frontiers | Research Topics

Artificial intelligence in cardiovascular research

Edited by
Ulrich Parlitz and Marta Varela

Published in
Frontiers in Network Physiology

ISSN 1664-8714
ISBN 978-2-8325-7234-4
DOI 10.3389/978-2-8325-7234-4

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Artificial intelligence in cardiovascular research

Topic editors
Ulrich Parlitz — Max Planck Institute for Dynamics and Self-Organization, Germany
Marta Varela — Imperial College London, United Kingdom

Citation
Parlitz, U., Varela, M., eds. (2025). *Artificial intelligence in cardiovascular research*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-7234-4

Table of contents

TYPE Original Research
PUBLISHED 19 July 2024
DOI 10.3389/fnetp.2024.1425871

OPEN ACCESS

EDITED BY
Marta Varela,
Imperial College London, United Kingdom

REVIEWED BY
Huiyi Wu,
Imperial College London, United Kingdom
Edoardo Occhipinti,
Imperial College London, United Kingdom

*CORRESPONDENCE
Christoph Schranz,
✉ christoph.schranz@salzburgresearch.at

RECEIVED 30 April 2024
ACCEPTED 17 June 2024
PUBLISHED 19 July 2024

CITATION
Schranz C, Halmich C, Mayr S and Heib DPJ (2024), Surrogate modelling of heartbeat events for improved J-peak detection in BCG using deep learning.
Front. Netw. Physiol. 4:1425871.
doi: 10.3389/fnetp.2024.1425871

Surrogate modelling of heartbeat events for improved J-peak detection in BCG using deep learning

Christoph Schranz[1]*, Christina Halmich[1,2], Sebastian Mayr[1] and Dominik P. J. Heib[3]

[1]Human Motion Analytics, Salzburg Research Forschungsgesellschaft mbH, Salzburg, Austria, [2]Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Salzburg, Austria, [3]Institut Proschlaf, Salzburg, Austria

Sleep, or the lack thereof, has far-reaching consequences on many aspects of human physiology, cognitive performance, and emotional wellbeing. To ensure undisturbed sleep monitoring, unobtrusive measurements such as ballistocardiogram (BCG) are essential for sustained, real-world data acquisition. Current analysis of BCG data during sleep remains challenging, mainly due to low signal-to-noise ratio, physical movements, as well as high inter- and intra-individual variability. To overcome these challenges, this work proposes a novel approach to improve J-peak extraction from BCG measurements using a supervised deep learning setup. The proposed method consists of the modeling of the discrete reference heartbeat events with a symmetric and continuous kernel-function, referred to as surrogate signal. Deep learning models approximate this surrogate signal from which the target heartbeats are detected. The proposed method with various surrogate signals is compared and evaluated with state-of-the-art methods from both signal processing and machine learning approaches. The BCG dataset was collected over 17 nights using inertial measurement units (IMUs) embedded in a mattress, together with an ECG for reference heartbeats, for a total of 134 h. Moreover, we apply for the first time an evaluation metric specialized for the comparison of event-based time series to assess the quality of heartbeat detection. The results show that the proposed approach demonstrates superior accuracy in heartbeat estimation compared to existing approaches, with an MAE (mean absolute error) of 1.1 s in 64-s windows and 1.38 s in 8-s windows. Furthermore, it is shown that our novel approach outperforms current methods in detecting the location of heartbeats across various evaluation metrics. To the best of our knowledge, this is the first approach to encode temporal events using kernels and the first systematic comparison of various event encodings for event detection using a regression-based sequence-to-sequence model.

KEYWORDS

ballistocardiogram, ballistocardiography, J-peak, heartrate estimation, event detection, peak detection, deep learning, ResNet

1 Introduction

Sleep has a profound influence on the physical and mental functioning of the following day, especially when it is lacking. Restful sleep promotes physical regeneration (Jakowski et al., 2023) and the performance of the immune system (Garbarino et al., 2021), as well as cognitive performance (Brownlow et al., 2020), motor dexterity (Craven et al., 2022) and emotional stability (Tomaso et al., 2021). Disturbed sleep can lead to serious chronic health problems such as cardiovascular disease, endocrinological dysregulation, and a range of psychological impairments (Itani et al., 2017). To elucidate the dynamics and interactions of sleep, a comprehensive understanding of sleep patterns and stages is essential. The gold standard for objectively measuring sleep is polysomnography (PSG), which integrates electroencephalography, electromyography, and electrooculography. However, PSG recordings are time-consuming, the equipment is costly, and trained personnel are required to ensure sufficient signal quality. These drawbacks limit the sample sizes of PSG studies and render them unsuitable for longitudinal studies. Recent trends in sleep research indicate that high-accuracy estimations of sleep stage fluctuations can be derived from variations in signals such as inter-beat interval (IBI) time-series (kranzinger et al., 2023) or alterations in respiratory effort over time when analyzed with machine learning models.

