020.0716
2. Barzilai O, Fisher CG, Bilsky MH. State of the Art Treatment of Spinal Metastatic Disease. *Neurosurgery* (2018) 82:757–69. doi: 10.1093/neuros/nyx567
3. Laur O, Nandu H, Titelbaum DS, Nunez DB, Khurana B. Nontraumatic Spinal Cord Compression: MRI Primer for Emergency Department Radiologists. *Radiographics* (2019) 39:1862–80. doi: 10.1148/rg.2019190024
4. Nair C, Panikkar S, Ray A. How Not to Miss Metastatic Spinal Cord Compression. *Br J Gen Pract* (2014) 64:e596–8. doi: 10.3399/bjgp14X681589
5. Bilsky MH, Laufer I, Fourney DR, Groff M, Schmidt MH, Varga PP, et al. Reliability Analysis of the Epidural Spinal Cord Compression Scale. *J Neurosurg Spine* (2010) 13:324–8. doi: 10.3171/2010.3.SPINE09459
6. Laufer I, Rubin DG, Lis E, Cox BW, Stubblefield MD, Yamada Y, et al. The NOMS Framework: Approach to the Treatment of Spinal Metastatic Tumors. *Oncologist* (2013) 18:744–51. doi: 10.1634/theoncologist.2012-0293
7. van Tol FR, Versteeg AL, Verkooijen HM, Öner FC, Verlaan JJ. Time to Surgical Treatment for Metastatic Spinal Disease: Identification of Delay Intervals. *Global Spine J* (2021) 18:2192568221994787. doi: 10.1177/2192568221994787
8. van Tol FR, Choi D, Verkooijen HM, Oner FC, Verlaan JJ. Delayed Presentation to a Spine Surgeon Is the Strongest Predictor of Poor Postoperative Outcome in Patients Surgically Treated for Symptomatic Spinal Metastases. *Spine J* (2019) 19:1540–7. doi: 10.1016/j.spinee.2019.04.011
9. Gourd E. UK Radiologist Staffing Crisis Reaches Critical Levels. *Lancet Oncol* (2017) 18:e651. doi: 10.1016/S1470-2045(17)30806–9
10. Care Quality Commission. *Radiology Review. A National Review of Radiology Reporting Within the NHS in England* (2018). Available at: https://www.cqc.org.uk/RadiologyReview.
11. The Royal College of Radiologists. *Unreported X-Rays, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) Scans: Results of a Snapshot Survey of English National Health Service (NHS) Trusts* (2015). Available at: https://www.rcr.ac.uk/sites/default/files/publication/Unreported_stu (Accessed 13 Dec 2021).
12. Griffin S. Covid-19: Failings in Imaging Services Have Put Cancer Patients at Risk, Watchdog Says. *BMJ* (2021) 374:n1749. doi: 10.1136/bmj.n1749
13. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI. *Radiology* (2021) 300:130–8. doi: 10.1148/radiol.2021204289
14. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, Battié MC, et al. Genodisc Consortium. ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of Reading of Radiological Features From Magnetic Resonance Images (MRIs) of the Lumbar Spine Without Human Intervention is Comparable With an Expert Radiologist. *Eur Spine J* (2017) 26:1374–83. doi: 10.1007/s00586-017-4956-3
15. Goedmakers CMW, Lak AM, Duey AH, Senko AW, Arnaout O, Groff MW, et al. Deep Learning for Adjacent Segment Disease at Preoperative MRI for Cervical Radiculopathy. *Radiology* (2021) 301:664–71. doi: 10.1148/radiol.2021204731
16. Wang J, Fang Z, Lang N, Yuan H, Su MY, Baldi P. A Multi-Resolution Approach for Spinal Metastasis Detection Using Deep Siamese Neural Networks. *Comput Biol Med* (2017) 84:137–46. doi: 10.1016/j.compbiomed.2017.03.024
17. Lindgren Belal S, Sadik M, Kaboteh R, Enqvist O, Ulén J, Poulsen MH, et al. Deep Learning for Segmentation of 49 Selected Bones in CT Scans: First Step in Automated PET/CT-Based 3D Quantification of Skeletal Metastases. *Eur J Radiol* (2019) 113:89–95. doi: 10.1016/j.ejrad.2019.01.028
18. England JR, Cheng PM. Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. *AJR Am J Roentgenol* (2019) 212:513–9. doi: 10.2214/AJR.18.20490
19. Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino JA, Kaiser J, et al. Reliability of Readings of Magnetic Resonance Imaging Features of Lumbar Spinal Stenosis. *Spine (Phila Pa 1976)* (2008) 33:1605–10. doi: 10.1097/BRS.0b013e3181791af3
20. Fardon DF, Williams AL, Dohring EJ, Murtagh FR, Gabriel Rothman SL, Sze GK. Lumbar Disc Nomenclature: Version 2.0: Recommendations of the Combined Task Forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. *Spine J* (2014) 14:2525–45. doi: 10.1016/j.spinee.2014.04.022
21. Snell J, Swersky K, Zemel RS. Prototypical Networks for Few-Shot Learning, in: *Advances in Neural Information Processing Systems* (2017). pp. 4077–87.
22. Yang HM, Zhang XY, Yin F, Liu CL. Robust Classification With Convolutional Prototype Learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018). pp. 3474–3482.
23. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016). pp. 770–778.
24. Ooi BC, Tan KL, Wang S, Wang W, Cai Q, Chen G, et al. SINGA: A Distributed Deep Learning Platform. In Proceedings of the 23rd ACM International Conference on Multimedia (2015). pp. 685–8.
25. Luo Z, Yeung SH, Zhang M, Zheng K, Zhu L, Chen G, et al. MLCask: Efficient Management of Component Evolution in Collaborative Data Analytics Pipelines. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (2021). pp. 1655–66. IEEE.
26. Miyato T, Maeda SI, Koyama M, Ishii S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* (2018) 41(8):1979–93.
27. Gwet KL. Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *Br J Math Stat Psychol* (2008) 61(Pt 1):29–48. doi: 10.1348/000711006X126600
28. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* (1977) 33:159–74. doi: 10.2307/2529310

29. Massaad E, Fatima N, Hadzipasic M, Alvarez-Breckenridge C, Shankar GM, Shin JH. Predictive Analytics in Spine Oncology Research: First Steps, Limitations, and Future Directions. *Neurospine* (2019) 16:669–77. doi: 10.14245/ns.1938402.201

30. Merali ZA, Colak E, Wilson JR. Applications of Machine Learning to Imaging of Spinal Disorders: Current Status and Future Directions. *Global Spine J* (2021) 11(1_suppl):23S–9S. doi: 10.1177/2192568220961353

31. Liu X, Li KW, Yang R, Geng LS. Review of Deep Learning Based Automatic Segmentation for Lung Cancer Radiotherapy. *Front Oncol* (2021) 11:717039. doi: 10.3389/fonc.2021.717039

32. Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A, et al. Deep Learning for Segmentation in Radiation Therapy Planning: A Review. *J Med Imaging Radiat Oncol* (2021) 65:578–95. doi: 10.1111/1754-9485.13286

33. Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, et al. Automatic Multiorgan Segmentation in Thorax CT Images Using U-Net-GAN. *Med Phys* (2019) 46:2157–68. doi: 10.1002/mp.13458

34. Merali Z, Wang JZ, Badhiwala JH, Witiw CD, Wilson JR, Fehlings MG. A Deep Learning Model for Detection of Cervical Spinal Cord Compression in MRI Scans. *Sci Rep* (2021) 11:10473. doi: 10.1038/s41598-021-89848-3

35. National Institute for Health and Care Excellence. *Metastatic Spinal Cord Compression: Diagnosis and Management of Adults at Risk of and With Metastatic Spinal Cord Compression NICE Guidelines (CG75)*. London: NICE (2008).

36. Gottumukkala S, Srivastava U, Brocklehurst S, Mendel JT, Kumar K, Yu FF, et al. Fundamentals of Radiation Oncology for Treatment of Vertebral Metastases. *Radiographics* (2021) 41:2136–56. doi: 10.1148/rg.2021210052

37. Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning. *IEEE Trans Neural Networks* (2009) 20:542. doi: 10.1109/TNN.2009.2015974

38. Zhu L, Yang K, Zhang M, Chan LL, Ng TK, Ooi BC. Semi-Supervised Unpaired Multi-Modal Learning for Label-Efficient Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2021 Sep 27 (pp. 394-404). Springer, Cham.

39. Zhang W, Zhu L, Hallinan J, Makmur A, Zhang S, Cai Q, Ooi BC. BoostMIS: Boosting Medical Image Semi-supervised Learning with Adaptive Pseudo Labeling and Informative Active Annotation. arXiv preprint arXiv:2203.02533. 2022 Mar 4.

40. van Tol FR, Massier JRA, Frederix GWJ, Öner FC, Verkooijen HM, Verlaan JJ. Costs Associated With Timely and Delayed Surgical Treatment of Spinal Metastases. *Global Spine J* (2021):2192568220984789. doi: 10.1177/2192568220984789

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fonc.2022.825353

Computational Analysis of Pathological Image Enables Interpretable Prediction for Microsatellite Instability

Jin Zhu[1], Wangwei Wu[1], Yuting Zhang[1], Shiyun Lin[2], Yukang Jiang[1], Ruixian Liu[3], Heping Zhang[4]* and Xueqin Wang[5]**

[1] Southern China Center for Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, China, [2] Center for Statistical Science, School of Mathematical Sciences, Peking University, Beijing, China, [3] Department of Clinical Laboratory, The Sixth Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China, [4] School of Public Health, Yale University, New Haven, CT, United States, [5] Department of Statistics and Finance/International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, China

OPEN ACCESS

Edited by:
*Haibin Shi,
Soochow University, China*

Reviewed by:
*Arif Ozgun Harmanci,
University of Texas Health Science
Center at Houston, United States
Yao Sun,
Central China Normal University, China*

***Correspondence:**
*Ruixian Liu
liurx25@mail.sysu.edu.cn
Heping Zhang
heping.zhang@yale.edu
Xueqin Wang
wangxq20@ustc.edu.cn*

Specialty section:
*This article was submitted to
Cancer Imaging and
Image-directed Interventions,
a section of the journal
Frontiers in Oncology*

Received: *30 November 2021*
Accepted: *21 June 2022*
Published: *22 July 2022*

Citation:
Zhu J, Wu W, Zhang Y, Lin S, Jiang Y, Liu R, Zhang H and Wang X (2022) Computational Analysis of Pathological Image Enables Interpretable Prediction for Microsatellite Instability. Front. Oncol. 12:825353. doi: 10.3389/fonc.2022.825353

Background: Microsatellite instability (MSI) is associated with several tumor types and has become increasingly vital in guiding patient treatment decisions; however, reasonably distinguishing MSI from its counterpart is challenging in clinical practice.

Methods: In this study, interpretable pathological image analysis strategies are established to help medical experts to identify MSI. The strategies only require ubiquitous hematoxylin and eosin–stained whole-slide images and perform well in the three cohorts collected from The Cancer Genome Atlas. Equipped with machine learning and image processing technique, intelligent models are established to diagnose MSI based on pathological images, providing the rationale of the decision in both image level and pathological feature level.

Findings: The strategies achieve two levels of interpretability. First, the image-level interpretability is achieved by generating localization heat maps of important regions based on deep learning. Second, the feature-level interpretability is attained through feature importance and pathological feature interaction analysis. Interestingly, from both the image-level and feature-level interpretability, color and texture characteristics, as well as their interaction, are shown to be mostly contributed to the MSI prediction.

Interpretation: The developed transparent machine learning pipeline is able to detect MSI efficiently and provide comprehensive clinical insights to pathologists. The comprehensible heat maps and features in the intelligent pipeline reflect extra- and intra-cellular acid–base balance shift in MSI tumor.

Keywords: cancer, microsatellite instability, interpretability, deep learning, random forest

INTRODUCTION

Microsatellite instability (MSI) is the condition of genetic hypermutability that results from impaired DNA mismatch repair. Cells with abnormally functioning mismatch repair are unable to correct errors that occur during DNA replication and consequently accumulate errors. MSI has been frequently observed within several types of cancer, most commonly in colorectal, endometrial, and gastric adenocarcinomas (1). The clinical significance of MSI has been well described in colorectal cancer (CC), as patients with MSI-high colorectal tumors have been shown to have improved prognosis compared with those with MSS (microsatellite stable) tumors (2). In 2017, the U.S. Food and Drug Administration approved anti–programmed cell death-1 immunotherapy for mismatch repair deficiency/MSI-high refractory or metastatic solid tumors, making the evaluation of DNA mismatch repair deficiency an important clinical task. However, in clinical practice, not every patient is tested for MSI, because this requires additional next-generation sequencing (3, 4), polymerase chain reaction (5), or immunohistochemical tests (1, 6, 7). Thus, it is in high demand for a cheap, effective, and convenient classifier to assist experts in distinguishing MSI vs. MSS.

Numerous publications have identified histologic features that are more commonly seen in MSI. By far, it is a well-known fact that tumors that have undifferentiated morphology, poor differentiation, and the high infiltration of TIL cells are more likely to be MSI (8–11). Unfortunately, it is still challenging to distinguish MSS from MSI based on pathologist's visual inspections from pathological images because the morphology of MSS is similar to that of MSI (12). The recent technical development of high-throughput whole-slide scanners has enabled effective and fast digitalization of histological slides to generate WSIs. More importantly, the thriving of various machine learning (ML) methods in image processing makes this task accessible. In recent years, ML has been broadly deployed as a diagnostic tool in pathology (13, 14). For example, Iizuka et al. built up convolutional neural networks (CNNs) and recurrent neural networks to classify WSI into adenocarcinoma, adenoma, and non-neoplastic (15). The study by Bar et al. demonstrated the efficacy of the computational pathology framework in the non-medical image databases by training a model in chest pathology identification (16). Notably, deep learning (DL) model has been used to predict MSI directly from H&E histology and reported the network achieved desirable performance in both gastric stomach adenocarcinoma (STAD) and CC (17). These studies attest to the great potential of ML methods in medical research and clinical practice.

There is no doubt that the ML revolution has begun, but the lack of the "interpretability" of ML is of particular concern in healthcare (18, 19). Here, the "interpretability" means that clinical experts and researchers can understand the logic of decision or prediction produced by ML methods (20). In essence, it urges ML systems to follow a fundamental tenet of medical ethics, that is, the disclosure of necessary yet meaningful details about medical treatment to patients (21). Unfortunately, to the best of our knowledge, most of the existing MSI diagnosis systems, especially DL-based systems, are non-interpretable. Therefore, there is an urgent need to establish a new research paradigm in applying an interpretable ML system in medical pathology field (22–26).

In this study, we used H&E-stained WSI from TCGA: 360 formalin-fixed paraffin-embedded (FFPE) samples of CC (TCGA-CC-DX) (27), 285 FFPE samples of STAD (TCGA-STAD) (28), and 385 snap-frozen samples of CC (TCGA-CC-KR). H&E-stained images in these databases have already been tessellated into 108,020 (TCGA-STAD), 139,147 (TCGA-CC-KR), and 182,403 (TCGA-CC-DX) color-normalized tiles (17), and all of them only target region with tumor tissue. The aims of the study are as follows: (i) to build an image-based ML method on MSI classification and post-process the fed image to a heat map to interpret the diagnosis of MSI at an image level; and (ii) to design a fully transparent feature extraction pipeline and understand the pathological features' importance and interactions for predicting MSI by training a feature-based ML model.

Our contributions are two folds. First, we developed ML models with decent power in the prediction of MSI. This model can exhibit a visual heatmap demonstrating high-contribution regions for MSI prediction in the H&E image. Second, we certified certain pathological features with non-trivial importance in MSI classification, which is not explicitly studied in the previous research. Therefore, our study facilitates MSI diagnosis based on H&E image and sheds light on the understanding of MSI at both image-level and features level.

MATERIALS AND METHODS

Histopathology Image Sources

The whole-slide H&E-stained histopathology images were obtained from TCGA, including three cancer subtype datasets. Dataset DX consisted of 295 MSS patients and 65 MSI patients from FFPE samples of CC. Dataset KR contained 316 MSS patients and 72 MSI patients from snap-frozen samples of CC. Dataset STAD collected 225 MSS patients and 60 MSI patients of FFPE STAD. Two criteria in the published study (17) classify patients as MSI: (i) all the patients who were previously defined as MSI were included in the MSI group (29); and (ii) some patients with unknown MSI status but with a mutation count of >1,000 were also defined as MSI (30).

All the images used in our models have already gone through tumor tissue detection and have been tessellated into small tiles in J.N. Kather's work (https://zenodo.org/record/2530835 and https://doi.org/10.5281/zenodo.2532612). The proceeding for getting the tiles is of two steps. First, the tumor region is identified from WSI image, and second, the tumor is divided into small square subregions, called tiles, where the edge of each tile is 256 μm. There are 108,020 tiles in TCGA-STAD cohort, 139,147 in TCGA-CC-KR, and 182,403 in TCGA-CC-DX. Color normalization has already been performed on every tile using the Macenko method (31), which converts all images to a reference color space. In all cases, training and test sets were split on a patient level, and no image tiles from test patients were present in any training sets.

Details of Deep Learning and Grad-CAM

The DL model that we considered is ResNet-18, which is one of the state-of-the-art CNNs (17, 32). We adopted all of the default settings in ResNet-18 and did not fine-tune any hyperparameters on it. ResNet-18 is built in Python 3.7 with TensorFlow-GPU 1.14.0 and Keras 2.3.0. Because the ResNet-18 is insensitive to the adversarial samples, we did not pre-process any image tiles in the three TCGA datasets. The patient-level areas under the curve (AUCs), receiver operating characteristic (ROC) curves, and 95% stratified bootstrap confidence intervals (CIs) for ROC curves were computed and visualized by two R packages: pROC (33) and ggplot2 (34). Gradient-weighted Class Activation Mapping (Grad-CAM) utilizes the gradient information abundant in the last convolutional layer of a CNN and generates a rough localization map of the important regions in the image. We apply the rectified linear unit to the linear combination of maps to generate localization maps of the desired class. Grad-CAM visualization was implemented in Python 3.7 with TensorFlow-GPU 1.14.0 and Keras 2.3.0.

Image Pretreatment

Before feature extraction, we apply pretreatments to the tiles before feature extraction and we summarized the pretreatments and associated implementation details in **Table 1**. First, white balance is performed on our cohorts because the natural appearance tone of the object may alter in the formation of images when exposed in a lightning condition of different color temperature (37). Because every tile has an area without cell organization, i.e., without H&E stain, we could view that part as the neutral reference in adjustment. In addition to the color cast, overexposure and underexposure also may result in the distortion of our features (38). Still, taking the unstained area as the reference, we regulated all tiles into the same level of brightness. In addition, to get the location of immune cells' nuclei, we similarly perform color deconvolution (39, 40) to separate color space from immunohistochemical staining on each tile. Finally, to extract the Haralick texture features (41, 42) of tumor cells, we used a positive cell detection algorithm to locate every tumor cell in each tile and use its batch process to get needed features.

Feature Extraction

In global color feature extraction, the region of interest (ROI) is a stained area. We recorded mean value, quantiles (25%, 50%, and 70%), and higher-order moments (variance, kurtosis, and skewness) in ROI of each channel in RGB and HSV as our global features. Moreover, with Gaussian mixture model (GMM) model (43), we perform image segmentation to each tile to divide the ROI into three clusters and record the corresponding features in every cluster as our local features. We located immune cells' nuclei after color deconvolution according to their size and grayscale and calculated the amount as the feature. As for the differentiation degree of tissue in tiles, we performed dilation, erosion, and circle Hough transforms (44) to identify outlines similar to circle in images and to decide their differentiation degree. Because the more regular shapes exist, the more highly the tissue differentiates. Because we have recorded the tumor cell's location, we extract Haralick features of each tumor cell in one tile and adopt the mean value of all cells' as this tile feature *via* QuPath software (45). In addition, we also recorded the count of a tumor cell as our feature.

Details of Random Forest and Benchmark Machine Learning Methods

Our RF method was built and tested using Python version 3.7.1 with RandomForestClassifier in sklearn.ensemble library (46). During training, 70% of patients in every dataset were randomly selected, and all of their tiles were used in training, whereas the rest of the tiles were held out and used as test sets. There are some anomalous tiles in each dataset, i.e., blurred or color disorder, resulting in the loss of the information contained in them. Therefore, we disposed of all of them in every dataset. In addition, we also delete the tiles owning an extreme immune cell number (a value that significant in 1% level) because an extremely small number may represent the non-tumor area, whereas a too large number represents lymphatic concentration area. In each forest, we set 500 trees in total and take Gini impurity as the criterion. For each forest, we tune the minimum node size of random forest (RF), which is an important parameter to prevent overfitting, and we keep other parameters with the default settings. We used a simple tuning criterion as follows: Consider the candidate minimum node size: 15, 16, …, 25, and then the size associated with the least out-of-bag error of RF is chosen. The selected minimum node size is 23 for the STAD cohort, 17 for both the KR and the DX cohorts. Again, we used pROC packages to compute AUC and assess 95% stratified bootstrapped CIs and ggplot2 package to visualize the model performance.

Out of comparison, we also consider two benchmarking ML methods suggested by a reviewer including support vector machine (SVM) (47) and generalized linear model (GLM) (48). The ridge regularization in GLM is selected *via* 10-fold cross validation. Because hundreds of thousands of tiles brought huge computational burden, SVM ran very slow even in the state-of-the-art implementation (49), and thus, we did not tune the parameters in SVM and set them as default.

TABLE 1 | Pre-treatment, software, and parameters used in each pre-treatment.

Pretreatment	Software	Parameters
White balance	OpenCV-Python	Default
Brightness Adjustment	OpenCV-Python	Target average brightness in RR: 240
Color Deconvolution	ImageJ (35)	Default
Tumor Cell Identification	scikit-image (36)	Objects with size: 5–17

Reference region (RR): an area without cell organization, whose values in RGB channels within (180, 255). The parameters are manually selected according to the experience of image analysis for H&E images.

Permutation Feature Importance and Conditional Minimal Depth

Permutation-based feature importance (50) is a widely used model inspection technique for RF. It is defined to be the decline in a model accuracy when one feature's values are randomly shuffled. The shuffle procedure cancels the relationship between the label and the feature, and thus, the drop in the model accuracy can serve as a measurement for the importance of the feature in RF. An alternative feature importance, minimal depth (51), is defined as the depth when a feature splits for the first time in a tree. For example, if a feature splits the root node in a tree, then its minimal depth is 0. The mean of minimal depths over all trees in a forest can measure the feature importance. The importance ordering of features under it keeps highly consistent with the result from the permutation-based method (**Figure S4**).

To investigate the interaction between two different features, we used a generalization of minimal depth, conditional minimal depth, that measures the depth of the second feature in a subtree with the root node where the first feature splits (52). Specifically, we recorded all of such splits with the first feature and calculated the mean of conditional minimal depths of the second features given the first feature. A large gap between the mean of conditional minimal depth and the mean of minimal depth implies possibilities for the second feature being used for splitting after the first feature. The occurrence of the large gap implies that the two features have a strong interaction. We used R version 3.5.1 with randomForest package (53) to rebuild that RF and analyze and visualize the relations between different features with randomForestExplainer package (52).

Ablation Experiment for Deep Learning

Ablation experiment (54–56) is conducted to investigate the contribution of pathological features in DL. Specifically, we eliminated the RGB mean differences between MSI and MSS groups in the test set by adjusting the mean value in each tile in the test set to the mean value of all the tiles as a whole. Then, we feed the adjusted tiles in the test set into the trained neural network. The drops of AUCs after reevaluation can verify the contribution of the RGB feature in the classification of the DL network.

Role of the Funding Source

The funder of this study had no role in study design, data collection, data analysis, data interpretation, and writing of the report. The corresponding author had full access to study data and final responsibility for the decision to submit for publication.

RESULTS

A Deep Learning Classifier and Image-Level Visual Interpretability

We used a commonly used end-to-end CNN, ResNet-18 (32) in the study. To fit this DL model for different cancer subtypes, we trained three ResNet-18 networks based on 70% of the tiles randomly sampled from three datasets, the remaining 30% of the tiles in each dataset were used for testing. In the testing cohort, a patient's slide was predicted to be MSI if at least half of the tiles were predicted to be MSI. The patient-level accuracy and AUC were 0.84 in the KR cohort, 0.81 in the DX cohort, and 0.80 in the STAD cohort (**Figure 1B**).

On the basis of the trained DL model, the Grad-CAM was used to make the convolutional-based model more transparent by generating localization maps of the important regions (57). To unveil the hidden logic behind the DL and provide visual interpretability, we deployed Grad-CAM to find out which part of the H&E image supports DL's classification. Two typical images for interpreting DL prediction logic are shown (**Figure 1A**). The region highlighted by Grad-CAM points out the important region for DL decision but not statistical correlation. Our pathologist noted that the highlighted region in **Figure 1A** tended to be where immune cells are mainly concentrated in the tumor organism; meanwhile, we also found that the highlighted region presented distinct color and texture characteristics. We were intrigued by this phenomenon and further examined this important region in great detail.

Transparent Pathological Image Analysis Workflow and Feature-Based Classification Model

The results from Grad-CAM suggested that certain features of the H&E-stained images might encode essential regions of the tumor organism. To further investigate this, we developed a multi-step, automatic and transparent workflow (**Figure 2**). In the first step, we standardized the three image datasets by standard image processing techniques (e.g., white balance and brightness adjustments). After the image pre-processing, we extracted visible pathological features. Motivated by the feedback from Grad-CAM and existing studies (9, 58, 59), we focused on these H&E feature characteristics: global and local color features in RGB and HSV channels, the numbers of infiltrating immune cells and tumor cells, the grading of differentiation, and the texture features from tumor cells. A total of 182 features were extracted from each image tile, and some representative ones are displayed in **Figure 3**.

We then applied RF (50), one of the most popular ML algorithms, to all three databases to classify MSI versus MSS on H&E-stained histology slides. We randomly selected 70% of patients in every dataset during training, and all their tiles were used in training, whereas the rest of the tiles were held out and used as test sets. In the test sets of each dataset, true MSS image tiles cohort had a median MSS score (the proportion of the prediction result judged to be MSS in each decision tree of the forest) that was significantly different from those of MSI tiles (the P-values of the two-tailed *t* test were 0.02, 0.0024, and 0.002 in the three datasets), indicating that our models can distinguish MSI from MSS. Because one patient may have many different tiles, we obtained the patient-level MSI scores by averaging the RF's prediction on all its tiles. AUCs for MSI detection were 0.78 (95% CI: 0.7–0.82) in KR cohort, 0.7 (95% CI: 0.65–0.74) in DX cohort, and 0.74 (95% CI: 0.65–0.79) in STAD cohort (see **Figures 4B**, **Figures S1B, S2B**). These results show that visible pathological features can be useful in MSI prediction. Comparing the AUCs of DL and RF, we can see that DL is superior to RF in

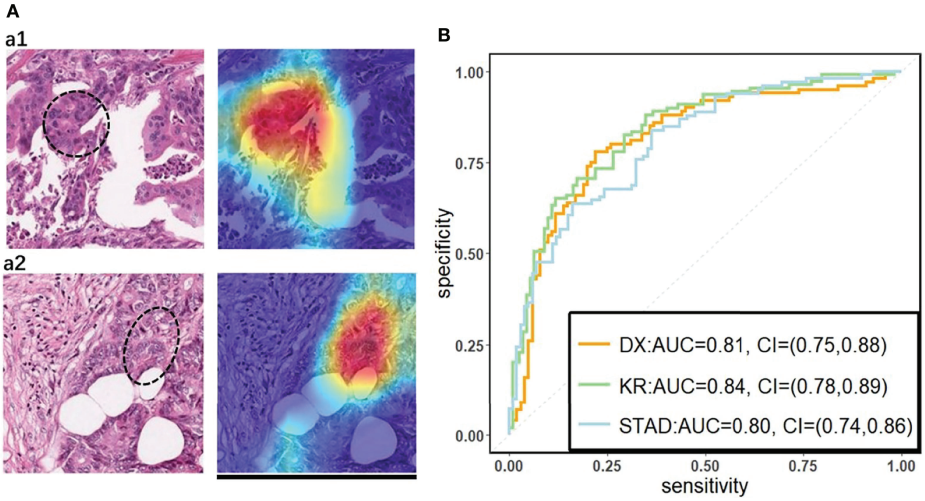


FIGURE 1 | **(A)** The original tile and the corresponding heatmap output by the GCAM. The image in the left of **(A1)** and **(A2)** display tiles from the TCGA-CC-DX dataset labeled with MSI and MSS, respectively. The ellipse upon the images corresponds the most contributed region revealed by GCAM. In the heatmaps, the brighter region contributes more to the classification. For instance, the red one is the most highlighted area, while the blue regions contribute limitedly. Scale bar, 256 μm **(B)** Patient-level receiver operating characteristic (ROC) curve for classifying MSI versus MSS in the three datasets with deep learning. The 95% confidence intervals (CI) were computed by the bootstrap method.

prediction, yet we would show that RF can reveal informative messages about the impact of pathological features on MSI prediction. From the comparison among RF, SVM, and GLM, we see that, from predictive power, RF surpasses the other benchmarking ML methods.

Feature-Level Visual Interpretability: Feature Importance and Interactions

One of the attractive advantages of RF is that it can evaluate the importance of the features. Therefore, we verify and quantify these features' power in distinguishing MSI from MSS by extracting information from a trained model. A representative pattern can be discovered from the visualization of permutation-based feature importance (50, 60) in the KR dataset (**Figure 4A**) . From the figure, we can deduce that the texture features play a dominant role. Because the texture features reflect the surface's average smoothness of the tumor cells in one tile, we deduce that the characteristics of the tumor surface are an important clinical indicator in automatic MSI diagnosis. Color features also have important contributions. In the global color feature, the higher-

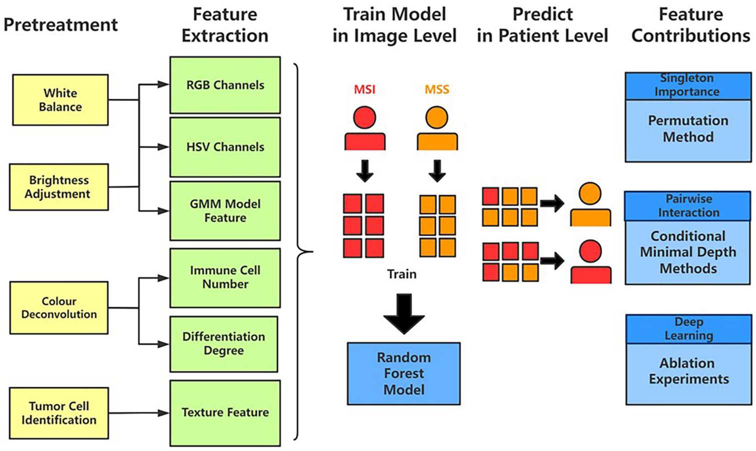


FIGURE 2 | The workflow of studying pathological features in discriminating against MSI from MSS. Five main steps—pretreatments, feature extraction, model training, patient-level predictions, and feature contributions analysis—were sequentially executed to improve image quality, generate pathological features, build statistical model, evaluate model performance, and measure features' contributions, respectively.

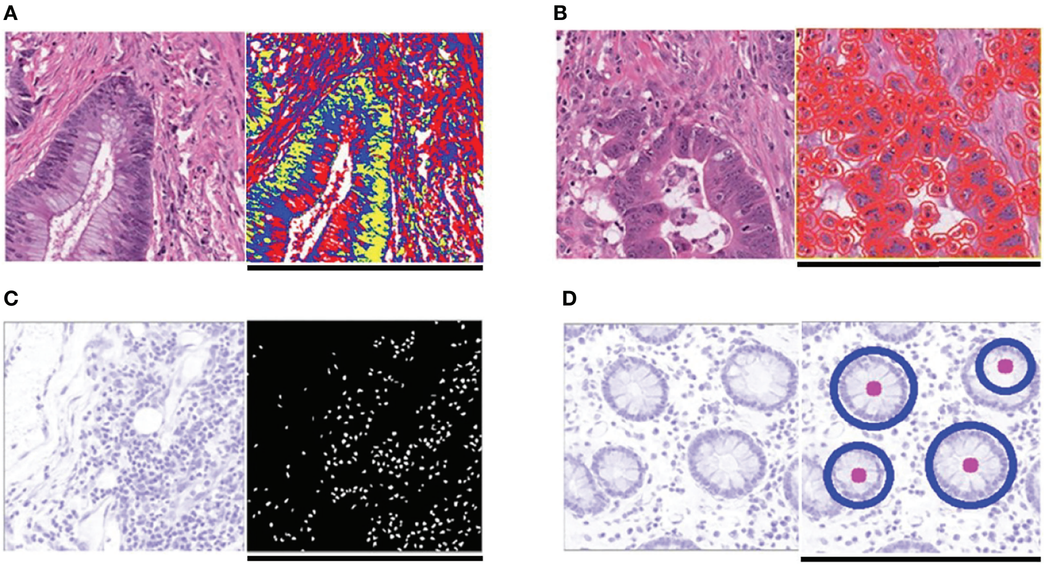


FIGURE 3 | Typical feature extraction result. **(A)** GMM model for image segmentation. The figure on the left is a tile from the TCGA-CC-DX dataset, and its image segmentation tiles processed by the GMM method are shown in the figure on the right. The green part whose grayscale is the lowest among the three parts tends to be tumor tissue, whereas the blue and red ones represent non-tumor tissue. **(B)** Tumor cell detection before Haralick texture identification. The figure on the left is an original tile, whereas the one on the right is processed with tumor identification. Each red circle in the tile on the right indicates the boundary of one tumor cell. **(C)** Infiltrating immune cells detection. The detection of immune cells allows us to calculate the connectivity domain. **(D)** The grading of differentiation. Detect the circularly similar arrangement in one slice and grade the degree of differentiation based on its amount. Scale bar, 256 μm.

order statistics (skew and kurtosis) contribute more than the first-order statistics (mean and quantile), indicating that some useful information contributing to classification are hidden in high-order features. Local color features also deserve our attention. Compared with global color features, the local ones were useful in image segmentation by dividing slices into different clusters, and we obtained the information in each cluster. **Figure 3** demonstrates the clinical utility of the clusters as they closely reflected tumor tissue versus non-tumor tissue. The number of infiltrating immune cells was also important as expected, whereas the differentiation grade contributed the least in every dataset.

It is widely accepted that feature interactions (i.e., the joint effect of features) can be important for the complex disease (61–64). Our feature-based RF models also allow us to exploit the pairwise feature interactions in MSI classification, and thus, we can attain a more clear understanding of the characteristics of MSI tiles and the mechanism of RF. Here, we use conditional minimal depth (51) to quantitatively assess feature interaction and then demonstrate the foremost 15 pairwise interactions (**Figures 4C**, **Figures S2C, S3C**). The feature types with the most effective interaction effect with other features in each dataset are the local color feature in KR, the global color feature in DX, and texture features in STAD. The three features enhanced the importance of the features interacting with them, even the features themselves may have a weak effect before. It is also worthy to note that interactions incline to occur more often between color features and texture features or between local color and global color features. To understand how the paired features jointly help the MSI diagnosis, we plot the prediction values of typical feature interaction on a grid diagram (**Figure 5** and **Figure S3**). In the KR dataset, a greater immune cell number and a lower value of the 75th percentile of red channel lead to a higher probability of MSS. In DX, a higher value of the max caliper in tumor cells and a fewer tumor cell number lead to a higher probability of MSS. In STAD, a lower value of the optical density range of tumor cells' nucleus in Hematoxylin stains and a higher value of texture feature correlation in eosin stains lead to a higher probability of MSS.

DISCUSSION

To our knowledge, this is the first study to not only build up a classification model in distinguishing MSI from MSS but also provide an interpretability analysis. Previous studies in investigating the pathologic predictors of MSI through feature extraction and logistics regression model suffered from the limited learning capability as well as the small sample size and thus could not achieve satisfactory performance (9). Other works on MSI classification paid attention to the enhancement of the prediction accuracy by establishing a DL network but did not provide a detailed description of the mechanism behind the model (17). In this study, we tackled these problems through using three different cancer types datasets from TCGA and

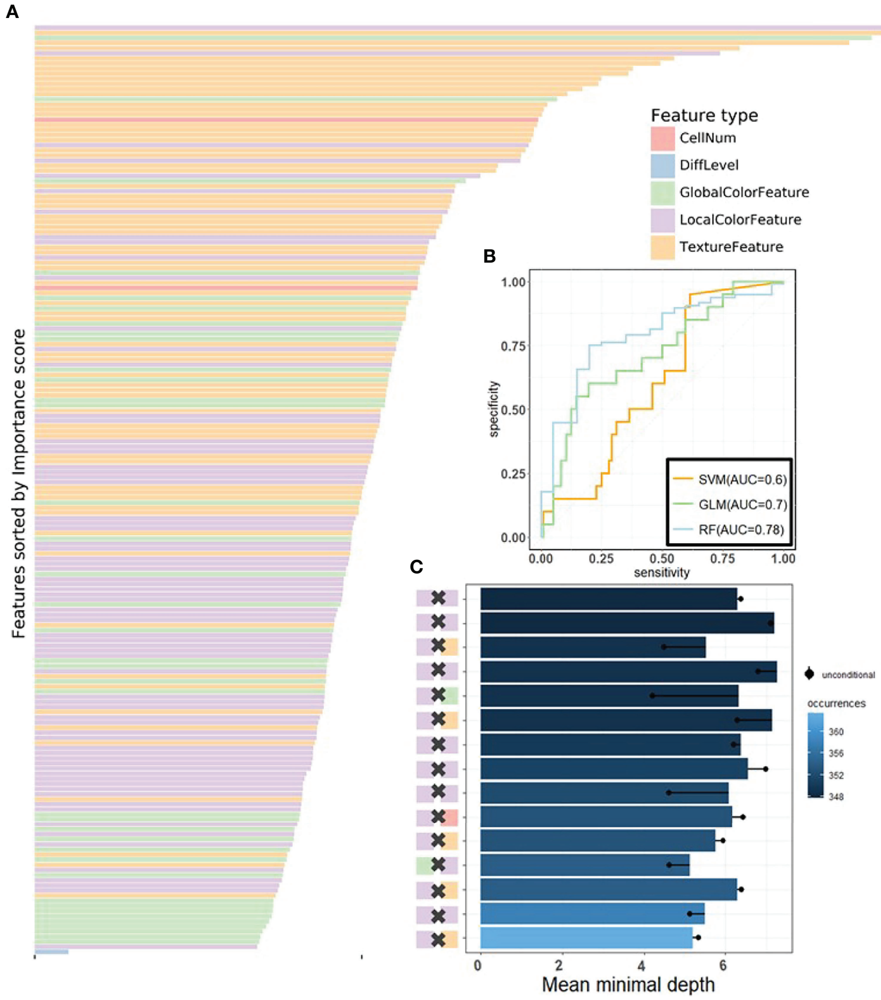


FIGURE 4 | The visualization of performance and interpretability of the RF in KR dataset. **(A)** The bar plot of permutation-based variable importance. Features are arranged from top to bottom in order of importance (the names of the features are provided in the order in **Table S2**). **(B)** The patient-level ROC curve for classifying MSI versus MSS with random forest. Three colors distinguish GLM, SVM, and RF. The 95% confidence intervals (CIs) computed by the bootstrap method are as follows: (0.53, 0.83) for GLM, (0.49, 0.72) for SVM, and (0.70, 0.82) for RF. **(C)** The bar plot of the mean of conditional minimal depth (the top 15 feature pairs of interaction are shown). A feature pair of interaction is listed as A × B, where A and B are one of feature type and their concrete names are listed in **Table S3**. Feature pairs are arranged from the bottom to top in the order of the occurrences, which are represented by the color intensity of the bars. The bar's length indicates the mean of conditional minimal depth and the distance from the dot to the y-axis measures the mean of minimal depth of **(B)** The length of the dot line implies the gap between them, measuring the effect of pairwise feature interaction. A large gap implies a strong interaction (see also **Figures S1–S3**).

following the framework of interpretability with two steps: first, built up a high-performance DL network with a visual explanation capacity as model-based interpretability; second, we further analyzed and confirmed features' power using a feature-based interpretable model.

To build an interpretable DL network, we trained residual learning CNNs and deployed Grad-CAM to the final convolutional layer of the network to produce the heatmap that reflects the highly contributed region. Notably, through its coarse localization map of the image's essential regions, it provided preliminary insight into highly contributed pathological features. It is worthy to note that the prediction performance of our method is also desirable, and it is comparable to the predictors proposed in other published research (17). Although Grad-CAM is also used in the recent literature, they just use it to quantify possible differences between real and synthetic images.

To understand the contribution of the pathological features on MSI classification, we manually extracted the clinically meaningful features *via* image processing methods, trained an RF classifier based on those features, assessed the importance of those features, and exploited their interaction. This procedure achieves feature level interpretability at the expense of prediction performance; however, we interestingly found that the texture and color of the H&E image

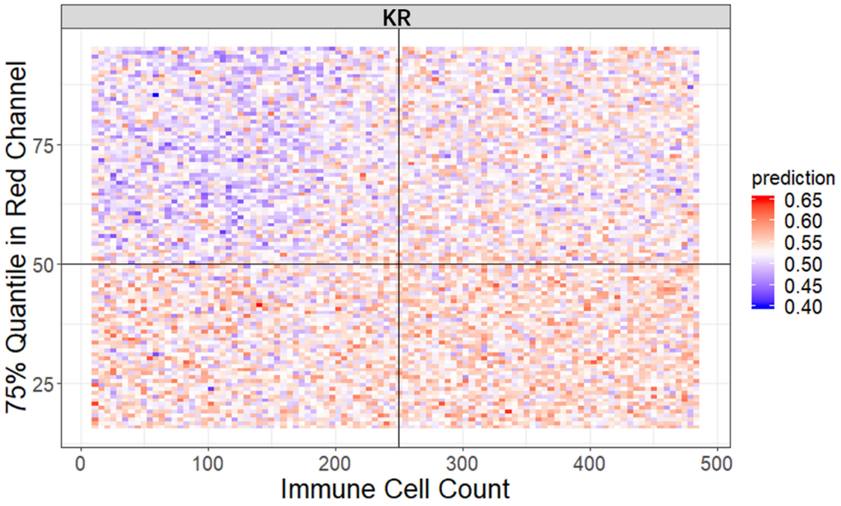


FIGURE 5 | The visualization of typical pairwise features' interaction in KR dataset. The prediction value ranges from 0 to 1 with color from blue to red. The bluer means a larger probability of MSI, whereas the redder tends to be MSS (see also **Figure S3**).

and the interactions among them were crucial for diagnosing MSI. To the best of our knowledge, this has not been noted before. From the widely studied underlying biology of immune infiltration in MSI, numerous pieces of evidence indicate that a high tumor mutational burden increases the likelihood that immunogenic neoantigens expressed by tumor cells induce increased immune infiltration (65–67). In addition, color feature is regard as an important feature for the diagnosis of TFE3 Xp11.2 translocation renal cell carcinoma *via* WSI (58). Finally, the pivotal roles of color and texture features found in our study reflect extra- and intracellular acid–base balance shift in MSI tumor (68). Another interesting fact is that the feature type that tends to interact with the other features has a clear difference in the three datasets due to the image heterogeneity raised from the diversity of cancer type (CC or STAD) and tissue preservation methods (snap-frozen or FFPE) (69), indicating that the feature interaction mode was influenced by preservation methods and tumor types. However, this insight would not be attained from "black-box" ML method. Moreover, we hypothesized that the dominant-role features such as color in RF models were also important in the DL model. To test our hypothesis, we eliminated the mean color differences between MSI and MSS groups and reevaluated our DL models' AUCs. Specifically, we calculated the RGB mean value of all tiles in both groups and centralized the RGB mean value of every tile into that population mean value. We found that the AUCs were reduced by 0.11, 0.12, and 0.14 in DX, KR, and STAD datasets, respectively, supporting our hypothesis that color features also contributed to the DL model.