Today, IBI time-series can be accurately recorded using inexpensive consumer devices, making inter-beat intervals a promising signal for large-scale sleep studies aimed at gaining new insights into sleep. State-of-the-art sensors for measuring heartbeats can be categorized into on-body (wearables) and off-body (contactless) solutions. On-body systems include devices directly attached to the body, such as electrocardiography and photoplethysmography to acquire electrocardiograms (ECG) and photoplethysmograms (PPG). PPG wearables, typically mounted on the wrist, arm, or earlobe, measure heartbeats by detecting periodic changes in the optical reflection of emitted light caused by blood pulses under the skin.

Contactless sensor systems mainly involve camera-based systems as well as ballistocardiography. Camera-based systems sense minor periodic changes in the skin color caused by blood pulses. Ballistocardiography is a sensor system that measures subtle accelerations of the human body, including cardiovascular and respiratory activity, that are plotted in ballistocardiograms (BCG). The heartbeat events in BCG are referred to as J-peaks and are caused by the contraction of the heart which results in the ejection of blood into the aortic arch where the direction of flow is changed, creating a momentum (Giovangrandi et al., 2011). Therefore, the systolic J-peak occurs after the electrical trigger, i.e. the R-peak, as accessed from an ECG. These events also occur closer to the heart and are sharper in their waveform than camera-based solutions or PPG, where the monitored pulses are measured at the skin or wrist.

The delay between the electrical activation of the heart muscle to the greatest vertical force as measured by a BCG-system is referred to as RJ-interval. A schematic ECG and BCG with their corresponding QRS- and IJK-complexes are illustrated in Figure 1. According to the literature, the RJ-intervals vary typically between 180 ms and 240 ms and change slowly over time (Casanella et al., 2012). J-peaks in BCG occur closer to the heart and are sharper in its waveform than PPG, which measures pulses at the skin or wrist.

The inter-subject variability is caused by different causes such as body mass, heart size, body placement relative to the fixed sensor position, body alignment, and also the physiological state of the subject. Additionally, the BCG also depends on the used sensor type and setup Sadek and Abdulrazak (2021). Some activities, such as paced respiration, can induce hemodynamic changes that affect the RJ-interval by 150 ms–300 ms (Casanella et al., 2012; Gomez-Clapers et al., 2014).

Ballistocardiogram can be implemented using different sensor technologies. The most commonly used are inertial measurement units (IMU) (Cathelain et al., 2020), electromechanical films (Sadek and Abdulrazak, 2021), or piezoelectric (Zhou et al., 2021; Liu et al., 2022), hydraulic- (Heise and Skubic, 2010) or pneumatic- (Pröll et al., 2019) pressure sensors. In each case, the sensor is integrated into the mattress, pillow, mat or chair underneath the person, allowing for an unobtrusive measurement. This type of unobtrusive measurement in particular offers a seamless recording of sleep data over several weeks at home without the need for a sleep laboratory, as no sensors or wearables need to be actively applied or activated. Data acquisition is activated simply by lying in bed.

Off-body measurement systems therefore offer a more elegant and unobtrusive means of recording physiological information over extended periods, as they typically require minimal interaction with the recording device, thereby reducing distress and potential user resistance associated with long-term use of wearables. However, as the indirect measurement leads to a decreased signal-to-noise ratio, detecting individual heartbeats from contactless sensors is significantly more challenging compared to wearables. This limitation of contactless devices is critical, as the accuracy of automatic sleep stage classification based on IBI time-series depends on the temporal precision of the captured IBIs. Therefore, advancements in heartbeat extraction from BCG are crucial.

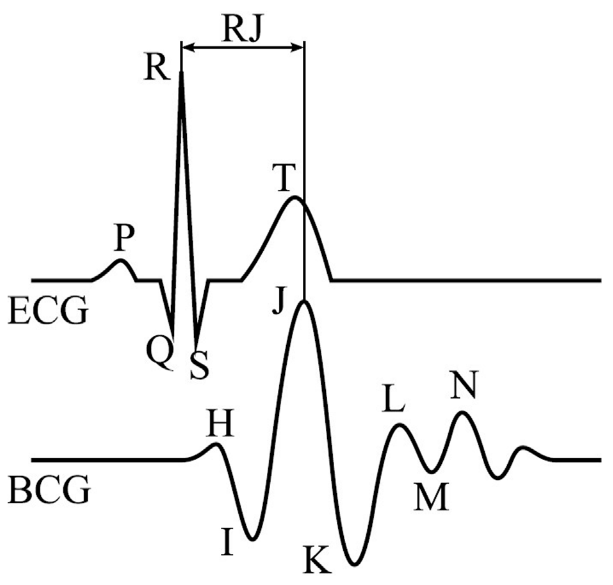


FIGURE 1
Single beat of an ECG (top) and BCG (bottom) with their annotated main waves and RJ-interval (Gomez-Clapers et al., 2014).