We note that our findings warrant replications through further biological experiments. The H&E stain is capable of highlighting the fine structures of cells and tissues. Most cellular organelles and extracellular matrix are eosinophilic, whereas the nucleus, rough endoplasmic reticulum, and ribosomes are basophilic. Our study shows that the spectrum, intensity, and texture of colors matter in distinguishing MSI from MSS, which needs further validation. We hypothesize that MSI tumor usually has distinct color/texture characteristics due to diverse gene mutation pattern (1, 70). Furthermore, the methodology of this study could be applied to the pathological analysis of other diseases, like infectious, in which color/texture characteristics of the H&E images are also crucial for disease diagnosis. One limitation of this study is that the cases in TCGA datasets may not be an unbiased collection from the real situation because pathologists may only upload the representative ones. Although our model performed well in these histopathology images, we should admit that their performance in the actual clinical settings requires further research. Therefore, one of our future direction is integrating more available datasets considered in (71), and we point out that it can naturally improve the specificity and control sensitivity simultaneously. Another limitation is that our study only focused on H&E-stained images, and we could not confirm whether the pattern in this study, especially the color features' contribution, works in other types of histopathology slices. The classifier models, which can be used for the diagnosis of other cancer types based on immunochemical stained images and *in vivo* images (72, 73), remain to be explored and established.

Further, our framework provides a positive feedback cycle in assisting pathologist's diagnosis of MSI (**Figure 6**). Specifically, the localization map outputted by our DL models can help experts to narrow their focus on the specific region of the whole H&E slide, thereby contributing to a more accurate and apprehensible diagnosis with the prediction result of our model. The features' distribution under our interpretable model can provide experts with more insight into analyzing the slices of MSI and MSS from clinical perspectives. Further, considering the similar feature distribution pattern in three datasets that we used, it is possible that, after running the same pipeline on MSI H&E slides under different cancer types, we can discover a generalization pattern

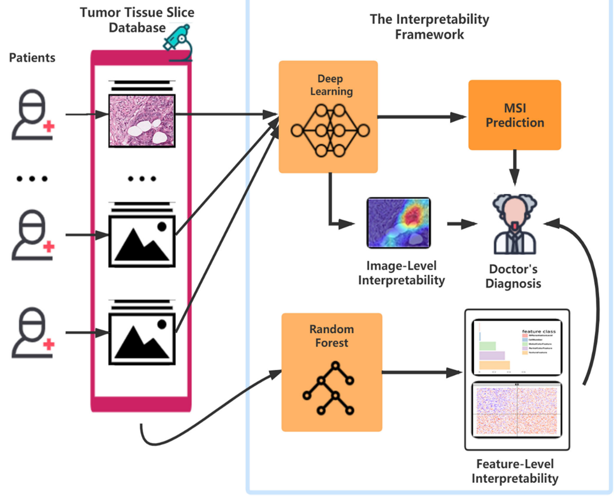

Tumor Tissue Slice Database
Patients
The Interpretability Framework
Deep Learning
MSI Prediction
Image-Level Interpretability
Doctor's Diagnosis
Random Forest
Feature-Level Interpretability

FIGURE 6 | The flowchart of the pattern in which our framework can assist the doctor's diagnosis. After surgery or biopsy, the embed cut H&E provided by each patient would go through MSI screening with deep learning. The doctor can make a critical diagnosis based on his insight combined with the prediction result and the deep learning model's visualization. Meanwhile, with the amplifying of the H&E datasets, the random forest could develop a more precise and interpretable model, which helps the doctors detect MSI.

behind them. After training on a larger dataset, the accuracy of the identification and the interpretability could improve, thereby contributing to accurate sample curation and treatment development of this aggressive cancer subtype.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://zenodo.org/record/2530835.

AUTHOR CONTRIBUTIONS

JZ, WW, RL, and XW conceived and designed the study. JZ, WW, YZ, and YJ performed the statistical and computational analysis. Funding acquisition: RL, XW, and HZ, SL helped manuscript editing. All co-authors review and modify the manuscript and approving its final version.

FUNDING

Wang's research is partially supported by NSFC(72171216, 71991474), the International Science & Technology cooperation program of Guangdong, China(2016B050502007), the Key Research and Development Program of Guangdong, China (2019B020228001), and Science and Technology Program of Guangzhou, China (202002030129). RL's research is supported by the National Natural Science Foundation of China (81902381). HZ's research is supported in part by U.S. National Institutes of Health (R01HG010171 and R01MH116527).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.825353/full#supplementary-material

REFERENCES

1. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and Characterization of Microsatellite Instability Across 18 Cancer Types. *Nat Med* (2016) 22(11):1342. doi: 10.1038/nm.4191
2. Popat S, Hubner R, Houlston RS. Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis. *J Clin Oncol* (2005) 23(3):609–18. doi: 10.1200/JCO.2005.01.086
3. Cohen R, Hain E, Buhard O, Guilloux A, Bardier A, Kaci R, et al. Association of Primary Resistance to Immune Checkpoint Inhibitors in Metastatic Colorectal Cancer With Misdiagnosis of Microsatellite Instability or Mismatch Repair Deficiency Status. *JAMA Oncol* (2019) 5(4):551–5. doi: 10.1001/jamaoncol.2018.4942
4. Cheng DT, Prasad M, Chekaluk Y, Benayed R, Sadowska J, Zehir A, et al. Comprehensive detection of germline variants by MSK-IMPACT, A Clinical Diagnostic Platform for Solid Tumor Molecular Oncology and Concurrent Cancer Predisposition Testing. *BMC Med Genomics* (2017) 10(1):33. doi: 10.1186/s12920-017-0271-4
5. Suraweera N, Duval A, Reperant M, Vaury C, Furlan D, Leroy K, et al. Evaluation of Tumor Microsatellite Instability Using Five Quasimonomorphic Mononucleotide Repeats and Pentaplex PCR. *Gastroenterology* (2002) 123(6):1804–11. doi: 10.1053/gast.2002.37070
6. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, et al. Performance Evaluation for Rapid Detection of Pan-Cancer Microsatellite Instability With MANTIS. *Oncotarget* (2016) 8(5):7452–63. doi: 10.18632/oncotarget.13918

7. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite Instability: A Review of What the Oncologist Should Know. *Cancer Cell Int* (2020) 20(1):16. doi: 10.1186/s12935-019-1091-8
8. Jenkins MA, Hayashi S, O'shea A-M, Burgart LJ, Smyrk TC, Shimizu D, et al. Pathology Features in Bethesda Guidelines Predict Colorectal Cancer Microsatellite Instability: A Population-Based Study. *Gastroenterology* (2007) 133(1):48–56. doi: 10.1053/j.gastro.2007.04.044
9. Greenson JK, Huang S-C, Herron C, Moreno V, Bonner JD, Tomsho LP, et al. Pathologic Predictors of Microsatellite Instability in Colorectal Cancer. *Am J Surg Pathol* (2009) 33(1):126. doi: 10.1097/PAS.0b013e31817ec2b1
10. Jass J. Classification of Colorectal Cancer Based on Correlation of Clinical, Morphological and Molecular Features. *Histopathology* (2007) 50(1):113–30. doi: 10.1111/j.1365-2559.2006.02549.x
11. Alexander J, Watanabe T, Wu T-T, Rashid A, Li S, Hamilton SR. Histopathological Identification of Colon Cancer With Microsatellite Instability. *Am J Pathol* (2001) 158(2):527–35. doi: 10.1016/S0002-9440(10)63994-6
12. Sagaert X, Cutsem EV, Tejpar S, Prenen H, Hertogh GD. MSI Versus MSS Sporadic Colorectal Cancers: Morphology, Inflammation, and Angiogenesis Revisited. *J Clin Oncol* (2014) 32(3):495–. doi: 10.1200/jco.2014.32.3_suppl.495
13. Wong TY, Bressler NM. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *J Am Med Assoc* (2016) 316(22):2366–7. doi: 10.1001/jama.2016.17563
14. Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin M-J, Diamond J, et al. Translational AI and Deep Learning in Diagnostic Pathology. *Front Med* (2019) 6(185). doi: 10.3389/fmed.2019.00185
15. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci Rep* (2020) 10(1):1–11. doi: 10.1038/s41598-020-58467-9
16. Bar Y, Diamant I, Wolf L, Greenspan H. Deep Learning With non-Medical Training Used for Chest Pathology Identification. *Med Imaging 2015: Computer-Aided Diagnosis; 2015: Int Soc Optics Photonics* (2015) 9414. doi: 10.1117/12.2083124
17. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep Learning can Predict Microsatellite Instability Directly From Histology in Gastrointestinal Cancer. *Nat Med* (2019) 25(7):1054–6. doi: 10.1038/s41591-019-0462-y
18. Towards Trustable Machine Learning. *Nat Biomed Eng* (2018) 2(10):709–10. doi: 10.1038/s41551-018-0315-x
19. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of Machine Learning-Based Prediction Models in Healthcare. *Wiley Interdiscip Reviews-Data Min Knowledge Discovery* (2020) 10(5):e1379. doi: 10.1002/widm.1379
20. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc Natl Acad Sci* (2019) 116(44):22071–80. doi: 10.1073/pnas.1900654116
21. Jacobson NC, Bentley KH, Walton A, Wang SB, Fortgang RG, Millner AJ, et al. Ethical Dilemmas Posed by Mobile Health and Machine Learning in Psychiatry Research. *Bull World Health Organ* (2020) 98(4):270. doi: 10.2471/BLT.19.237107
22. Schaumberg AJ, Juarez-Nicanor WC, Choudhury SJ, Pastrián LG, Pritt BS, Pozuelo MP, et al. Interpretable Multimodal Deep Learning for Real-Time Pan-Tissue Pan-Disease Pathology Search on Social Media. *Modern Pathol* (2020) 1–17. doi: 10.1038/s41379-020-0540-1
23. Vellido A. Societal Issues Concerning the Application of Artificial Intelligence in Medicine. *Kidney Dis* (2019) 5(1):11–7. doi: 10.1159/000492428
24. Piano SL. Ethical Principles in Machine Learning and Artificial Intelligence: Cases From the Field and Possible Ways Forward. *Humanities Soc Sci Commun* (2020) 7(1):1–7. doi: 10.1057/s41599-020-0501-9
25. Elshawi R, Al-Mallah MH, Sakr S. On the Interpretability of Machine Learning-Based Model for Predicting Hypertension. *BMC Med Inf decision making* (2019) 19(1):146. doi: 10.1186/s12911-019-0874-0
26. Lee E, Choi J-S, Kim M, Suk H-I. Initiative AsDN. Toward an Interpretable Alzheimer's Disease Diagnostic Model With Regional Abnormality Representation *via* Deep Learning. *NeuroImage* (2019) 202:116113. doi: 10.1016/j.neuroimage.2019.116113
27. Network CGA. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* (2012) 487(7407):330. doi: 10.1038/nature11252
28. Network CGAR. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* (2014) 513(7517):202–9. doi: 10.1038/nature13480
29. Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, et al. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* (2018) 33(4):721–35.e8. doi: 10.1016/j.ccell.2018.03.010
30. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* (2018) 173(2):371–85.e18. doi: 10.1016/j.cell.2018.02.060
31. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A Method for Normalizing Histology Slides for Quantitative Analysis. In Boston: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, vol. 2009. IEEE (2009). p. 1107–10.
32. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Las Vegas *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 2016. IEEE (2016). p. 770–8.
33. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinf* (2011) 12(1):77. doi: 10.1186/1471-2105-12-77
34. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. springer, New York (2016).
35. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 Years of Image Analysis. *Nat Methods* (2012) 9(7):671–5. doi: 10.1038/nmeth.2089
36. Van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. Scikit-Image: Image Processing in Python. *PeerJ* (2014) 2:e453. doi: 10.7717/peerj.453
37. Linden MA, Sedgewick GJ, Ericson M. An Innovative Method for Obtaining Consistent Images and Quantification of Histochemically Stained Specimens. *J Histochem Cytochem* (2015) 63(4):233–43. doi: 10.1369/0022155415568996
38. Kuru K. Optimization and Enhancement of H&E Stained Microscopical Images by Applying Bilinear Interpolation Method on Lab Color Mode. *Theor Biol Med Model* (2014) 11(1):1–22. doi: 10.1186/1742-4682-11-9
39. Ruifrok AC, Johnston DA. Quantification of Histochemical Staining by Color Deconvolution. *Analytical quantitative cytol Histol* (2001) 23(4):291–9.
40. Yi F, Huang J, Yang L, Xie Y, Xiao G. Automatic Extraction of Cell Nuclei From H&E-Stained Histopathological Images. *J Med Imaging* (2017) 4 (2):027502. doi: 10.1117/1.JMI.4.2.027502
41. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Trans Systems Man Cybernetics* (1973) SMC-3(6):610–21. doi: 10.1109/TSMC.1973.4309314
42. Azevedo Tosta TA, de Faria PR, Neves LA, do Nascimento MZ. Evaluation of Statistical and Haralick Texture Features for Lymphoma Histological Images Classification. *Comput Methods Biomechanics Biomed Engineering: Imaging Visualization* (2021) 1–12. doi: 10.1080/21681163.2021.1902401
43. McLachlan GJ, Peel D. *Finite Mixture Models*. John Wiley & Sons, New York (2004).
44. Yuen H, Princen J, Illingworth J, Kittler J. Comparative Study of Hough Transform Methods for Circle Finding. *Image Vision computing* (1990) 8 (1):71–7. doi: 10.1016/0262-8856(90)90059-E
45. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open Source Software for Digital Pathology Image Analysis. *Sci Rep* (2017) 7(1):1–7. doi: 10.1038/s41598-017-17204-5
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
47. Bishop CM. *Pattern Recognition and Machine Learning*. (2006) 128(9):326–345.
48. McCullagh P, Nelder JA. *In: London Generalized Linear Models*. Routledge (2019).
49. Wen Z, Shi J, Li Q, He B, Chen J. ThunderSVM: A Fast SVM Library on GPUs and CPUs. *J Mach Learn Res* (2018) 19(1):797–801.
50. Breiman L. Random Forests. *Mach Learn* (2001) 45(1):5–32. doi: 10.1023/A:1010933404324
51. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. *J Am Stat Assoc* (2010) 105(489):205–17. doi: 10.1198/jasa.2009.tm08622

52. Paluszynska A, Biecek P. *Randomforestexplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. R Package Version 09*. (2017).
53. Wiener A. Classification and Regression by Randomforest. *R News* (2002) 2:18–22.
54. Meyes R, Lu M, de Puiseau CW, Meisen T. Ablation Studies in Artificial Neural Networks. *arXiv preprint arXiv:190108644* (2019).
55. Sheikholeslami S, Meister M, Wang T, Payberah AH, Vlassov V, Dowling J. AutoAblation: Automated Parallel Ablation Studies for Deep Learning. In: New York *Proceedings of the 1st Workshop on Machine Learning and Systems*, vol. 2021. ACM (2021). p. 55–61.
56. Du L. How Much Deep Learning Does Neural Style Transfer Really Need? An Ablation Study. In: New York *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, vol. 2020. AMC (2020). p. 3150–9.
57. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization. In: Venice *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017. Kluwer Academic Publishers (2017). p. 618–26.
58. Cheng J, Han Z, Mehra R, Shao W, Cheng M, Feng Q, et al. Computational Analysis of Pathological Images Enables a Better Diagnosis of TFE3 Xp11. 2 Translocation Renal Cell Carcinoma. *Nat Commun* (2020) 11(1):1–9. doi: 10.1038/s41467-020-15671-5
59. Echle A, Laleh NG, Schrammen PL, West NP, Trautwein C, Brinker TJ, et al. Deep Learning for the Detection of Microsatellite Instability From Histology Images in Colorectal Cancer: A Systematic Literature Review. *ImmunoInformatics* (2021) 100008. doi: 10.1016/j.immuno.2021.100008
60. Biau G, Scornet E. A Random Forest Guided Tour. *Test* (2016) 25(2):197–227. doi: 10.1007/s11749-016-0481-7
61. Denisko D, Hoffman MM. Classification and Interaction in Random Forests. *Proc Natl Acad Sci* (2018) 115(8):1690–2. doi: 10.1073/pnas.1800256115
62. Hao N, Zhang HH. Interaction Screening for Ultrahigh-Dimensional Data. *J Am Stat Assoc* (2014) 109(507):1285–301. doi: 10.1080/01621459.2014.881741
63. Cordell HJ. Detecting Gene–Gene Interactions That Underlie Human Diseases. *Nat Rev Genet* (2009) 10(6):392–404. doi: 10.1038/nrg2579
64. Zhao Y, Chung M, Johnson BA, Moreno CS, Long Q. Hierarchical Feature Selection Incorporating Known and Novel Biological Information: Identifying Genomic Features Related to Prostate Cancer Recurrence. *J Am Stat Assoc* (2016) 111(516):1427–39. doi: 10.1080/01621459.2016.1164051
65. Schumacher TN, Schreiber RD. Neoantigens in Cancer Immunotherapy. *Science* (2015) 348(6230):69–74. doi: 10.1126/science.aaa4971
66. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade. *Science* (2017) 357(6349):409–13. doi: 10.1126/science.aan6733
67. Lee C-H, Yelensky R, Jooss K, Chan TA. Update on Tumor Neoantigens and Their Utility: Why It Is Good to Be Different. *Trends Immunol* (2018) 39 (7):536–48. doi: 10.1016/j.it.2018.04.005
68. McGrail DJ, Garnett J, Yin J, Dai H, Shih DJH, Lam TNA, et al. Proteome Instability Is a Therapeutic Vulnerability in Mismatch Repair-Deficient Cancer. *Cancer Cell* (2020) 37(3):371–86.e12. doi: 10.1016/j.ccell.2020.01.011
69. Braun M, Menon R, Nikolov P, Kirsten R, Petersen K, Schilling D, et al. The HOPE Fixation Technique - a Promising Alternative to Common Prostate Cancer Biobanking Approaches. *BMC Cancer* (2011) 11(1):511. doi: 10.1186/1471-2407-11-511
70. Chang S-C, Lan Y-T, Lin P-C, Yang S-H, Lin C-H, Liang W-Y, et al. Patterns of Germline and Somatic Mutations in 16 Genes Associated With Mismatch Repair Function or Containing Tandem Repeat Sequences. *Cancer Medicine* (2020) 9(2):476–86. doi: 10.1002/cam4.2702
71. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep Learning Model for the Prediction of Microsatellite Instability in Colorectal Cancer: A Diagnostic Study. *Lancet Oncol* (2021) 22(1):132–41. doi: 10.1016/S1470-2045(20)30535-0
72. Xu Y, Li C, Lu S, Wang Z, Liu S, Yu X, et al. Design of a Metallacycle-Based Supramolecular Photosensitizer for *In Vivo* Image-Guided Photodynamic Inactivation of Bacteria. *Angewandte Chemie* (2022) 134(5):e202110048. doi: 10.1002/anie.202110048
73. Xu Y, Li C, Lu S, Wang Z, Liu S, Yu X, et al. Construction of Emissive Ruthenium (II) Metallacycle Over 1000 Nm Wavelength for *In Vivo* Biomedical Applications. *Nat Commun* (2022) 13(1):1–13. doi: 10.1038/s41467-022-29572-2

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 30 September 2022
DOI 10.3389/fonc.2022.879376

Predicting IDH subtype of grade 4 astrocytoma and glioblastoma from tumor radiomic patterns extracted from multiparametric magnetic resonance images using a machine learning approach

Pashmina Kandalgaonkar[1,2†], Arpita Sahu[1,2*†], Ann Christy Saju[2,3], Akanksha Joshi[1,2], Abhishek Mahajan[1,2], Meenakshi Thakur[1,2], Ayushi Sahay[2,4], Sridhar Epari[2,4], Shwetabh Sinha[2,3], Archya Dasgupta[2,3], Abhishek Chatterjee[2,3], Prakash Shetty[2,5], Aliasgar Moiyadi[2,5], Jaiprakash Agarwal[2,3], Tejpal Gupta[2,3] and Jayant S. Goda[2,3*]

[1]Department of Radiodiagnosis, Tata Memorial Center, Mumbai, India, [2]Homi Bhabha National Institute, Mumbai, India, [3]Department of Radiation Oncology, Tata Memorial Center, Mumbai, India, [4]Department of Pathology, Tata Memorial Center, Mumbai, India, [5]Department of Neurosurgery, Tata Memorial Center, Mumbai, India

Check for updates

OPEN ACCESS

EDITED BY
Ellen Ackerstaff,
Memorial Sloan Kettering Cancer Center, United States

REVIEWED BY
Marica Eoli,
IRCCS Carlo Besta Neurological Institute Foundation, Italy
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan

*CORRESPONDENCE
Arpita Sahu
drarpitasahu@gmail.com
Jayant S. Goda
jgoda@actrec.gov.in;
godajayantsastri@gmail.com

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

RECEIVED 19 February 2022
ACCEPTED 09 September 2022
PUBLISHED 30 September 2022

CITATION
Kandalgaonkar P, Sahu A, Saju AC, Joshi A, Mahajan A, Thakur M, Sahay A, Epari S, Sinha S, Dasgupta A, Chatterjee A, Shetty P, Moiyadi A, Agarwal J, Gupta T and Goda JS (2022) Predicting IDH subtype of grade 4 astrocytoma and glioblastoma from tumor radiomic patterns extracted from multiparametric magnetic resonance images using a machine learning approach. *Front. Oncol.* 12:879376. doi: 10.3389/fonc.2022.879376

Background and purpose: Semantic imaging features have been used for molecular subclassification of high-grade gliomas. Radiomics-based prediction of molecular subgroups has the potential to strategize and individualize therapy. Using MRI texture features, we propose to distinguish between IDH wild type and IDH mutant type high grade gliomas.

Methods: Between 2013 and 2020, 100 patients were retrospectively analyzed for the radiomics study. Immunohistochemistry of the pathological specimen was used to initially identify patients for the IDH mutant/wild phenotype and was then confirmed by Sanger's sequencing. Image texture analysis was performed on contrast-enhanced T1 (T1C) and T2 weighted (T2W) MR images. Manual segmentation was performed on MR image slices followed by single-slice multiple sampling image augmentation. Both whole tumor multislice segmentation and single-slice multiple sampling approaches were used to arrive at the best model. Radiomic features were extracted, which included first-order features, second-order (GLCM—Grey level co-occurrence matrix), and shape features. Feature enrichment was done using LASSO (Least Absolute Shrinkage and Selection Operator) regression, followed by radiomic classification using Support Vector Machine (SVM) and a 10-fold cross-validation strategy for model development. The area under the Receiver Operator Characteristic (ROC) curve and predictive accuracy were used as diagnostic metrics to evaluate the model to classify IDH mutant and wild-type subgroups.

Results: Multislice analysis resulted in a better model compared to the single-slice multiple-sampling approach. A total of 164 MR-based texture features were extracted, out of which LASSO regression identified 14 distinctive GLCM features for the endpoint, which were used for further model development. The best model was achieved by using combined T1C and T2W MR images using a Quadratic Support Vector Machine Classifier and a 10-fold internal cross-validation approach, which demonstrated a predictive accuracy of 89% with an AUC of 0.89 for each IDH mutant and IDH wild subgroup.

Conclusion: A machine learning classifier of radiomic features extracted from multiparametric MRI images (T1C and T2w) provides important diagnostic information for the non-invasive prediction of the IDH mutant or wild-type phenotype of high-grade gliomas and may have potential use in either escalating or de-escalating adjuvant therapy for gliomas or for using targeted agents in the future.

KEYWORDS

glioblastoma, radiomics, molecular subgroups, machine learning, texture feature

Introduction

High-grade gliomas, especially grade 4 astrocytomas and glioblastomas, are not only the most common primary malignant brain tumors in the adult population but are also associated with intrinsic heterogeneity and invasive properties and are clinically associated with high morbidity and lethality (1). With a better understanding of biology and the advent of newer molecular techniques, researchers have been able to develop unique biomarkers that could predict treatment response and predict these tumors with a high degree of accuracy, paving the way for a more personalized treatment approach. The two molecular biomarkers of significant interest that have translated into clinical practice are Isocitrate Dehydrogenase (IDH) and MGMT (O (2)-methylguanine-DNA methyltransferase), both of which are responsible for epigenetic alterations in glioblastomas. The evaluation of these biomarkers has now become the norm in tailoring therapy and disease prediction.

Glioblastomas, although previously categorized under grade 4 gliomas, are now considered biologically and molecularly distinct entities, namely, glioblastoma IDH-wildtype and IDH-mutant grade 4 astrocytoma, based on 'the present' World Health Organization classification of brain tumors. IDH mutations are identified in approximately 5%–13% of glioblastomas and are associated with a significantly better prognosis, particularly when resection includes the non-enhancing tumor component, which is traditionally left unresected (3). Therefore, it is essential to distinguish the IDH mutation status for planning the most appropriate management strategies, as IDH-mutated tumors have more prolonged overall survival and a higher chance of responding to chemotherapy or radiotherapy (4, 5).

Currently, IDH mutation status is assessed by immunohistochemistry (IHC) or DNA sequencing techniques of the tumor specimen, which is invasive, and given the morphological heterogeneity and invasiveness of high-grade gliomas, the full extent of intratumoral phenotypic/genotypic heterogeneity may not be represented in the tumor specimen. Additionally, the widespread use of these biomarkers remains a challenge due to either a lack of expertise or cost issues associated with their testing. For these reasons, accurate preoperative assessment of the IDH mutation from radiological images is important for prognostic evaluation and optimizing therapy for high-grade gliomas (which in our study are grade 4 astrocytoma, IDH-mutant, and glioblastoma, IDH-wildtype).

Studies have demonstrated that certain quantitative image features, like texture features, can be used to predict both IDH mutations on preoperative imaging of gliomas (6). Tumor radiomics based on texture analysis of MR images represent a quantitative approach in which several individual imaging features that are not easily perceived by the unaided eye are processed using advanced algorithms to reveal measurable indices. Given the inherent tumor heterogeneity in histopathological tissues and the universal availability of MRI, we expected the use of machine learning classifiers of the tumor texture features extracted from multiparametric magnetic resonance imaging (MRI) in a large cohort of GBM patients to subclassify them based on the IDH status as confirmed by immunohistochemistry and/or gene sequencing as the gold standard. The study aimed to explore the accuracy of MR-based tumor radiomics and develop a robust model using a machine learning approach to classify GBM into two distinct molecular subgroups of IDH wild and IDH mutant types in a fairly large cohort of patients.

In this retrospective single-center study, we developed a simple radiomics model using a Support Vector Machine algorithm, based on a minimal set of tumor features obtained using a single and multislice tumor segmentation approach on multiparametric MRI sequences for pretreatment prediction of IDH1 status in high-grade glioma patients.

Materials and methods

Patient population

The study was initiated at a tertiary cancer care center through an institutional intramural grant (Grant no. TRAC/1016/1710/001) after obtaining due approval from the Institutional Ethics Committee (IEC). All histologically confirmed high-grade glioma patients, patients who had complete clinical and pretreatment imaging data in Digital Imaging and Communications in Medicine (DICOM) format, and patients whose IDH status was determined by immunohistochemistry and/or Sanger sequencing were included in the study for radiomic feature extraction, classification, and building the model for sub-classifying the high-grade gliomas based on their IDH status.

Molecular subtyping

IDH mutant or wild phenotype was classified by initial screening using immunohistochemistry (IHC) of the paraffin-embedded tissue followed by DNA sequencing in cases where IHC results were equivocal as per the institutional protocol. The IDH R132H mutation was tested by IHC for all the glial tumors. The antibody used for IDH immunohistochemistry was mouse monoclonal anti-IDH1R132H, clone H09 from Dianova GMBH (Hamburg, Germany). Tumors that stained for IDH antibody were considered positive for IDH mutations, while tumors that did not stain for IDH were subjected to Sanger sequencing, considered the gold standard for detecting IDH mutations. Sanger sequencing for IDH1R132 and IDH2R172 loci was performed by PCR using specific primers from Sigma-Aldrich. On sequencing, other alterations besides the commonest R132H were identified. If sequencing was negative, an absence of IDH mutation was confirmed, and such tumors were deemed IDH wild-type GBM. If IHC was negative and sequencing was positive, such tumors were considered IDH mutant (2).

Radiomics pipeline

A visualization of the steps in the radiomics workflow is depicted in Figure 1. Initially, the brain tumor images were acquired from two different MRI machines (1.5 Tesla Philips^{TM} and 3 Tesla General Electric^{TM}). The DICOM compatible images were imported into the TexRad software^{TM} and reconstructed. The reconstructed images were preprocessed using spatial scaled filters (SSFs) to reduce the background noise and increase the sharpness of the tumor edges. The preprocessed images were used to contour the region of interest (ROI). The segmented images were augmented to increase the number of image data sets. Shape, first order (or histogram), and second order texture (GLCM) features were then extracted from the region of interest. The extracted features were then scaled down using the LASSO regression method. Finally, the data analysis step involved building a model from the selected radiomic features to predict the endpoint of interest (IDH wild vs. IDH mutant high-grade glioma).

Image acquisition protocol

Magnetic resonance imaging sequences of 100 patients were obtained at our institution using Philips Ingenia 1.5T and GE Signa 3T MRI with a pre-fixed standard scanning protocol for brain tumor imaging. Axial T1 contrast (T1C) and T2W images were obtained from the vertex to the skull base, encompassing the whole brain, where the primary tumor is visible in its entirety. These sequences were archived in the institutional Picture Archival and Communication System (PACS) and transferred to the radiomics (texture) analysis system (TexRAD^{TM}). The radiological features on the T2W and contrast-enhanced T1W MR images were evaluated and discerned by an experienced neuro-radiologist, and the texture features were extracted on the TexRad^{TM} console.

MR image preprocessing, segmentation (ROI generation), and augmentation

Magnetic Resonance Imaging of the brain was acquired on two different MRI machines (1.5 Tesla Phillips^{TM} and 3 Tesla General Electric^{TM}). The acquisition details of the MR images for the brain imaging protocol for both machines have been explicitly described in Table 1. The resultant imaging protocol

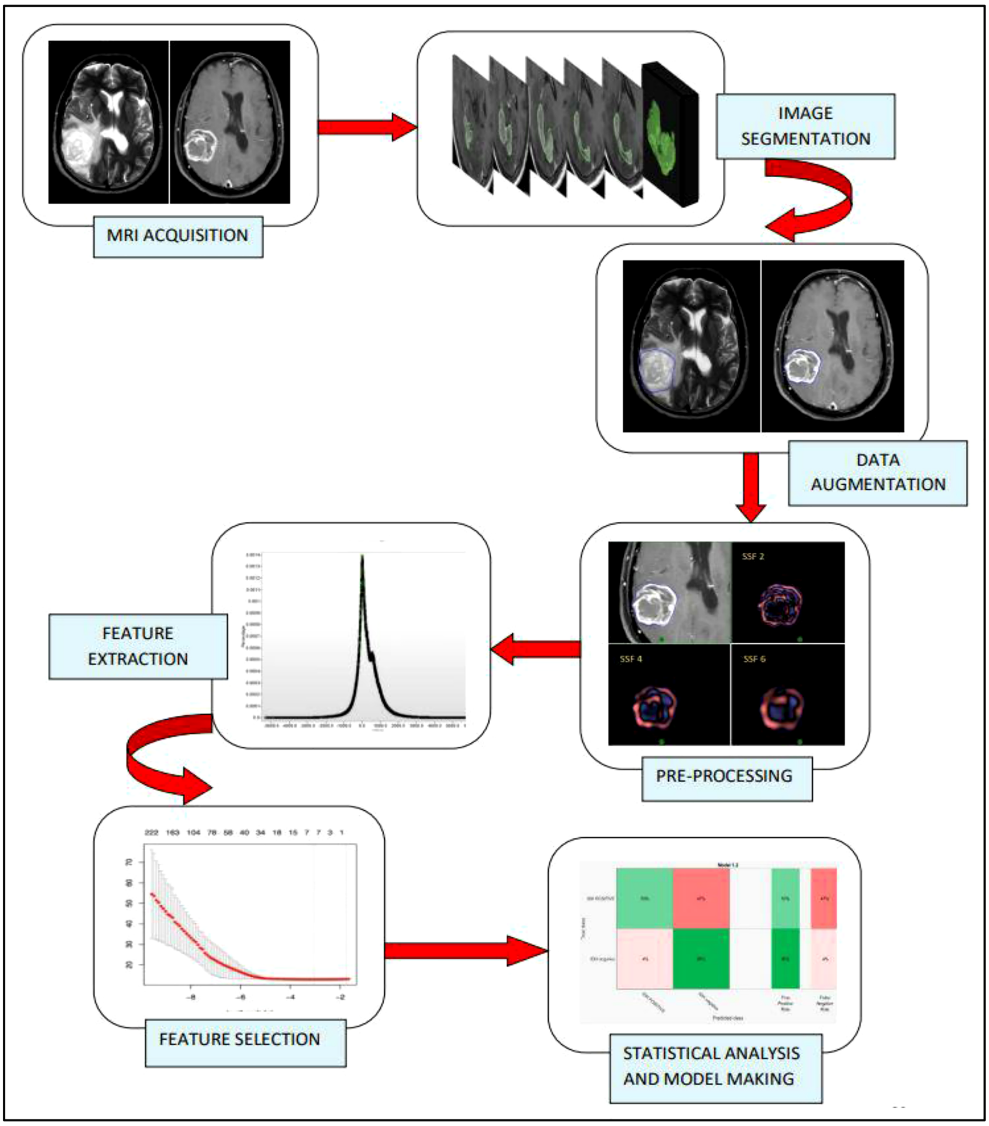


FIGURE 1
Radiomics study flow. The radiomic workflow involves MR brain imaging and data acquisition, followed by slice by slice image segmentation, data augmentation by single slice multiple sampling technique, Image pre-processing by spatial scale filters which involve the use of LoG (Laplacian of Gaussian) bandpass filter, extraction of first order, and second-order features from the texture analysis software, feature selection using LASSO regression and statistical analysis and model development using Support Vector Machine (SVM) and a 10-fold cross-validation strategy.

will result in some imaging heterogeneity. Therefore, before segmentation and ROI delineation, image preprocessing was performed using the Laplacian of Gaussian (LOG) bandpass filters to remove the background noise (Gaussian filter) and enhance the tumor edges (Laplacian filter). This allowed for the extraction of specific structures corresponding to the filter width. Spatial scale Filters (SSF) used filtration values of 0, 2 mm, 3 mm, 4 mm, 5 mm, and 6 mm in width (radius), representing the increasingly coarser level of texture scales for first-order statistics. The use of a filtration algorithm before radiomic feature extraction helps in nullifying some of the effects of heterogeneous acquisition protocols and improves the robustness of the feature selection by removing the features affected by MR noise and imaging heterogeneity.

Tumor segmentation and region of interest (ROI) delineation were performed manually with the freehand drawing function (polygon tool) of the software. The ROI contours and segmentation were separately verified by a neuro-oncologist with 10 years of experience and a neuroradiologist with 10 years of experience. The segmentation was verified by them individually, and any discrepancy was resolved by a consensus. For analysis, the

TABLE 1 MR image acquisition protocol.

MRI Machine	Sequences	FOV (cm)	Matrix	NEX	Slice thickness (mm): Slice gap (mm)
GE Signa 3T	Axial T2	24	320 × 224	1	5:1.5
	Axial T1 + C	24	320 × 190	1	5:1.5
Philips Ingenia 1.5T	Axial T2	23 (AP) 18.5 (RL)	448 × 304	2	5:1
	Axial T1 + C	23 (AP) 18.5 (RL)	232 × 104	2	5:1

FOV, Field-of-view; NEX, Number of excitations; AP, Anteroposterior; RL, Right left.

final contours as verified by the neuroradiologist were considered. Two types of segmentation techniques were used, i.e., whole tumor segmentation (volumetric) as well as single slice with multiple sampling segmentation methods, which in turn were used for data augmentation as described in prior literature (7, 8). A total of 831 Axial T1C and 831 T2 image datasets were obtained for analysis from the study population.

MR texture analysis

The radiomic features were extracted from the segmented images using proprietary texture analysis research software (TexRAD^{TM} Research Version 3.10, TexRAD Ltd, Cambridge, UK), and the machine learning algorithm (SVM) developed a predictive model for molecular sub-classification of high-grade gliomas and was blinded to molecular diagnosis. Eighty-two radiomic features were extracted separately for T1W + C and T2W images using the TexRAD tool, which included 36 first-order features at various SSFs (0, 2, 3, 4, and 6) (Figure 3). Second-order features such as Gray Level Co-occurrence Matrices (GLCM) and topographic features were extracted without applying filters. Twenty GLCM features each for pixel pairs spaced 1 pixel (GLCM1) and 4 pixels apart (GLCM4) respectively, and 6 Shape features (Figure 2**).** The details of all the texture features are provided in Table 2.

Radiomic feature selection

The least absolute shrinkage and selection operator (LASSO) logistic regression algorithm was used for reducing the excessive dimensionality of data and selecting the most significant features in the training data set. Radiomic features with non-zero coefficients were selected from the training data. The analysis was performed using R^{TM} software version 3.6.3, Vienna, Austria, and R Studio^{TM} version 1.2.5033, Boston, USA using the "glmnet" package.

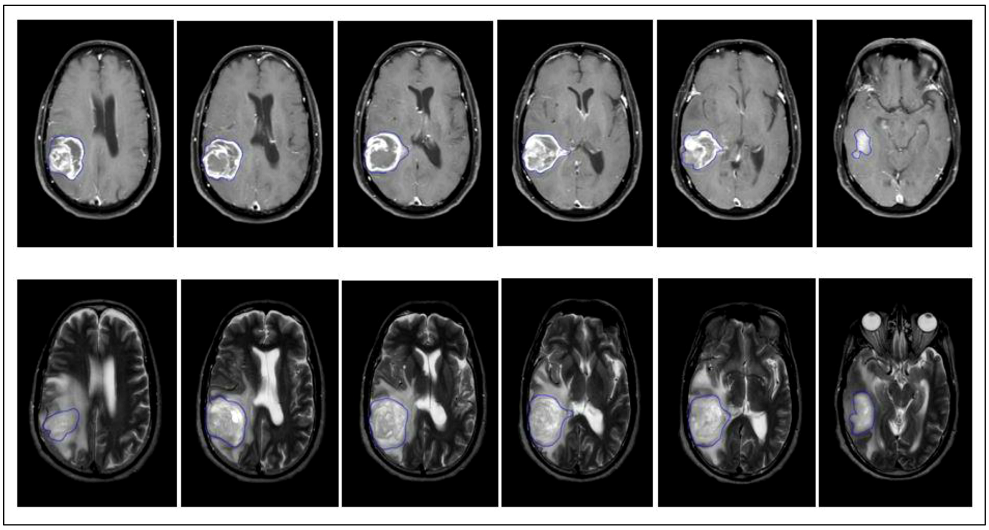


FIGURE 2
Representative multi-slice region of interest (ROI) of an IDH wild-type GBM done on axial T1 + C and T2 MR Images using slice-by-slice image segmentation.

TABLE 2 Demographic, tumor and treatment profile of grade 4 IDH mutant astrocytoma and IDH wild type glioblastoma.

		Overall (Total N = 100)	IDH Wild (N = 83)	IDH Mutant (N = 17)	p-value
BASELINE CHARACTERISTICS					
AGE					
Median age		52 years	54 years	34 years	< 0.001
Range		19–71 years	19–71 years	23–68 years	
IQR		38–59 years	46–59 years	27–43 years	
GENDER					
Male		70	58 (69.9%)	12 (70.6%)	0.954
Female		30	25 (30.1%)	5 (29.4%)	
CENTRICITY					
Unicentric		94	78 (94%)	16 (94%)	0.982
Multicentric		6	5 (6%)	1 (6%)	
LATERALITY					
Right		37	33 (39.8%)	4 (23.5%)	0.450
Left		53	42 (50.6%)	11 (64.%)	
Central/Bilateral		10	8 (9.6%)	2 (11.8%)	
LOCATION					
Cerebellum		2	2 (2.4%)	0 (0%)	0.479
Frontal		31	20 (24.1%)	11 (64.7%)	
Insular		2	2 (2.4%)	0 (0%)	
More than two		32	29 (34.9%)	3 (17.6%)	
Occipital		2	2 (2.4%)	0 (0%)	
Parietal		17	15 (18.1%)	2 (11.8%)	
Temporal		14	13 (15.7%)	1 (5.9%)	
HISTOPATHOLOGY					
MGMT					
Unmethylated		36	32 (48.5%)	4 (33.3%)	0.333
Methylated		42	34 (51.5%)	8 (66.7%)	
Unknown		22			
ATRX					
Retained		73	71 (88.8%)	3 (17.6%)	< 0.001
Lost		15	5 (6.3%)	10 (58.8%)	
Non-contributory		8	4 (5.0%)	4 (23.5%)	
Unknown		3			
Overall					
P53					
Negative		2	2 (2.4%)	0 (0%)	0.518
Positive		98	81 (97.6%)	17 (100%)	
Median Mib 1 index (%)		17.5 (IQR 4%–55.5%)	17.5 (IQR 13.5–22.5)	17.5 (IQR 8–23.75)	0.188
TREATMENT DETAILS					
EXTENT OF SURGERY (n = 99)					
Gross total resection		34	31 (37.8%)	3 (17.6%)	0.271
Near-total resection		26	20 (24.4%)	6 (35.3%)	
Subtotal resection		39	31 (37.8%)	8 (47.1%)	
RADIOTHERAPY					
RT received	Yes	88	72 (86.7%)	16 (94.1%)	0.451
	No	12	11 (13.3%)	1 (5.9%)	
Median RT dose		59.4 Gy, Range (56.5 Gy to 59.4 Gy)	59.4 Gy, Range (56.5 Gy to 59.4 Gy)	59.4 Gy, Range (56.7 Gy to 59.4 Gy)	0.781

(Continued)

TABLE 2 Continued

		Overall (Total N = 100)	IDH Wild (N = 83)	IDH Mutant (N = 17)	p-value
Median RT fractions		33 (IQR 30 to 33 fractions)	33 (IQR 30 to 33 fractions)	33 (IQR 31 to 33 fractions)	0.451
ADJUVANT TMZ (Temozolomide)					
Adj. TMZ Received	Yes	74	60 (72.3%)	14 (82.4%)	0.389
	No	26	23 (27.7%)	3 (17.6%)	
Median cycles of adjuvant TMZ		6 (IQR 4.25–11)	6 (IQR 4–6.50)	11 (IQR 6–12)	0.038

IQR, Inter quartile range; TMZ, Temozolomide; RT, Radiation Therapy; ATRX, Alpha-Thalassemia/Mental Retardation Syndrome, X-Linked; MGMT, O^6-Methylguanine-DNA Methyltransferase.

Radiomic feature classification and modeling

The features selected by LASSO were used as a training set for model development. A Support Vector Machine (SVM) classifier with a 10-fold cross-validation strategy was used in the prediction of the two main molecular subgroups. The performance of the model was assessed using the Area Under Curve (AUC). Multiple models were sequentially evaluated by the system using a combination of selected texture features to arrive at the best model. The SVM analysis was conducted with MATLAB^{TM} version 9.0 (R2016a), The MathWorks, Inc., Natick, MA, USA. Standardization (z-score normalization) was done on the extracted features before SVM analysis as the predictors were of different scales.