Given the suitability of ballistocardiography for unobtrusive long-term sleep measurement, and the availability of improved machine learning algorithms and computational resources, there is an increasing amount of research focused on heartbeat extraction from BCG. Research has shown that J-peaks can be used to predict the subject's sleep stages and therefore sleep quality (Kranzinger et al., 2023). This work proposes a novel approach to model heartbeat events as a continuous signal, thereby improving the accuracy of heartbeat extraction within a supervised deep learning framework. The primary objective is to evaluate various heartbeat event representations in combination with different deep learning network architectures for J-peak detection in BCG signals and to compare them with existing methods.

Section 2 introduces the state-of-the-art methods and reasons why machine learning approaches might offer advantages in overcoming existing limitations of current contact-less methods. Section 3 provides a formal introduction to the problem from a theoretical perspective, and Section 4 presents the proposed method and evaluation of the methods. In Section 5 the results of the method comparison are presented. Finally, the findings are discussed in Section 6 and concluded in Section 7.

2 Related work

The classical approach for the detection of heartbeats in ballistocardiogram (BCG) is the Pan-Tompkins algorithm (Pan and Tompkins, 1985). This algorithm was initially developed for ECG and is based on classical signal processing techniques, such as low and high-pass filtering, derivates, functional mappings, and averaging. Using the thereby processed signal, a peak detection algorithm is applied to detect the characteristic R-peak of the ECG. However, the low signal-to-noise ratio of BCG data limits the detection of heartbeats using the same approach. Hence, the Pan-Tompkins algorithm was adapted for the application on BCG data. Most solutions employ a bandpass filter as the initial processing step, with a recommended system bandwidth ranging from 1.5 Hz to 22.5 Hz. This frequency band encompasses all relevant cardiovascular signals while filtering out respiratory activity and movements (Gomez-Clapers et al., 2014). For instance, Pröll et al. (2019), applied the following sequential processing steps: bandpass filter, cubic function, low-pass filter, second order derivate, absolute value function, and low-pass filter (Pröll et al., 2019). This processing pipeline transforms the raw BCG signal into a signal that exhibits the characteristic J-peak of the BCG more significantly. A subsequent peak detection identifies the IJK-complex that is analogous to the QRS-complex in ECG.

Other classical signal processing approaches apply wavelet transformations, template matching, or signal envelopes (Pino et al., 2017; Sadek and Abdulrazak, 2021). Additionally, some approaches apply methods in the frequency domain (Brüser et al., 2011). Analogously, classical signal processing approaches for R-peak extraction in noisy ECG and PPG measurements are based on a similar combination of algorithms (Nguyen et al., 2019; Yun et al., (2022)).

As pointed out in Section 1, the substantial inter- and intrasubject variability of BCG as well as the low signal-to-noise ratio remain major challenges of the J-peak extraction in BCG measurements. In order to address these challenges, neural networks can be used that are effective in capturing the variability of BCG within a data-driven supervised machine learning setting. Pröll et al. (2021); Sadek and Abdulrazak (2021) demonstrated that their deep learning approach, using a combination of convolutional neural networks (CNN) and recurrent network layers (LSTM and GRU) of different sizes, has improved the accuracy of estimating the mean heart rate within 8 s epochs by more than 50% in terms of MAE compared to five state-of-the-art digital signal processing approaches (from 4.24 to 2.07). This approach estimates the heart rate from a BCG signal and compares it with a reference heart rate as accessed from an ECG. Other approaches apply deep learning models to approximate a signal with characteristic J-peaks. For example, Cathelain et al. (2020); Zhou et al. (2021) have applied the U-net architecture and (Liu et al., 2022) a combination of Residual Networks (ResNet) and long-short term memory (LSTM). Most of these methods, both based on traditional digital signal processing and neural networks, have in common that they process the input BCG data to emphasize the J-peaks. Moreover, for all deep learning models found in literature review, the discrete J-peak-events are represented as a time-series encoded with a binary masking. This binary masking, however, may lead to inaccurate peak detection, as further discussed in Section 3.

In this work, a method based on deep learning is proposed, which transforms a BCG measurement into a one-dimensional time-series, from which the discrete heartbeat events can be detected more precisely. The objective of the work is to compare the effect of different heartbeat encodings on J-peak detection accuracy using a fixed neural network architecture, and to compare the proposed method against state-of-the-art approaches. Thereby, we investigate improved encodings of heartbeat events in order to facilitate an optimized J-peak detection.

3 Problem formulation

The objective of J-peak detection in BCG is to estimate the timestamps of heartbeat events P in time-series data X. Consequently, a J-peak detector implements an algorithm that maps X to P. The input BCG are single- or multichannel measurements, which are represented as $X \in \mathbb{R}^{n \times k}$ with k being the number of channels and n the number of equidistantly sampled measurement points. In order to detect heartbeats present in X, each deep learning approach for J-peak detection mentioned in Section 2 models the J-peaks, a set of event timestamps $P = \{p \in \mathbb{R}\}$, with a binary