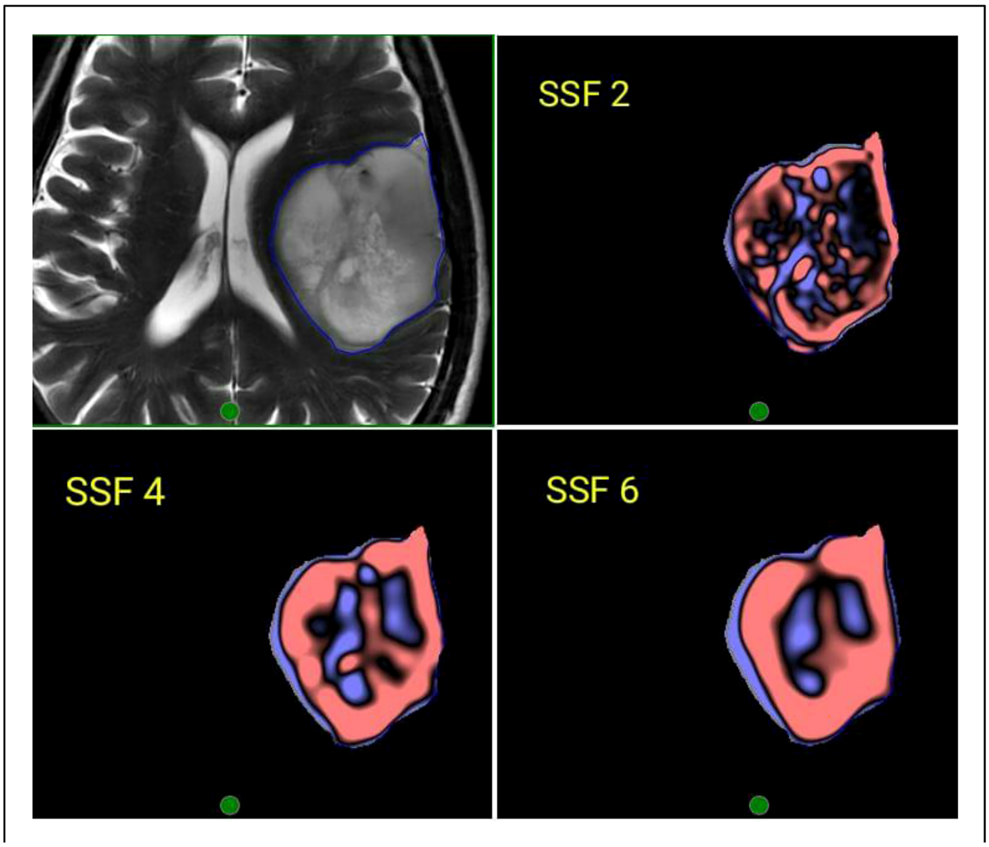


FIGURE 3
Representative image of the region of interest (ROI) contoured on a T2W MRI and corresponding filtered images using Laplacian of Gaussian (LOG) bandpass filtration algorithm showing SSF-2 (fine texture), SSF 4 (Medium texture), and SSF 6 (Coarse texture).

Statistical analysis

Quantitative variables were expressed as mean and/or median. The Student t-test for independent samples was used for the comparison of two different groups. In the case of variables that were not distributed normally, the Mann–Whitney rank sum test was used. The diagnostic accuracy for IDH genotype prediction by textural features was evaluated by analyses of receiver-operating characteristic (ROC) curves using immunohistochemistry/gene sequencing results as the gold standard. The area under the ROC curve (AUC) was evaluated to assess the performance of the developed model. The diagnostic metrics used to assess the model were the AUC, sensitivity, specificity, and overall accuracy as reported in various literature studies investigating Machine Learning-Based Radiomics Signatures for different types of cancers (9–11).

Radiomics quality assurance score and the image biomarker standardization initiative

Imaging data for extracting radiomic features have been used as a tool for testing medical hypotheses. However, the radiomic features extracted from the image data had high dimensionality, requiring complex models to predict or correlate with the endpoints of interest. This limits its usage for only research purposes without real-world application in the clinics and guides the clinical decision-making process, resulting in a huge translational gap. Therefore, Lambin et al. developed a standardized radiomic quality assurance score (RQS) for evaluating the performance, reproducibility, and/or clinical utility of radiomic biomarkers. The RQS is a reporting system of metrics used to validate the robustness of radiomic studies (12). The RQS comprises 16 components, as represented in Supplementary Table 1.

Apart from the RQS, our study tried to adhere to the Image Biomarker Standardization Initiative (IBSI) guidelines which were initiated to address the challenges in utilizing radiomics as an image-based biomarker (13) For this study, we evaluated all the processing steps from image processing, segmentation, and ROI delineation to the computation of radiomic features were evaluated in this study (Supplementary Table 2**)**.

Results

Baseline demographics and tumor and treatment characteristics of the study cohort

One hundred and thirty-three patients with a histological diagnosis of high-grade gliomas (CNS WHO grade 4 of adult type diffuse gliomas) were screened for the radiomic study. Based on the inclusion criteria, only a hundred patients were eligible for the study. Seventeen patients had IDH mutations and 83 patients had IDH wild-type glioblastoma. The median age of patients at presentation was 52 years (a range of 19 to 71 years) and the majority of them were males (70%), The demographic details of the study population are presented in detail in Table 2. All but one patient underwent maximal safe resection of the tumor, whereas one patient underwent only biopsy, followed by risk-based adjuvant therapy incorporating both radiotherapy and chemotherapy as deemed appropriate after discussion in a joint multidisciplinary clinic Table 3.

Molecular subgrouping

Of the 100 patients who were studied, IHC for **IDH1R132H** was done on all the cases. IDH1/2 sequencing was performed on cases that were deemed negative on IHC for **IDH1R132H** but showed loss of expression for ATRX. The cases which were negative for **IDH1R132H** on IHC and showed retained expression of ATRX were taken as IDH wild type (14). A total of 13 patients (13%) were positive for **IDH1R132H** on IHC.

TABLE 3 Radiomic features extracted.

Texture Features Used	
1st Order Features	Mean Standard Deviation Mean of Positive pixels Entropy Skewness Kurtosis
GLCM features	Autocorrelation Cluster prominence Cluster shade Cluster tendency Contrast Correlation Dissimilarity Homogeneity Joint average Joint energy Joint entropy Idm (inverse difference moment) Diffentropy Diffvariance Idmn (inverse difference moment normalized) Idn (inverse difference normalized) Inverse variance Sum entropy Sum squares Join tmax
Shape Features	Perimeter Area Elongation Sphericity Long axis Short axis

Eighty-seven patients (87%) were negative for IDH1R132H on immuno-histochemistry. Among the 87 patients, six showed loss of expression of ATRX and underwent Sanger sequencing for confirmation of IDH status. Of these six patients, four showed IDH mutations on Sanger sequencing: two patients were positive for **IDH1R132C** only, while one patient had an **IDH1R132L** mutation and another patient showed an **IDH1R132H** mutation. Two of the six patients showed no point mutation for IDH1 or IDH2 and were considered IDH-wild type. Therefore, of the 100 patients, 17 patients were considered IDH mutant subtype, while 83 patients were IDH wildtype.

Performance of the binary classification model

Out of a total of 82 texture features each in T1W + C and T2W images, LASSO regression for feature selection elucidated seven discriminant features for T1W + C images and seven discriminant features for T2W images, which were used for further model development.

A combination of LASSO selected first order texture features, second order (GLCM) features, and topographic features were used to create different models using both T1W + C and T2W images in an attempt to arrive at the best SVM model Table 5. Among various models evaluated, a combination of 14 GLCM features from combined T1W + C and T2W images resulted in the best classifier, as depicted in Table 4. The model based on a Combined Multi-slice Texture Analysis of T1 + C and T2 weighted MR imaging using a Quadratic Support Vector Machine Classifier and a 10-fold internal cross-validation approach, resulted in the best performance in predicting the molecular subtypes with a predictive accuracy of 89% and a Receiver Operator Characteristic (ROC) analysis demonstrating an AUC of 0.89 for each IDH positive and IDH negative subtype (Figure 4)**.** Of the 83 IDH negative cases, 80 tumors were true positive while three tumors were false negative, resulting in a very high sensitivity of 96%, but at the same time, the model specificity was 52.9%. This low specificity is due to the unbalanced classification of IDH subtypes. Similarly, for 17 IDH positive cases, nine tumors were true positives while eight tumors were false negatives, resulting in a sensitivity of only 53% but a high specificity of 96.4% as depicted in the confusion matrix (Figure 5), Table 6.

Discussion

We developed a Support Vector Machine (SVM) based classification model with satisfactory performance to probe the genomic profile (IDH mutant vs. IDH wild type) of grade 4 adult diffuse gliomas, based on MR image phenotypes. The SVM classifier had an overall accuracy of 89% for predicting IDH wild-type tumors from IDH mutants. Our results suggest the use of multiparametric MR radiomics along with machine-learning models to classify the molecular subtype of grade 4 adult type diffuse gliomas consistent with the new 2021 WHO classification. By employing a specific ML classifier, several clinical applications for the detection of IDH status in high-grade gliomas can be achieved with or without histopathology of the tumor specimen.

IDH mutations are considered to be an early event in gliomagenesis and are one of the most critical genetic biomarkers for high-grade gliomas having prognostic implications (improved survival with IDH mutant than wild-type glioblastomas {31 months vs 15 months}) (15). Additionally, IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype, which is a powerful determinant of tumor pathogenicity (16). Therefore, having a preoperative assessment of IDH gene mutation status in glioma may help in optimizing glioma therapeutics. While immunohistochemistry is considered a routine screening method for detecting IDH mutations in the majority of cases, Sanger sequencing is considered to be a confirmatory test for identifying IDH mutations. However, high-grade gliomas, especially glioblastomas, show marked intratumoral heterogeneity in IDH status. Pathological tissue biopsies from the different parts of tumors may yield varied results regarding the IDH status as these high-grade gliomas are considered to be heterogeneous. Therefore, a non-invasive method like magnetic resonance imaging could be put to effective use for objectively quantifying structural heterogeneity within the tumor using image-based radiomic analysis. Radiomics is a novel approach for the high-throughput extraction of quantitative image features from a specified ROI (17). These quantitative features (radiomic features) have been successfully used to develop models using sophisticated machine learning algorithms for identifying image biomarkers with the capability to predict the genotype of a tumor (18). Published studies have leveraged machine learning classifiers to develop radiomic signatures to predict IDH mutation status in gliomas (11, 19, 20). Within the framework of radiomics, tumor texture features as extracted from MR images of brain tumors are predefined and quantitative features are derived by computational methods that describe the spatial variations in the

TABLE 4 A combination of LASSO selected features that resulted in the best classification model.

T1W + C TEXTURE FEATURES (N = 7)	T2W TEXTURE FEATURES (N = 7)
KURTOSIS_0_T1C	MEAN_0_T2
ENTROPY_2_T1C	MPP_0_T2
KURTOSIS_2_T1C	KURTOSIS_0_T2
MEAN_5_T1C	MEAN_4_T2
KURTOSIS_5_T1C	GLCM1_clusterShade_T2
SKEWNESS_6_T1C	GLCM1_idn_T2
GLCM4_correlation_T1C	GLCM1_sumEntropy_T2

GLCM 1, GLCM features of pair of pixels which are 1 pixel apart; **GLCM 4**, GLCM features of a pair of pixels which are 4 pixels apart; **T1C**, Contrast-enhanced T1 weighted images; **idn**, inverse difference normalized.

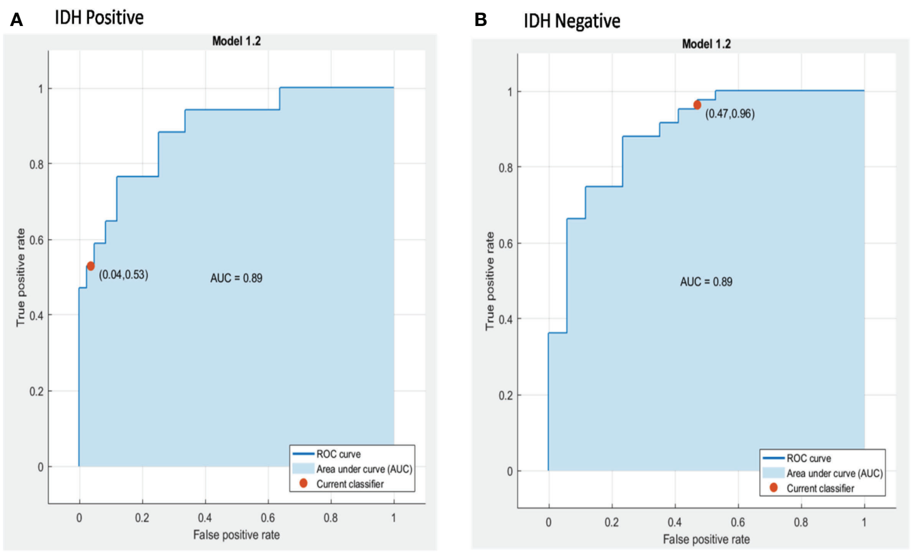


FIGURE 4
ROC curves of the best model for prediction of the two molecular subgroups using combined multi-slice T1 + C and T2w GLCM features using Quadratic SVM, **(A)** IDH positive and **(B)** IDH negative.

intensity of the images along the entire cross-section of the tumor that is beyond visual perception. These features have the potential to yield additional information not only about the tumor biology but also about the genomic profile. Thus, they allow the prediction of the IDH genotype in glioma patients with a high degree of accuracy (21). The present study was done to investigate the feasibility of using machine learning-based radiomic signatures to predict the IDH subtype in high-grade gliomas in a high throughput setting.

Radiomics-based machine learning tools or deep learning tools have been used for subclassifying various grades of gliomas

TABLE 5 Showing the molecular classification (IDH mutant and IDH wild type) of grade-IV GBM modeled by using Support Vector Machine as the radiomics classifier on MRI-based sequences.

ImageSingle slice v/s Multi slice	MRI sequence	IDHClassification	Radiomics classifier	Diagnostic Metrics		Validation Process	
				AUC	Accuracy	10-fold internal cross-validation	Hold Validation
Single slice analysis	T1C	IDH –VE (694)	Linear SVM	0.91	89.8%	YES	NO
		IDH +VE (137)		0.91			
	T2W	IDH –VE (689)	Cubic SVM	0.84	86.9%	YES	NO
		IDH +VE (149)		0.84			
Multi-slice analysis	T1C	IDH –VE (83)	Linear SVM	0.87	87%	YES	NO
		IDH +VE (17)		0.87			
	T2W	IDH –VE (83)	Quadratic SVM	0.80	91%	YES	NO
		IDH +VE (17)		0.80			
	T1C + T2W	IDH –VE (83)	Quadratic SVM	0.89	89%	YES	NO
		IDH +VE (17)		0.89			
	T1C + T2W	IDH –VE (83)	Cubic SVM	0.81	90%	NO	YES (90:10)
		IDH +VE (17)		0.81			

AUC, Area under the curve; SVM, Support Vector Machine.

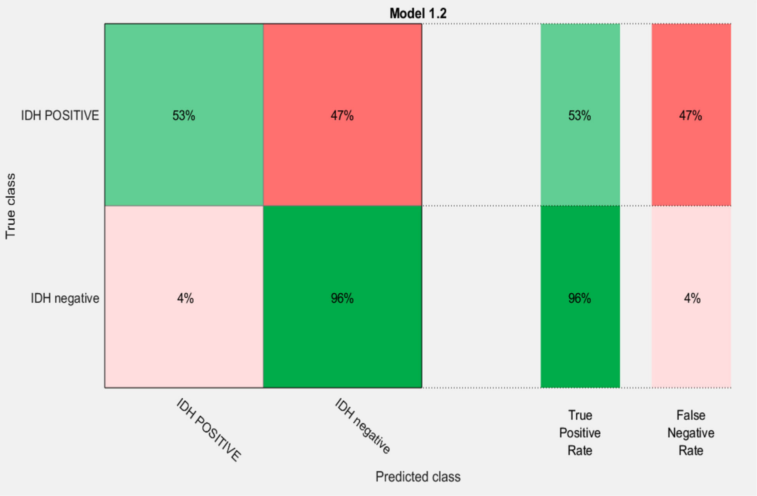


FIGURE 5
Confusion matrix of the best model for prediction of the two molecular subgroups using GLCM features of combined multi-slice T1 + C and T2w images using Quadratic SVM.

into IDH wild-type or mutant-type entities (22). However, the literature on this subject is quite sparse (Table 7). A Taiwanese group used radiomic features consisting of morphological, intensity, and textural features to develop a prediction model for IDH mutation (26) and textural features yielded the best accuracy of 85%. Going further, the group used the same set of patients to interpret the status of IDH status in glioblastomas from transformed magnetic resonance imaging patterns (26). By ranklet transformation of collected images from 39 patients (32 IDH wild and seven IDH mutant cases), three feature sets were extracted, with each feature set having 14 GLCM textural features. They achieved an accuracy of 90%, a sensitivity of 57%, and a specificity of 97%. In contrast to the Taiwanese group, our study used both axial T2 and axial post-contrast T1 + C images, and unlike the largest single slice that was used in this study, we incorporated tumor contours on each axial slice of both the sequences wherever the tumor was present. This took into account the heterogeneity present within the entire tumor volume, which has an advantage over core biopsy methods, which target only a limited section of the tumor for histopathology.

Comparative models studying the predictive abilities of radiomic features have been rarely performed in the literature. A multicentric study compared various machine learning classifiers to predict the genetics of GBM on different MRI sequences. This study was done on 156 adult patients with a pathologic diagnosis of GBM. Radiomic features were extracted using various extraction tools like NET, CET, and NEC with a custom version of Pyradiomics and selected through the Boruta algorithm. The investigators used various radiomic classifiers like AdaBoost (AB), Extreme Gradient Boosting (xGB), Gradient Boosting (GB), Decision Tree (DT), and Random Forest (RF), Logistic Regressor (LR), two stacking classifiers (ST, ST_ABC), and K Neighbors (KN). It is used to classify IDH mutants from the IDH wild subtype of GBM. Based on the results, the AB classifier performed the best, with a reported accuracy for classifying the IDH phenotype. (overall accuracy of 89% and ROC-AUC of 87.7%) (27). The SVM classifier we used to predict the IDH subtype performed relatively well (ROC-AUC of 89% and overall accuracy of 89%, similar to the above study) (27).

Isocitrate dehydrogenase (IDH) mutations are quite common in low-grade gliomas, unlike in higher grade gliomas. Machine learning-based radiomic feature modeling has been

TABLE 6 Performance of best classification model.

Diagnostic metrics	IDH –VE (n = 83)	IDH +VE (n = 17)
AUC	**0.89**	**0.89**
TP	80	9
TN	9	80
FP	8	3
FN	3	8
Sensitivity	96%	53%
Specificity	52.9%	96.4%
FNR	4%	47%
PPV	90.9%	75%
NPV	75%	90.9%
Overall Accuracy	**89%**	

AUC, Area under the curve;
TP, True positive; TN, True negative; FP, False positive; FN, False negative; FNR, False negative rate; PPV, Positive predictive value; NPV, Negative predictive value.

tried in various grades of gliomas (28). Sakai et al. in a heterogeneous cohort of gliomas [n = 100 (grade-I I = 11; grade-3 = 8 and grade IV: 81)] used MRI-based radiomic features to predict IDH1 Mutation Status in Gliomas using a gradient tree boost machine learning classifier. The best performance was seen with a DWI-trained XG Boost model, which achieved ROC with an Area Under the Curve (AUC) of 0.97, an accuracy of 0.90 on the test set. They used the same machine learning classifier (XG boost) on the FLAIR-MR images used as a test set and achieved a ROC with an AUC of 0.95 and an accuracy of 0.90. Their results showed that the model that was trained on combined FLAIR-DWI radiomic features did not provide an increment in terms of accuracy. Using multiparametric radiomic features derived from preoperative MRI can predict IDH1 mutation status with approximately 90% accuracy (28).

Although a single institutional study, the radiomic analysis and model development were done on a relatively small sample size. In our study, we used two approaches to analyze the texture data: a volumetric approach and a single slice multiple sampling approach. Analysis was done using a Support Vector Machine classifier based on features selected by LASSO regression, which selected the best of all the features. Support Vector Machine utilizes the concept of a hyperplane, which is a plane that has the maximum margin, and considers the furthest of the points falling on either side of the hyperplane and is less vulnerable to overfitting as compared to other simple classifiers like logistic regression. Moreover, outliers have less impact on the SVM as opposed to other machine learning algorithms, especially when in higher dimensional data. Various classifier models were used and validation was done using 10-fold internal cross-validation as well as hold-out validation at ratios of 9:1, 8:2, and 7:3, and the latter yielded suboptimal results due to a lack of adequate sample size. The texture features analyzed included first-order and GLCM features. To overcome the limitation of the small sample size, an augmentation strategy called the single slice multiple sampling approach was evaluated. This approach enabled us to reduce the potential overfitting of data, which is known to happen in machine learning approaches, and this approach also yielded appreciable results. Although the SVM classifier has several advantages that have been elucidated, it does have some limitations and uncertainties when it comes to building models for very large data sets. Moreover, the algorithm does not perform well for datasets where target classes are overlapping. It also underperforms in situations where the number of radiomic features for each data point exceeds the number of training data samples. The SVM will underperform in these situations.

Our study was a single institutional study with a quality-controlled central pathological laboratory and uniform radiology and radiomic review. One of the strengths of the study was that all the image delineation was verified by an experienced neuro-oncologist with 10 years of experience, blinded to the results of the molecular subgrouping. Being a tertiary cancer institute, it

TABLE 7 Literature review of studies using radiomics and or semantic features for glioblastoma molecular subgroup classification using various diagnostic metrics.

No.	Author (year) No. of patients	MRI sequences	Model used for subgroup classification		AUC	Sensitivity	Specificity	PPV
3	Hsieh et al. (23) (2020), (n = 39)	Feature-based with use of ranklet transformation on axial T1 + C MR images	KNN and SVM	Test Cohort	–	0.57	–	–
	Pasquini et al. (24) (2021), (n = 100)	Featureless radiomics on MPRAGE, FLAIR, T1W, T2W, DWI with ADC, PWI) with DSC sequence	4 block 2D-CNN architecture	Training and test (80:20) set.	0.86 ± 0.05, the highest achieved using rCBV maps	0.76 ± 0.05	–	–
2	Calabrese et al. (25) (2020), (n = 199)	Fully automated deep learning-based tumor segmentations using T1W, T2W, T2W/FLAIR, DWI, SWI, HARDI fractional anisotropy (HARDI FA), ASL, and T1C.	Automated dCNN segmentation	10-fold stratified shuffle split cross-validation strategy with a train/test split of 60:40	0.95 ± 0.03	0.93 ± 0.08	–	–
4	Pashmina et al. (Present study) (n = 100)	Feature-based radiomics using axial T1 + C and T2W MR images	LASSO regression and SVM	10-fold internal cross-validation	0.89	0.96 for IDH wild, 0.80 for IDH mutant	0.53 for IDH wild, 0.03 for IDH mutant	0.91 for IDH wild, 0.75 for IDH mutant

SVM, Support Vector Machine; LASSO, Least absolute shrinkage and selection operator; CNN, Convolutional neural network; IDH, Isocitrate dehydrogenase; ADC, apparent Diffusion Coefficient; DSC, Dynamic Susceptibility Contrast; PWI, perfusion-weighted images.

catered to a large and diverse pool of patients. The use of the single slice multiple sampling methods in this study not only helped in data augmentation but also prevented data loss. The main presumed weakness lies in the heterogeneity of MRI acquisition parameters in the study population and the fact that uniformity in image acquisition is necessary for radiomic analysis was acknowledged (29). Regardless of the heterogeneity in MR acquisition parameters, we were able to achieve a fair bit of accuracy, suggesting that this would consequently have a good implication if validated in a large cohort of patients in real-world clinical practice. Additionally, the current methodology of using internal cross-validation has the limitation of inflating the performance metrics. However, with a limited sample size, we thought that the internal 10-fold cross-validation would be the best strategy to utilize for model development. We are accruing more patients to evaluate the model on an external dataset, and this will be done in future studies.

In addition to radiomics features, our study did not include semantic features as those established by "The Visually AcceSAble Rembrandt Images" (VASARI) project could have potentially improved the performance of the model. Next, the study was limited by its small sample size with a skewed distribution of the various molecular subgroups. The relatively small sample size of our study also limited the use of deep learning algorithms, such as convolutional neural network (CNN) analysis, which requires a massive number of image datasets, which would not have been possible without the pooling of image data from multiple institutions, which in itself could have introduced a confounding factor of image heterogeneity, resulting in variability and generalization gaps in the predictive model. Although we did 10-fold internal cross-validation, the lack of an external validation cohort limits its robustness. These create future opportunities to incorporate clinical parameters and semantics features to complement the radiomic signatures to develop a more robust predictive model with better diagnostic metrics to classify the molecular subgroups of glioblastoma. The model developed in the current study is planned to be tested on an independent validation cohort and subsequently on a larger imaging dataset.

Conclusion

The results of the study affirm that a texture feature-based radiomic model of multiparametric MR images can effectively classify molecular subgroups of GBM with an acceptable degree of accuracy using a machine learning approach. The proposed image-based radiomic approach provides an alternative non-invasive and efficient method to sub-classify the molecular subgroup and can aid in optimizing the adjuvant therapy of glioblastomas. Given that radiogenomics is rapidly evolving, machine learning approaches combined with clinical and radiological semantic (VASARI) features may show superior outcomes. The field of radiomics needs to be further researched to translate findings into an interpretable format for presurgical prediction of the molecular genotype of GBM.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

This study was reviewed and approved by the Institutional Ethics Committee,Tata Memorial Centre. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

PK, JSG, and AS contributed to the conception and design of the study. PK, ACS organized the database. PK and ACS performed the statistical analysis. PK, ACS, AS, AM, wrote the first draft of the manuscript. SE and AyS wrote sections of the manuscript and did the immunohistochemistry and Sanger sequencing to classify IDH mutant vs wild type. All authors contributed to the manuscript revision, read and approved the submitted version. AS, JPA and JSG supervised the entire study.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.879376/full#supplementary-material

SUPPLEMENTARY TABLE 1
Shows the components of the Radiomic Quality score (RQS) with the total score of our study.

SUPPLEMENTARY TABLE 2
Shows the IBSI reporting of the study.

References

1. Thakkar JP, Dolecek TA, Horbinski C, Ostrom QT, Lightner DD, Barnholtz-Sloan JS, et al. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol Biomarkers Prev* (2014) 23(10):1985–96. doi: 10.1158/1055-9965.EPI-14-0275

2. Epari S, Gurav M. P03.03 IDH1/2 mutations in gliomas: A single tertiary cancer institutional experience. *Neuro Oncol* (2017) 19(Suppl 3):iii33. doi: 10.1093/neuonc

3. Houillier C, Wang X, Kaloshi G, Mokhtari K, Guillevin R, Laffaire J, et al. IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. *Neurology* (2010) 75(17):1560–6. doi: 10.1212/WNL.0b013e3181f96282

4. Tamimi AF, Juweid M. Epidemiology and outcome of glioblastoma. In: De Vleeschouwer S, editor. *Glioblastoma*, vol. 8 . Brisbane (AU: Codon Publications (2017).

5. SongTao Q, Lei Y, Si G, YanQing D, HuiXia H, XueLin Z, et al. IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci* (2012) 103(2):269–73. doi: 10.1111/j.1349-7006.2011.02134.x.Epub2011Nov28

6. Zhang B, Chang K, Ramkissoon S, Tanguturi S, Bi WL, Reardon DA, et al. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro Oncol* (2017) 19(1):109–17. doi: 10.1093/neuonc/now121

7. Kocak B, Durmaz ES, Ates E, Ulusan MB. Radiogenomics in clear cell renal cell carcinoma: Machine learning–based high-dimensional quantitative CT texture analysis in predicting PBRM1 mutation status. *Am J Radiol* (2019) 212:55–63. doi: 10.2214/AJR.18.20443

8. Kocak B, Durmaz ES, Erdim C, Ates E, Kaya OK, Kilickesmez O. Radiomics of renal masses: Systematic review of reproducibility and validation strategies. *Am J Roentgenology* (2020) 214(1):129–36. doi: 10.2214/AJR.19.21709

9. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-Small-Cell lung cancer. *Int J Mol Sci* (2021) 22(17):9254. doi: 10.3390/ijms22179254

10. Jeong J, Wang L, Ji B, Lei Y, Ali A, Liu T, et al. Machine-learning based classification of glioblastoma using delta-radiomic features derived from dynamic susceptibility contrast enhanced magnetic resonance images: Introduction. *Quant Imaging Med Surg* (2019) 9(7):1201–13. doi: 10.21037/qims.2019.07.01

11. Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, et al. Machine learning-based radiomics for molecular subtyping of gliomas. *Clin Cancer Res* (2018) 24(18):4429–36. doi: 10.1158/1078-0432.CCR-17-3445

12. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (.2017) 14(12):749–62. doi: 10.1038/nrclinonc.2017.141

13. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* (2020) 295(2):328–38. doi: 10.1148/radiol.2020191145

14. Santosh V, Sravya P, Gupta T, Muzumdar D, Chacko G, Suri V, et al. ISNO consensus guidelines for practical adaptation of the WHO 2016 classification of adult diffuse gliomas. *Neurol India* (2019) 67(1):173–82. doi: 10.4103/0028-3886.253572

15. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, et al. IDH1 and IDH2 mutations in gliomas. *N Engl J Med* (2009) 360(8):765–73. doi: 10.1056/NEJMoa0808710

16. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* (2012) 483(7390):479–83. doi: 10.1038/nature10866

17. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology* (2016) 278(2):563–77. doi: 10.1148/radiol.2015151169

18. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin* (2019) 69(2):127–57. doi: 10.3322/caac.21552

19. Choi YS, Bae S, Chang JH, Kang SG, Kim SH, Kim J, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas *via* deep learning and radiomics. *Neuro Oncol* (2021) 23(2):304–13. doi: 10.1093/neuonc/noaa177

20. Zhou H, Chang K, Bai HX, Xiao B, Su C, Bi WL, et al. Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low- and high-grade gliomas. *J Neurooncol* (2019) 142(2):299–307. doi: 10.1007/s11060-019-03096-0

21. Li Z-C, Bai H, Sun Q, Zhao Y, Lv Y, Zhou J, et al. Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma. *Cancer Med* (2018) 7(12):5999–6009. doi: 10.1002/cam4.1863

22. Manikis GC, Ioannidis GS, Siakallis L, Nikiforaki K, Iv M, Vozlic D, et al. Multicenter DSC-MRI-Based radiomics predict IDH mutation in gliomas. *Cancers (Basel)* (2021) 13(16):3965. doi: 10.3390/cancers13163965

23. Hsieh KL, CY C, Lo CM. Radiomic model for predicting mutations in the isocitrate dehydrogenase gene in glioblastomas. *Oncotarget* (2017) 8(28):45888–97. doi: 10.18632/oncotarget.17585

24. Pasquini L, Napolitano A, Tagliente E, Tagliente E, Dellepiane F, Lucignani M, Vidiri A, et al. Deep learning can differentiate IDH-mutant from IDH-wild GBM. *J Pers Med* (2021) 11(4):290. doi: 10.3390/jpm11040290

25. Calabrese E, Villanueva-Meyer JE. Cha S.A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. *Sci Rep* (2020) 10(1):11852. doi: 10.1038/s41598-020-68857-8

26. Lo CM, Weng RC, Cheng SJ, Wang HJ, Hsieh KL. Computer-aided diagnosis of isocitrate dehydrogenase genotypes in glioblastomas from radiomic patterns. *Med (Baltimore)* (2020) 99(8):e19123. doi: 10.1097/MD.0000000000019123

27. Pasquini L, Napolitano A, Lucignani M, Tagliente E, Dellepiane F, Rossi-Espagnet MC, et al. Comparison of machine learning classifiers to predict patient survival and genetics of GBM: Towards a standardized model for clinical implementation. *ArXiv* doi: 10.48550/arXiv.2102.06526

28. Sakai Y, Yang C, Kihira S, Tsankova N, Khan F, Hormigo A, et al. MRI Radiomic features to predict IDH1 mutation status in gliomas: A machine learning approach using gradient tree boosting. *Int J Mol Sci* 21(21):8004. doi: 10.3390/ijms21218004

29. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magnetic Resonance Imaging* (2004) 22(1):81–91. doi: 10.1016/j.mri.2003.09.001

frontiers | Frontiers in Oncology

TYPE Systematic Review
PUBLISHED 03 October 2022
DOI 10.3389/fonc.2022.884173

Classifying primary central nervous system lymphoma from glioblastoma using deep learning and radiomics based machine learning approach - a systematic review and meta-analysis

Amrita Guha[1]*, Jayant S. Goda[1]*, Archya Dasgupta[2], Abhishek Mahajan[1], Soutik Halder[3], Jeetendra Gawde[3] and Sanjay Talole[3]

[1]Department of Radio Diagnosis, Tata Memorial Centre, Homi Bhaba National Institute, Mumbai, India, [2]Department of Radiation Oncology, Tata Memorial Centre, Homi Bhaba National Institute, Mumbai, India, [3]Department of Biostatistics, Tata Memorial Centre, Homi Bhaba National Institute, Mumbai, India

OPEN ACCESS

EDITED BY
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan

REVIEWED BY
Yuanbo Liu,
Capital Medical University, China
Yasuo Iwadate,
Chiba University, Japan

*CORRESPONDENCE
Amrita Guha
amritaguha85@gmail.com
Jayant S. Goda
jgoda@actrec.gov.in

SPECIALTY SECTION
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

RECEIVED 25 February 2022
ACCEPTED 07 September 2022
PUBLISHED 03 October 2022

CITATION
Guha A, Goda JS, Dasgupta A, Mahajan A, Halder S, Gawde J and Talole S (2022) Classifying primary central nervous system lymphoma from glioblastoma using deep learning and radiomics based machine learning approach - a systematic review and meta-analysis.
Front. Oncol. 12:884173.
doi: 10.3389/fonc.2022.884173

Background: Glioblastoma (GBM) and primary central nervous system lymphoma (PCNSL) are common in elderly yet difficult to differentiate on MRI. Their management and prognosis are quite different. Recent surge of interest in predictive analytics, using machine learning (ML) from radiomic features and deep learning (DL) for diagnosing, predicting response and prognosticating disease has evinced interest among radiologists and clinicians. The objective of this systematic review and meta-analysis was to evaluate the deep learning & ML algorithms in classifying PCNSL from GBM.

Methods: The authors performed a systematic review of the literature from MEDLINE, EMBASE and the Cochrane central trials register for the search strategy in accordance with PRISMA guidelines to select and evaluate studies that included themes of ML, DL, AI, GBM, PCNSL. All studies reporting on ML algorithms or DL that for differentiating PCNSL from GBM on MR imaging were included. These studies were further narrowed down to focus on works published between 2018 and 2021. Two researchers independently conducted the literature screening, database extraction and risk bias assessment. The extracted data was synthesised and analysed by forest plots. Outcomes assessed were test characteristics such as accuracy, sensitivity, specificity and balanced accuracy.

Results: Ten articles meeting the eligibility criteria were identified addressing use of ML and DL in training and validation classifiers to distinguish PCNSL from GBM on MR imaging. The total sample size was 1311 in the included studies. ML approach was used in 6 studies while DL in 4 studies. The lowest reported sensitivity was 80%, while the highest reported sensitivity was 99% in studies in which ML and DL was directly compared with the gold standard histopathology. The lowest reported specificity was 87% while the highest reported specificity was 100%. The highest reported balanced accuracy was 100% and the lowest was 84%.

Conclusions: Extensive search of the database revealed a limited number of studies that have applied ML or DL to differentiate PCNSL from GBM. Of the currently published studies, Both DL & ML algorithms have demonstrated encouraging results and certainly have the potential to aid neurooncologists in taking preoperative decisions in the future leading to not only reduction in morbidities but also be cost effective.

KEYWORDS

machine learning, deep learning, predictive analytics, primary central nervous system (CNS) lymphoma, glioblastoma, magnetic resonance imaging, metaanalysis, systematic review

Introduction

Primary Central Nervous System Lymphomas (PCNSL) and Glioblastomas(GBM) are tumours of the adults and elderly, however, they are distinct entities in terms of their cell of origin, incidence, natural history, treatment protocols and prognosis (1). Even though these tumours are different, they appear radiologically appear similar on Magnetic resonance imaging(MRI) with only a few discerning features (2). Although there are a few semantic MR imaging features that help the radiologist to differentiate PCNSL from GBM (3), these features are subjective and dependant on the expertise and the experience of the radiologist with a resultant dependence on the gold standard histopathology of the tumour specimen (4). Certain special MRI sequences such as Diffusion Weighted Imaging (DWI), MR spectroscopy (MRS) may complement the semantic features (5), and could be useful in differentiating the two tumours but these special MR protocols are resource intense and their use is limited due to lack of widespread availability and associated cost escalations have practise implications in high throughput cancer centres.

Tumor radiomics based on texture feature analysis of MR images represents an abstract mathematical quantitative approach whereby multiple individual imaging features not easily discerned by the naked eye are processed by means of sophisticated algorithms to reveal quantifiable indices (6). Radiomics maximizes the number of quantitative image features from digital images and as a result, can overcome intratumoral heterogeneities in both the molecular and histopathological assessment of various tumour histologies using measurable values that contribute to tumor diagnosis, pre-surgical grading, response to treatment, prognostication of cancers and predicting gene mutation. Moreover, with quantified analyses of images, it has also been incorporated with various novel computer technologies, such as machine learning and deep learning algorithms like deep convolutional neural network (dCNN) (7–15)

Even before deep learning methods were available, majority of ML based radiology studies used texture features extracted from manually segmented tumour images followed by application of conventional ML tools such as random forests and support vector machines (15–17) The advent of advanced computational methods like deep learning algorithms brought a paradigm shift in the image based classification of tumours and their biology (18). The development of the convolutional neural network (CNN), that comprises of convolution and pooling layers, has led to automation in identifying relevant image features for various classification tasks (19).

Although, various ML tools like random forest or support vector machine models and DL algorithms like CNN have been used to classify PCNSL from GBM, the results have been heterogeneous in terms of the specificity, sensitivity and accuracy of the various computational methods in differentiating these tumours precluding their use in clinical practice. Therefore, there remains a need for systematic and thorough review of all the existing literature that have looked into the classification aspect PCNSL vs GBM by various ML and DL tools.

Thus, the purpose of this systematic review and metanalysis was to estimate the diagnostic accuracy of ML-based radiomics and DL models in classifying PCNSL and GBM in an endeavour to eventually help neurooncologists in their management decisions upfront. In addition, we evaluated different combinations of selection methods and classifiers, trying to make comparison of models' performances.

Methods

Literature review

This study was conducted in concordance with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Figure 1). Quality of primary studies was assessed using the QUADAS 2 tool (Figure 2).

Literature Search Strategy:

Eligible studies reporting on the diagnostic yield of machine learning or big data in differentiating PCNSL from GBM were identified through a systematic search of the medical literature using a validated search strategy. An electronic search of Medline *via* PubMed, EMBASE and Cochrane database was conducted without any language restrictions from January 1990 till December 2021 to identify potentially relevant articles. Different key-words including Medical Subject Heading (MeSH) terms were combined using Boolean operations 'AND' and 'OR,' namely, "Magnetic Resonance Imaging" [MeSH] OR "MRI" AND "primary central nervous system lymphoma" [MeSH] OR "brain lymphoma" OR "PCNSL" AND "diagnosis" OR "accuracy" OR "yield" AND "radiomics" OR "Machine learning" OR "deep learning" OR "Artificial Intelligence" OR "AI." The Cochrane Central Register of Controlled Trials (CENTRAL) and Database of Abstracts of Reviews of Effectiveness (DARE) were also searched electronically from inception until December 2021. Electronic search was further supplemented by hand-searching of review articles, cross references, and conference proceedings.

Eligibility criteria

a. Selection of studies

All studies reporting on ML algorithms that aimed to differentiate between GBM and PCNSL on MR imaging were included. Studies that compared ML with radiologists were excluded in this meta-analysis in order to maintain homogeneity, and we intend to explore this in a subsequent paper. Articles were also excluded if they were commentaries, editorials, letters, or case reports Two reviewers (AG and JSG) extracted relevant data from each selected article, including study characteristics and findings of test results using a standardized data extraction sheet that was verified independently by the third reviewer (A.M). Any discrepancy was resolved by consensus. Quality of individual primary study in the meta-synthesis was assessed using the

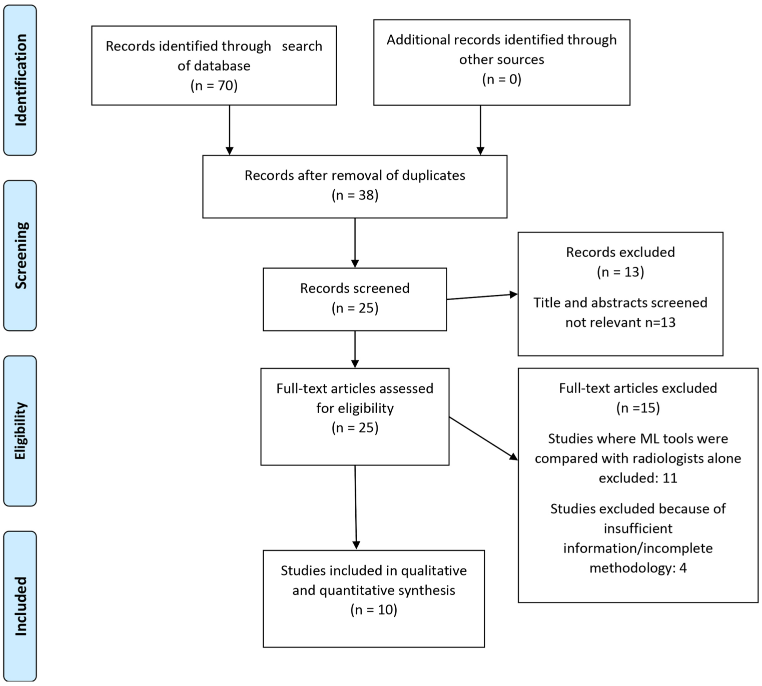

Identification
Screening
Eligibility
Included
Records identified through search of database (n = 70)
Additional records identified through other sources (n = 0)
Records after removal of duplicates (n = 38)
Records screened (n = 25)
Records excluded (n = 13)
Title and abstracts screened not relevant n=13
Full-text articles assessed for eligibility (n = 25)
Full-text articles excluded (n =15)
Studies where ML tools were compared with radiologists alone excluded: 11
Studies excluded because of insufficient information/incomplete methodology: 4
Studies included in qualitative and quantitative synthesis (n = 10)

FIGURE 1
PRISMA 2009 Flow Diagram.

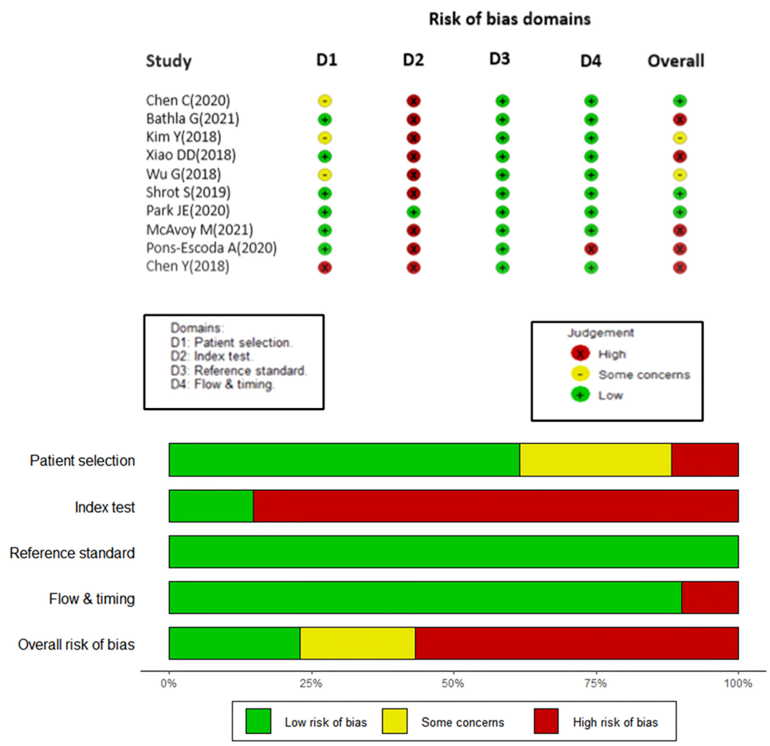


FIGURE 2
Studies included in the meta-analysis with the quality of diagnostic accuracy studies (QUADAS) scores.

QUADAS 2 quality assessment tool for studies that uses criteria scored as ‘yes, unclear, or no’ risk of bias, and assigns overall quality rating as ‘low, high, or unclear’ to each individual study. Furthermore, we also used the Radiomic Quality score (RQS), a quality assessment tool specifically developed to evaluate quality of radiomics in neuro-oncology studies (20). Studies were scored upto a maximum of 36, involving six key domains.

b. Type of participants

Patients of PCNSL & GBM with pathological confirmation of disease. In addition, all the patients had DICOM MR images of the tumour.

c. Diagnostic metrics

The diagnostic metrics included the Sensitivities and specificities of all the included studies. If papers described performance using receiver operating characteristic curves, we back-calculated possible sensitivities and specificities.

Quality assessment

Quality of primary studies was assessed using the QUADAS 2 tool and the Radiomic Quality Score (RQS) by two independent reviewers [AG and JSG]. The QUADAS-2 tool is recommended by the agency for healthcare research and Quality, the Cochrane Collaboration and the United Kingdom National Institute for Health and Clinical excellence in order to assess the risk of bias among 4 domains (patient selection, index test, reference standard and flow & timing). Any disagreement between the two reviewers were solved by mutual consensus, and then independently scored by a third reviewer (AD). Four main domains including patient selection, index test, reference standard, and flow and timing were evaluated and plotted for various risk bias domains (Figure 2).

Statistical analyses

We performed a meta-analysis of the performance of ML and DL algorithms in differentiating PCNSL from GBM. Reference standard was pathologic confirmation on biopsy of concomitant primary CNS lymphoma or GBM. Results for studies pooled in the quantitative analysis were calculated as proportions, with meta-analysis performed using the generalized linear mixed model (random-effects model) to produce summary estimates with 95% confidence intervals (CIs).All statistical analyses were performed on R Studio version

3.6.1.The 'meta', 'mada' and R-packages were used to draw forest plots. We also used the mada package, a freely available package to construct hierarchical summary receiver operating characteristic (HSROC) models, as recommended by the Cochrane Collaboration for meta-analyses of diagnostic tests. The 'robvis' package was used for QUADAS analysis. Balanced accuracy was also calculated using the average of sensitivity and specificity for all the studies. The I (2) was estimated to test level of heterogeneity.

Results

Literature search

Seventy studies of interest were found, of which 38 were duplicates. Of the remaining 32, seven were rejected based on title and abstract. Of the twenty-five full-text manuscripts retrieved, ten were selected for this meta-analysis after considering the inclusion and exclusion criteria (Figure 1). The total sample size in the 10 studies was 1311 and the overall accuracy (9 studies), sensitivity, and specificity values of each study was documented. There was no available accuracy value in one of the studies (21), nor were we able to reverse calculate it with the given information. 5 studies used a 3T MRI scanner, while 3 studies used both 3T and 1.5 T. Two studies (22, 23) did not provide details on the scanner used or the scanning protocol. None of the studies had a prospective design.

All eligible studies were relatively recent, and conducted between 2018 to 2021. 60% of the studies were conducted from hospitals in Asia (China/South Korea). The metananalysis included ten studies that compared PCNSL from GBM. A summary of the general characteristics of included studies is presented in Table 1, while the method-related information is summarized in Table 2. All studies reported at least one of the following: accuracy, sensitivity, specificity, or AUC (Table 3). Half of the studies used SVM as part of their ML algorithm, while 40% used CNN (23) (23–25), and one paper (26) used step wise selection with unsupervised learning. All of the studies performed some version of internal independent or internal cross-validation to train their ML algorithms, but only one of the studies externally validated their model (24).

Among 10 studies, seven studies were from single centre, and 3 studies were from multicentre data source (25–27).

Risk of bias assessment

The QUADAS tool assessment of risk of bias in the included studies are shown in Figure 2. In domain 1, 60% studies reported well-documented image acquisition protocols or use of publicly available image databases, with one study having a high risk of bias and two others with unclear risk in accruing for patient selection. The patient selection for these trials was based on a case-control design because outcomes were known prior to implementation of ML.

Additionally, in the second domain ("index tests"), the study designs for the papers examined had prior knowledge of the reference standard prior to implementing the index test, which introduces a high risk of bias. Hence, the authors decided to evaluate only the results of validation/test data set to conduct the statistical analysis in this study. Only one study was externally validated (21), therefore, all the other included studies were assigned a high risk of bias. As noted previously, future studies of ML should attempt to remove this risk of bias as much as possible, ideally by utilizing a prospective design and external validation.

As judged in domain 3, the reference standard of histological diagnosis was considered to provide an accurate classification of the target condition, although this reporting could be improved if the authors provided details regarding how the histological samples were obtained and processed and the specific histological characteristics that determined the diagnosis.

Finally, most of the studies apparently included all eligible patients in the analysis and had clearly defined inclusion and exclusion criteria, with a resultant low amount of bias in the fourth domain, "flow and timing".

Overall, a high risk of bias was estimated in the studies as summarized in Figure 2. Consequently, the quality assessment was limited regarding the applicability of ML based radiomics analysis.

Assessment of the radiomics quality score

The median RQS score of the 10 studies was 16.0, which was 44.4% of the ideal score of 36 (Table 4). The lowest score was 13 and the highest score was 18 (50% of the ideal quality score). Compared with the ideal score, the RQS of the selected studies was lowest in the high level of evidence domain and open science and data domain (0%), followed by biological/clinical validation, and feature reproducibility in image and segmentation.

Feature reduction was missing from the study with the lowest score (28). Meanwhile, studies with the highest score earned additional points by using validation based on a dataset from another institute.

Subgroup analysis

Data extraction

Two of the ten studies (29) (30), utilized a single MRI sequence acquired by either conventional imaging, while the remaining studies implemented both conventional and

TABLE 1 Summary of the study profile & methodology of the reviewed studies.

Sr. No	Author/ Year/ country	Patient cohort		Classifier/algorithm used	Internal validation set	External validation set	MRI used for Imaging	MRI sequences	Image Segmentation & Feature extraction tool	Reference Standard test	Type of study
		PCNSL	GBM								
1	Chen (1) et al/China/ 2018	30	66	Convolutional neural network,	Yes, cross validation	Yes	3 Tesla	T1 contrast enhanced	Scale Invariant Feature Transformation (SIFT)	Histopathology	Retrospective study
2	Xiao (2) et al/2018/ China	22	60	Machine learning: Naïve Bayes (NB), SVM, LR, Random Forest model	Yes ,10-fold internal cross validation	No	1.5 & 3 Tesla	T1W, T2W & T1W contrast enhanced	Pyradiomics	Histopathology	Retrospective study
3	Guoqing (3) et al/2018/ china	32	70	Convolutional neural network	Yes, Independent set	No	3 Tesla	Not reported	Patch based Sparse representation method	Histopathology	Retrospective study
4	Kim (4) et al/2018/ S Korea	65	78	Logistic regression, SVM, Random Forest model	No	Yes, Independent set	3 Tesla	T1W contrast, DWI, T2W	Pyradiomics	Histopathology	Retrospective study
5	Shrot (5) et al/2019/ Israel	12	41	Machine learning: Binary SVM	Yes, Leave one out cross validation	No	1.5 & 3 Tesla	MR perfusion using DSC images & DTI	3D slicer, intensity-based feature extraction from MR maps	Histopathology	Single Institutional retrospective study
6	Chen (6) et al/2020/ China	62	76	**5 selections**: Distance correlation, RF, LASSO, XG boost, GBDT; **3 Classifiers:** LDA, SVM, LR	Yes, Independent set	No	3 Tesla	T1W contrast	lifeX	Histopathology	Single Institutional retrospective study
7	Park (7) et al/Korea/ 2020	95	165	Convolutional Neural network	Yes, Internal Validation set	Yes, Independent set	3 Tesla	T1W contrast, T2 FLAIR, DSC	Segmentation done semiautomatically by two neuroradiologists	Histopathology	Retrospective study
8	Escoda (8) et al /2020/ Spain	47	48	Logistic binary regression	Yes, Independent set	No	1.5 & 3 Tesla	T1 Contrast, DSC-PWI	3D Slicer, Time intensity curve normalization	Histopathology	Retrospective study
9	Bathla (9) et al /2021/ USA	34	60	Machine learning SVM with Polynomial kernel, SVM with radial kernel ,neural network, MLP, Random Forest Model, GBRM, Adaboost	Yes (5 fold cross validation)	No	Not Reported	T1 W contrast, & FLAIR, ADC	Pyradiomics	Histopathology	Single Institutional retrospective study
10	McAvoy (10) et al/ 2021/USA	135	113	Convolutional Neural network	yes, independent	No	3 Tesla	T1 Contrast	Not Reported	Histopathology	Single Institutional retrospective study

FLAIR, Fluid Attenuation & Recovery; MLP, Multilayer Perception; SVM, Support Vector Machine; GBRM, Generalised Boosted regression Model; DTI, Diffusion tensor Imaging; DSC, Dynamic Susceptibility; ML , Machine learning; CNN, convolutional Neural network; RF, Random forest; LASSO, Least Absolute Shrinkage and Selection Operator; PCA, Principal Component Analysis; LR, Logistic Regression; SVM, Support Vector Machine; LOGISMOS, Layered Optimal Graph Image Segmentation for Multiple Objects and Services; MLP, Multilayer perceptron; SIFT, Scale invariant feature transform; mRMR, Minimum redundancy maximum relevance; CNN, Convolutional neural network; LOOCV, Leave One Out Cross Validation.

TABLE 2 Summary of the results of the reviewed studies.

Sr. No.	Author/ Year/ country	Diagnostic metrics of the best performing model from the validation set				Study limitations reported	Study strengths reported
		Accuracy	Sensitivity	Specificity	AUC		
1	Chen (1) et al/2018/ China	0.906	0.8	0.955	0.982	- Single MR sequence was used	Calculation methods are fast
2	Xiao (2) et al/2018/ China	0.82	0.78	0.91	0.9	- non enhancing & multiple lesions were excluded - Different scanners were used for image acquisition resulting in imaging protocol heterogeinity	- Image pre-processing technique used
3	Guoqing (3) et al/2018/ China	0.945	0.9	0.96	NA	Not reported	-Completely automated
4	Kim (4) et al/2018/ S Korea	0.947	0.966	0.929	0.956	- Retrospective study with patient selection bias - MR images of validation & discovery cohort were obtained from the same machine thereby may not be generalizable to other MR machines -Features were chosen empirically	Not reported
5	Shrot (5) et al/2019/ Israel	NA	1.00	1.00	NA	-ROI tracing was done manually leading to intra & interobserver variability -Small sample size - Non enhancing part of the tumour was excluded - Impact of each MR sequence on the classification model not reported	Not reported
6	Chen (6) et al/2020/ china	0.979	0.982	0.976	0.978	-Isolated evaluation of T1C images -Diagnostic Performance of radiomics based machine learning was not compared with other MR technology -Small sample size -No external validation	Not reported
7	Park (7) et al/Korea/ 2020	NA	0.95	0.76	0.89	-Diagnostic performance dropped in external data set due to overfitting - Spatial heterogeneity - Differences in contrast preloading & Image acquisition protocol results in variability of time signal intensity curves	Not reported
8	Escoda (8) et al/2020/ Spain	0.93	0.93	0.92	NA	-Retrospective nature of the study. - Wide range of MR sequences	-Near Homogenous Imaging protocol. - Balancing of tumour types - semi-automation in image segmentation & co-registration - Objective approach to classification process
9	Bathla (9) et al /2021/ USA	0.934	0.97	0.871	0.977	Small sample size -Absence of external validation set -Did not assess deep neural networks	-Well documented Imaging protocol - use of feature selection techniques, discrimination and nested cross validation
10	McAvoy (10) et al/ 2021/USA	0.93 0.94	1 0.87	0.86 1	0.94 (GBM) 0.95 (PCNSL)	- Retrospective study with small number of patients - Loss of data while exporting the image data sets	Not reported

advanced perfusion and Diffusion Weightage Imaging (DWI) sequences. An imbalance in the ratio of sample size between PCNSL cohort and GBM cohort was observed in all the studies with a ratio of almost 2:1 and 3:1 in favour of GBM cohort. However, two of the studies had a balanced sample size between PCNSL and GBM cohort (27, 29).

Heterogeneity assessment

Significant heterogeneity was present amongst the included studies regarding their scanning protocols, image sequences selected for analysis, methods of drawing ROI, feature engineering, and methodology of using ML/DL algorithms.

TABLE 3 Summary of the diagnostic metrics of all the studies included in the meta-analysis.

Author	Year	Sample Size (N)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced Accuracy (%)	AUC (%)
Chen Y (1)	2018	96	90.6	80.0	95.5	87.8	98.2
Xiao DD (2)	2018	82	82.0	78.0	91.0	84.5	90.0
Wu G (3)	2018	102	94.5	90.0	96.0	93.0	NA
Shrot S (5)	2019	53	93.6	100	100	100	NA
Chen C (6)	2020	138	97.9	98.2	97.6	97.9	97.8
Park JE (7)	2020	260	NA	95.0	76.0	85.5	89.0
Escoda A (8)	2020	95	93.0	93.0	92.0	92.5	NA
Bathla G (9)	2021	94	93.4	97.0	87.1	92.1	97.7
McAvoy (10) M	2021	248	94.0	87.0	100	93.5	95.0
Kim Y (4)	2018	143	94.7	96.6	92.9	94.7	95.6

The forest plots for balanced accuracy, sensitivity, and specificity were plotted based on the total sample size, and the forest plot for accuracy was plotted based on 1051 samples (excluding the study conducted by Park JE (21) as accuracy data was not available). The I (2) was estimated to test the level of heterogeneity; and since this was greater than 50%, random effect model for meta-analysis was used.

A large difference between the confidence region and 95% prediction regions in the Hierarchical Summary Receiver Operator Curve (HSROC) plot curve represents the heterogeneity across the studies in Figure 3. A forest plot was drawn to estimate the heterogeneity in sensitivity, specificity, accuracy and balanced accuracy as represented in Figures 4A–D. Significant heterogeneity was found in both sensitivity (I (2) 83%, $p < 0.01$), specificity (I (2) 87%, $p \leq 0.01$) and accuracy (I (2) 65%, $p \leq 0.01$).

Threshold effect assessment (HSROC)

The Spearman correlation coefficient between the sensitivity and false-positive rate was − 0.16 ($p = 0.66$), indicating the absence of a threshold effect. A threshold effect indicates a positive correlation between sensitivities and the false-positive rate that leads to a "shoulder arm" plot in the summary receiver-operating characteristic curve space. However, the visual assessment of the HSROC indicates the absence of a threshold effect as shoulder is absent in the HSROC space.

Data analysis

The HSROC based on a random effect model was applied to account for both intra- and interstudy variances in analysing the diagnostic accuracy of the ML and DL algorithms utilizing radiomic features for classifying PCNSL from GBM. The area under the curve (AUC) data available from 7 studies showed a ranged between 0.89 to 0.98 in the validation data set indicating high diagnostic performance.

Subgroup analysis

The pooled sensitivity, specificity, and accuracy were combined using a random effects model because of the heterogeneity across the reviewed studies in Figures 4A–D.

TABLE 4 Summary of Radiomics Quality Score (RQS) of individual studies.

Sr. No	Name	RQS score	%	RQS checkpoint 1 (image protocol quality)	RQS checkpoint 12	RQS checkpoint 3
1	Chen Y (1) (2018)	16	44.44%	1	1	14
2	Xiao DD (2) (2018)	16	44.44%	1	1	14
3	Wu G (3) (2018)	15	41.67%	1	1	13
4	Short S (5) (2019)	13	36.11%	1	1	11
5	Chen C (6) (2020)	17	47.22%	1	1	15
6	Park JE (7) (2020)	18	50.00%	1	1	16
7	Pons-Escoda A (8) (2020)	16	44.44%	1	1	14
8	Bathla G (9) (2021)	16	44.44%	1	1	14
9	McAvoy (10) M (2021)	16	44.44%	1	1	14
10	Kim Y (4) (2018)	17	47.22%	1	1	15

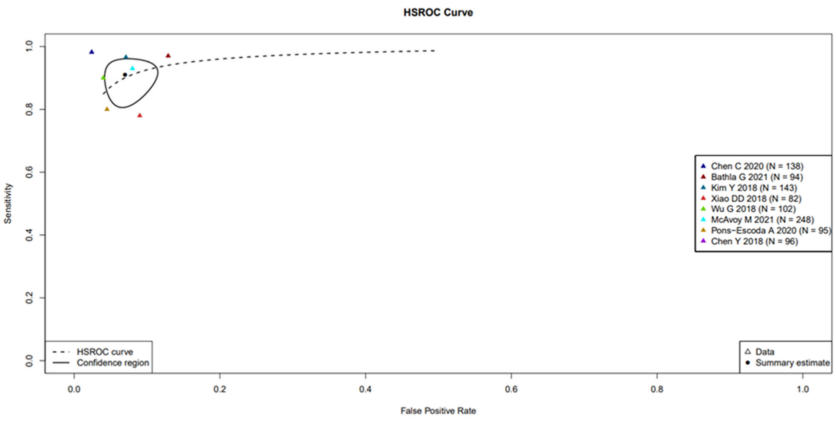
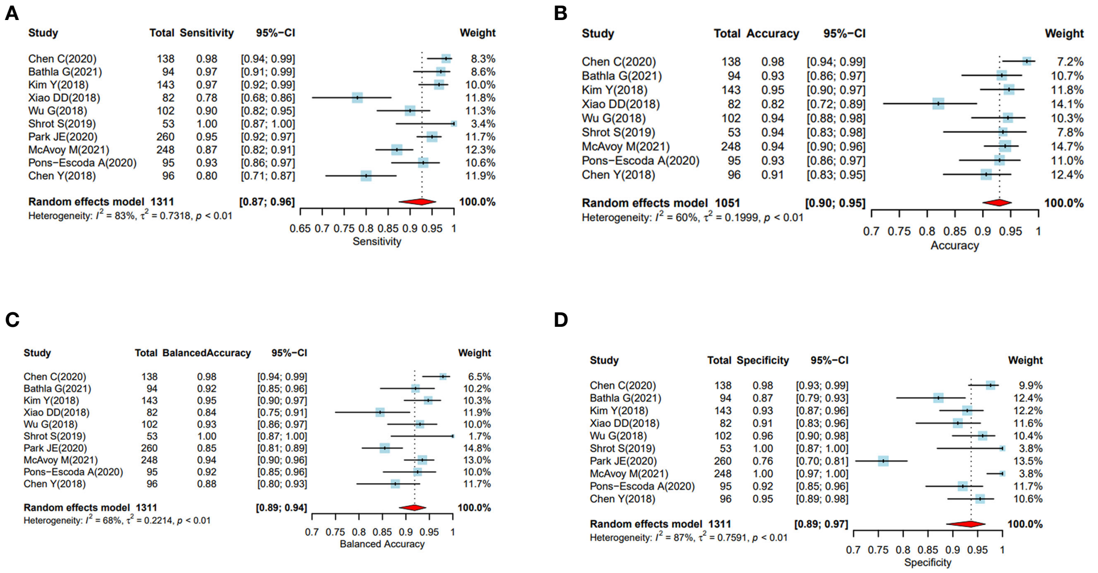


FIGURE 3
Hierarchical Summary Receiver Operator Curve (HSROC) plot displaying the diagnostic performance of radiomic based ML tools & DL tools in differentiating PCNSL from GBM. Hierarchical Summary Receiver Operator Curve (HSROC) plot displaying the diagnostic performance of radiomic based ML tools & DL tools in differentiating PCNSL from GBM. Each coloured triangle represents each of the studies in the meta-analysis. The plotted curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and the use of perfect reference standards. The "TP", "FP", "FN", "TN" rates for the two studies (Park JE 2020 and Shrot S 2019 studies) as the former study has no available accuracy value and the latter one has both sensitivity and specificity equal to one.

In the subgroup analysis, the overall sensitivity of diagnosing PCNSL was lower (92% (95% CI, 0.88, 0.95)) than the specificity (94% (95% CI, 0.89, 0.97)). We did not find any significant differences in sensitivity, specificity or accuracy based on sample sizes less than or greater than 100.

Discussion

This systematic review and meta-analysis evaluated the efficacy of deep learning/machine learning based algorithms in differentiating PCNSL from GBMs, a dilemma often



A

Study	Total	Sensitivity	95%−CI	Weight
Chen C(2020)	138	0.98	[0.94; 0.99]	8.3%
Bathla G(2021)	94	0.97	[0.91; 0.99]	8.6%
Kim Y(2018)	143	0.97	[0.92; 0.99]	10.0%
Xiao DD(2018)	82	0.78	[0.68; 0.86]	11.8%
Wu G(2018)	102	0.90	[0.82; 0.95]	11.3%
Shrot S(2019)	53	1.00	[0.87; 1.00]	3.4%
Park JE(2020)	260	0.95	[0.92; 0.97]	11.7%
McAvoy M(2021)	248	0.87	[0.82; 0.91]	12.3%
Pons−Escoda A(2020)	95	0.93	[0.86; 0.97]	10.6%
Chen Y(2018)	96	0.80	[0.71; 0.87]	11.9%
Random effects model	**1311**		**[0.87; 0.96]**	**100.0%**

Heterogeneity: $I^2 = 83\%$, $\tau^2 = 0.7318$, $p < 0.01$

B

Study	Total	Accuracy	95%−CI	Weight
Chen C(2020)	138	0.98	[0.94; 0.99]	7.2%
Bathla G(2021)	94	0.93	[0.86; 0.97]	10.7%
Kim Y(2018)	143	0.95	[0.90; 0.97]	11.8%
Xiao DD(2018)	82	0.82	[0.72; 0.89]	14.1%
Wu G(2018)	102	0.94	[0.88; 0.98]	10.3%
Shrot S(2019)	53	0.94	[0.83; 0.98]	7.8%
McAvoy M(2021)	248	0.94	[0.90; 0.96]	14.7%
Pons−Escoda A(2020)	95	0.93	[0.86; 0.97]	11.0%
Chen Y(2018)	96	0.91	[0.83; 0.95]	12.4%
Random effects model	**1051**		**[0.90; 0.95]**	**100.0%**

Heterogeneity: $I^2 = 60\%$, $\tau^2 = 0.1999$, $p < 0.01$

C

Study	Total	BalancedAccuracy	95%−CI	Weight
Chen C(2020)	138	0.98	[0.94; 0.99]	6.5%
Bathla G(2021)	94	0.92	[0.85; 0.96]	10.2%
Kim Y(2018)	143	0.95	[0.90; 0.97]	10.3%
Xiao DD(2018)	82	0.84	[0.75; 0.91]	11.9%
Wu G(2018)	102	0.93	[0.86; 0.97]	10.0%
Shrot S(2019)	53	1.00	[0.87; 1.00]	1.7%
Park JE(2020)	260	0.85	[0.81; 0.89]	14.8%
McAvoy M(2021)	248	0.94	[0.90; 0.96]	13.0%
Pons−Escoda A(2020)	95	0.92	[0.85; 0.96]	10.0%
Chen Y(2018)	96	0.88	[0.80; 0.93]	11.7%
Random effects model	**1311**		**[0.89; 0.94]**	**100.0%**

Heterogeneity: $I^2 = 68\%$, $\tau^2 = 0.2214$, $p < 0.01$

D

Study	Total	Specificity	95%−CI	Weight
Chen C(2020)	138	0.98	[0.93; 0.99]	9.9%
Bathla G(2021)	94	0.87	[0.79; 0.93]	12.4%
Kim Y(2018)	143	0.93	[0.87; 0.96]	12.2%
Xiao DD(2018)	82	0.91	[0.83; 0.96]	11.6%
Wu G(2018)	102	0.96	[0.90; 0.98]	10.4%
Shrot S(2019)	53	1.00	[0.87; 1.00]	3.8%
Park JE(2020)	260	0.76	[0.70; 0.81]	13.5%
McAvoy M(2021)	248	1.00	[0.97; 1.00]	3.8%
Pons−Escoda A(2020)	95	0.92	[0.85; 0.96]	11.7%
Chen Y(2018)	96	0.95	[0.89; 0.98]	10.6%
Random effects model	**1311**		**[0.89; 0.97]**	**100.0%**

Heterogeneity: $I^2 = 87\%$, $\tau^2 = 0.7591$, $p < 0.01$

FIGURE 4
(A–D) Performance evaluation of the ML and DL algorithms of all the studies in distinguishing PCNSL from GBM as represented by the random forest plots. **(A)** Forest plots of sensitivity. **(B)**: Forest plots of accuracy. **(C)** Forest plot of balanced accuracy, **(D)** Forest plots of specificity.

encountered with neuroradiologists and the neurosurgeons, often requiring invasive biopsies to classify the above entities. Apart from the dilemma of differentiating the above malignant lesions, neuroradiologists often also have difficulty in differentiating PCNSL from inflammatory conditions (multiple sclerosis and tumifactive demyelination) often having therapeutic implications. Therefore, having non-invasive tools like radiomics and AI would help increase the diagnostic ability of the neuroradiologists to differentiate not only malignant from benign inflammatory conditions but also classify malignant lesions like PCNSL and GBM which are hitherto difficult to distinguish using radiological semantic features. This could be further useful in patients where histopathological examination cannot be done due to a multitude of reasons such as deep location within the brain and poor performance status.IN such a scenario, non- invasive methods like radiomics and deep learning from the MR images may help the clinician.

Although radiomics and deep learning algorithms have been used for a multitude of neurological conditions (31–34) its use in classifying malignant conditions and differentiating them is of paramount importance as the therapy and prognosis changes across the spectrum of brain tumours. The present study highlights the use of ML and DL algorithms for discriminating PCNSL and GBM on radiological imaging. We identified 10 studies that trained predictive models using ML or DL algorithms to classify PCNSL from GBM against the reference gold standard histopathology. All the studies, used classifiers that trained on radiomic features extracted from MR images or classifiers using deep learning algorithms like convolutional neural network (CNN).

The pooled analysis of all the studies showed encouraging results with ML or DL classifiers performing extremely well with highest accuracy of 97.9% and the lowest of 82% in differentiating PCNSL from GBM. The area under the curve (AUC) ranged between 89%-98.2% among all the studies that were reviewed. (Table 4 and Figure 4)

The diagnostic metrics from the pooled analysis of the results of the 10 studies showed a high degree of concordance in classifying PCNSL from GBM as against the reference histopathology. However, these positive results must be interpreted with caution as a multitude of factors such as small sample size, heterogenous imaging protocols, patient selection criteria into the training and the validation set may have led to overfitting of the data at the time of model development. Overfitting is common in radiomic studies involving machine learning and deep earning classifiers that reduces its potential for immediate incorporation into clinical practise and use it for treatment decisions (29–31).

Therefore, ML and DL classifiers need to be trained in large data sets using highly heterogenous population. Further, these models (classifiers) show variations with subtle changes in the methods of segmentation, pre-processing of MR images acquired from heterogenous MR machines. A previously conducted systematic review and meta-analysis in 2018 included 8 studies which used ML based classifiers for differentiating PCNSL from GBM. Seven of the eight studies did not have any external validation except for one study in which the ML classifier modelled on the training set was validated on an external data set (32). Similar to the above metaanalysis, our metaanalysis also had a single study that was externally validated on a different data set. There has been a spurt in publications on ML/DL models in classifying PCNSL versus GBM since 2018, and we found a total of 24 articles investigating role of ML and DL algorithms for classifying PCNSL and GBM. In order to make the meta-analysis more robust, we focussed on studies reporting on the performance of their ML/DL models exclusively against the reference standard (histopathology).

A recent systematic review from 23 studies also investigated the role of DL & ML in differentiating PCNSL from all grades (Grade-II-IV) of Gliomas (31). However, significant differences exist in the methodology and the search strategy of our metaanalysis. Moreover, our metaanalysis included only those studies that used ML or AL for differentiating PCNSL from Glioblastoma against the gold standard histopathology. By combining all the studies, on DL/ML in differentiate ng PCNSL from GBM, they were left with a heterogenous dataset precluding any further mathematical analysis to derive a meaningful data, and hence had to contend with only a systematic review of the available literature.

However, at the very outset and literature search strategy stage of our manuscript, we identified the heterogeneity in methodology of the conducted studies, and realised they could be broadly classified into 2 types- those comparing ML/DL with histopathology as a gold standard, and then those comparing ML/DL models with radiologists performance. We found around 12 papers under each category, and analysed them separately. This current manuscript deals with the performance of ML/DL methods versus histopathology as a gold standard. Hence, mathematical analysis in the form of statistical tests for a meta-analysis were performed to evaluate the proof of performance of advanced computing methods in differentiating PCNSL from GBM and not other gliomas.

To summarise, ML and DL tools may complement the radiologic features to differentiate PCNSL from GBM. These tools may have the potential to assist radiologists in approaching cases that may have features common to both PCNSL and GBM. Presently these algorithms may have certain deficiencies, however with refinement in the computing processes, ML/DL based models will likely help the neurosurgeons improve the quality of managing patients of brain tumours by optimizing the use of invasive diagnostic procedures in the future, thereby reducing the incidence of complications that compromise patient quality of life and life expectancy while expediting initiation of intervention.

Strengths of the study

We assessed ML and DL performance in both internal and external validation data sets which enhanced the credibility of the review. Being able to compare both the analytic methods to the gold standard histopathology in the test cohort and validation cohort have produced fairly clear results.

Limitations of the Study

Application of ML in neuroradiology for solving the dilemma of whether an image depicts GBM or PCNSL is relatively new. There is currently a limited number of publications that address this scientific inquiry. Our search strategy for the present study only included limited databases (PubMed, EMBASE & Cochrane database). All the studies that were reviewed varied in terms of the imaging protocols used, types of MRI machines used, MR sequences used (i.e., T1-weighted, T2-weighted, diffusion-weighted, etc.), method of tumour segmentation, tools for feature selection and reduction and ultimately the types of classifiers used for training the image datasets. Future studies that address distinguishing GBM from PCNSL should prospectively evaluate the performance of their model and also consider the utility of newer MRI techniques that may improve differentiation of these two pathologies. Additionally, our assessment of bias revealed inherent issues with applying the QUADAS-2 to ML studies. Despite these limitations, we maintain that assessment of bias is an absolute necessity.

Future directions:

Prospective multicentre trials are the need of the hour to generate more robust data so that results from an independent external validation dataset are available. The inherent variability across studies with regard to the process of conducting each step leading to the radiomics model could be attributed to high bias and heterogeneity, not necessarily underlying biologic effects, standardization in image acquisition, segmentation methodology, feature selection and classification, statistical analysis, and the reporting format should be established for reproducibility and the generalization of ML-based radiomics studies (33). Essential steps for standardization include optimizing the standard imaging acquisition process, fully automating the process for segmentation and feature engineering, reducing the redundancy of feature numbers, enhancing the reproducibility of radiomics features, and reporting the results transparently. The guidelines suggested by the relevant professional societies, such as the Society of Nuclear Medicine and Molecular Imaging, the Quantitative Imaging Network, Radiology Society of North America, and the European Society of Radiology that lead the field in imaging methods, including radiomics, should be considered (34).

Conclusion

The systematic review of studies investigating ML & DL based algorithms to differentiate PCNSL from GBM have demonstrated encouraging results and certainly have the potential to aid neurooncologists in taking preoperative treatment decisions in the future leading to not only reduction in morbidities but also be cost effective. It is likely that predictive analytics using ML or DL based algorithms will help optimize diagnostic decision-making process and individualise patient management. Although studies had limited sample size, formal predictive analytics, using these models may have the potential to improve clinician performance complementing human expertise and experience with the computational power. However, one must keep in mind the pitfalls associated with overfitting the data due to limited image data sets and resultant lack of training these algorithms to maximize the generalizability and their utility. Therefore, prospective multicentric trials with large data sets should be initiated to train the models on large heterogeneous and real-world data sets that account for the heterogeneity encountered in acquisition of images in the real-world clinical practice.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

AG & JG: Conceptualisation and designing the study, writing the manuscript, analysing the results. ST, SH, JG: statistical analysis. AD: verification of results, bias risk assessment, AM: review of manuscript. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Surawicz TS, McCarthy BJ, Kupelian V, Jukich PJ, Bruner JM, Davis FG, et al. Descriptive epidemiology of primary brain and CNS tumors: Results from the central brain tumor registry of the united states, 1990-1994. *Neuro Oncol* (1999) 1(1):14–25. doi: 10.1093/neuonc/1.1.14

2. Malikova H, Koubska E, Weichet J, Klener J, Rulseh A, Liscak R, et al. Can morphological MRI differentiate between primary central nervous system lymphoma and glioblastoma? *Cancer Imaging* (2016) 16(1):40. doi: 10.1186/s40644-016-0098-9

3. Han Y, Wang Z-J, Li W-H, Yang Y, Zhang J, Yang X-B, et al. *Differentiation Between Primary Cent Nervous System Lymphoma Atypical Glioblastoma Based MRI Morphological Feature Signal Intensity Ratio: A Retrospective Multicenter Study. Front Oncol* (2022) 12:811197. doi: 10.3389/fonc.2022.811197

4. Al-Okaili RN, Krejza J, Woo JH, Wolf RL, O'Rourke DM, Judy KD, et al. Intraaxial brain masses: MR imaging-based diagnostic strategy–initial experience. *Radiology.* (2007) 243(2):539–50. doi: 10.1148/radiol.2432060493

5. Ozturk K, Soylu E, Cayci Z. Differentiation between primary CNS lymphoma and atypical glioblastoma according to major genomic alterations using diffusion and susceptibility-weighted MR imaging. *Eur J Radiol* (2021) 141:109784. doi: 10.1016/j.ejrad.2021.109784

6. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging–"how-to" guide and critical reflection. *Insights Imaging* (2020) 11(1):1–16. doi: 10.1186/s13244-020-00887-2/tables/3

7. Li Y, Liu X, Qian Z, Sun Z, Xu K, Wang K, et al. Genotype prediction of ATRX mutation in lower-grade gliomas using an MRI radiomics signature. *Eur Radiol* (2018) 28(7):2960–8. doi: 10.1007/S00330-017-5267-0

8. Ditmer A, Zhang B, Shujaat T, Pavlina A, Luibrand N, Gaskill-Shipley M, et al. Diagnostic accuracy of MRI texture analysis for grading gliomas. *J Neurooncol* (2018) 140(3):583–9. doi: 10.1007/S11060-018-2984-4

9. Zhou H, Vallières M, Bai HX, Su C, Tang H, Oldridge D, et al. MRI Features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol* (2017) 19(6):862–70. doi: 10.1093/NEUONC/NOW256

10. Lee JH, Han SS, Hong EK, Wei Y, Zhang Y, Chen J, et al. Predicting lymph node metastasis in pancreatobiliary cancer with magnetic resonance imaging: A prospective analysis. *Eur J Radiol* (2019) 116:1–7. doi: 10.1016/J.EJRAD.2019.04.007

11. Xie J, Liu R, Luttrell J, Zhang C. Deep learning based analysis of histopathological images of breast cancer. *Front Genet* (2019) 10:80/BIBTEX. doi: 10.3389/FGENE.2019.00080/BIBTEX

12. van IJzendoorn DGP, Szuhai K, Briaire-De Bruijn IH, Kostine M, Kuijjer ML, Bovée JVMG. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PloS Comput Biol* (2019) 15(2):e1006826. doi: 10.1371/JOURNAL.PCBI.1006826

13. Ko SY, Akahata W, Yang ES, Kong W-P, Burke CW, Honnold SP, et al. A virus-like particle vaccine prevents equine encephalitis virus infection in nonhuman primates. *Sci Transl Med* (2019) 11(492). doi: 10.1126/SCITRANSLMED.AAV3113

14. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. *Am J Roentgenol* (2017) 208(4):754–60. doi: 10.2214/AJR.16.17224

15. Kunimatsu A, Kunimatsu N, Yasaka K, Akai H, Kamiya K, Watadani T, et al. Machine learning-based texture analysis of contrast-enhanced MR imaging to differentiate between glioblastoma and primary central nervous system lymphoma. *Magn Reson Med Sci* (2019) 18(1):44–52. doi: 10.2463/MRMS.MP.2017-0178

16. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* (2020) 295(2):328–38. doi: 10.1148/radiol.2020191145

17. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Stiphout van RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48(4):441–6. doi: 10.1016/J.EJCA.2011.11.036

18. Tang B, Pan Z, Yin K, Khateeb A. Recent advances of deep learning in bioinformatics and computational biology. *Front Genet* (2019) 10:214/BIBTEX. doi: 10.3389/FGENE.2019.00214/BIBTEX

19. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* (2018) 9 (4):611–29. doi: 10.1007/S13244-018-0639-9/FIGURES/15

20. Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, et al. A systematic review reporting quality of radiomics research in neuro-oncology: Toward clinical utility and quality improvement using high-dimensional imaging features. *BMC Cancer.* (2020) 20(1):1–11. doi: 10.1186/S12885-019-6504-5/TABLES/3

21. Park JE, Kim HS, Lee J, Cheong EN, Shin I, SooAhn S, et al. Deep-learned time-signal intensity pattern analysis using an autoencoder captures magnetic resonance perfusion heterogeneity for brain tumor differentiation. *Sci Rep* (2020) 10(1):1–11. doi: 10.1038/s41598-020-78485-x

22. Bathla G, Priya S, Liu Y, Ward C, Le NH, Soni N, et al. Radiomics-based differentiation between glioblastoma and primary central nervous system lymphoma: a comparison of diagnostic performance across different MRI sequences and machine learning techniques. *Eur Radiol* (2021) 31(11):8703–13. doi: 10.1007/s00330-021-07845-6

23. Chen Y, Li Z, Wu G, Yu J, Wang Y, Lv X, et al. Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features. *Int J Neurosci* (2018) 128(7):608–18. doi: 10.1080/00207454.2017.1408613

24. Kang D, Park JE, Kim Y-H, Kim JH, Oh J, Kim J, et al. Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. *Neuro Oncol* (2018) 20(9):1251–61. doi: 10.1093/neuonc/noy021

25. McAvoy M, Calvachi Prieto P, Kaczmarzyk JR, Fernández IS, McNulty J, Smith T, et al. Classification of glioblastoma versus primary central nervous system lymphoma using convolutional neural networks. *Sci Rep* 11:15219 doi: 10.1038/s41598-021-94733-0

26. Pons-Escoda A, Garcia-Ruiz A, Naval-Baudin P, Cos M, Vidal N, Plans G, et al. Presurgical identification of primary central nervous system lymphoma with normalized time-intensity curve: A pilot study of a new method to analyze DSC-PWI. *AJNR Am J Neuroradiol* (2020) 41(10):1816–24. doi: 10.3174/AJNR.A6761

27. Kim Y, Cho H-H, Kim ST, Park H, Nam D, Kong D-S. Radiomics features to distinguish glioblastoma from primary central nervous system lymphoma on multi-parametric MRI. *Neuroradiology* (2018) 60(12):1297–305. doi: 10.1007/s00234-018-2091-4

28. Shrot S, Salhov M, Dvorski N, Konen E, Averbuch A, Hoffmann C. Application of MR morphologic, diffusion tensor, and perfusion imaging in the classification of brain tumors using machine learning scheme. *Neuroradiology* (2019) 61(7):757–65. doi: 10.1007/s00234-019-02195-z

29. Chen C, Zheng A, Ou X, Wang J, Ma X. Comparison of radiomics-based machine-learning classifiers in diagnosis of glioblastoma from primary central nervous system lymphoma. *Front Oncol* (2020) 10:1151. doi: 10.3389/fonc.2020.01151

30. Xiao D-D, Yan P-F, Wang Y-X, Osman MS, Zhao H-Y. Glioblastoma and primary central nervous system lymphoma: Preoperative differentiation by using MRI-based 3D texture analysis. *Clin Neurol Neurosurg* (2018) 173:84–90. doi: 10.1016/j.clineuro.2018.08.004

31. Cassinelli Petersen GI, Shatalov J, Verma T, Brim WR, Subramanian H, Brackett A, et al. Machine learning in differentiating gliomas from primary CNS lymphomas: A systematic review, reporting quality, and risk of bias assessment. *Am J Neuroradiol* (2022) 43(4):526–33. doi: 10.3174/ajnr.A7473

32. Sotoudeh H, Sadaatpour Z, Rezaei A, Shafaat O, Sotoudeh E, Tabatabaie M, et al. The role of machine learning and radiomics for treatment response prediction in idiopathic normal pressure hydrocephalus. *Cureus* (2021) 13(10). doi: 10.7759/CUREUS.18497

33. Ristow I, Madesta F, Well L, Shenas FY, Wright F, Molwitz I, et al. Evaluation of magnetic resonance imaging-based radiomics characteristics for differentiation of benign and malignant peripheral nerve sheath tumors in neurofibromatosis type 1. *Neuro Oncol* (2022), 1–9. doi: 10.1093/NEUONC/NOAC100

34. Park JE, Kim HS, Lee J, Cheong EN, Shin I, Ahn SS, et al. Deep-learned time-signal intensity pattern analysis using an autoencoder captures magnetic resonance perfusion heterogeneity for brain tumor differentiation. *Sci Rep* (2020) 10(1):21485. doi: 10.1038/s41598-020-78485-xx

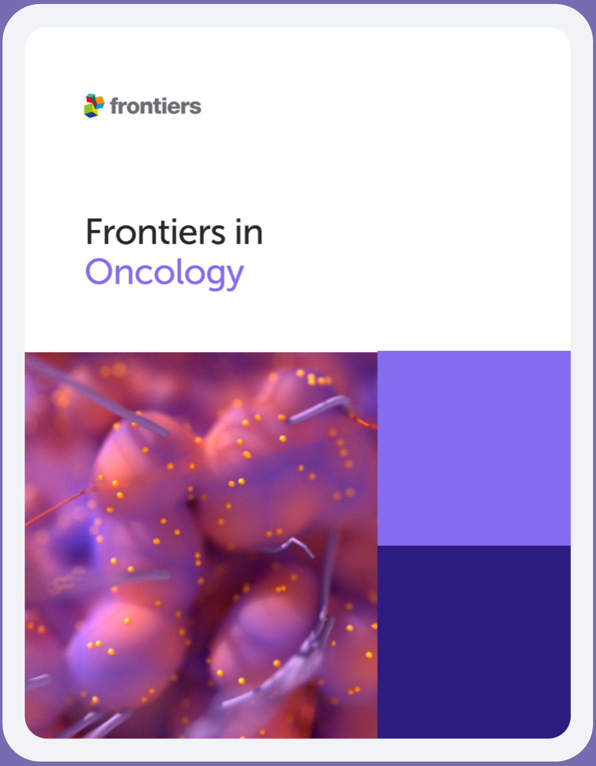
Frontiers in Oncology

Advances knowledge of carcinogenesis and tumor progression for better treatment and management

The third most-cited oncology journal, which highlights research in carcinogenesis and tumor progression, bridging the gap between basic research and applications to imrpove diagnosis, therapeutics and management strategies.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Oncology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact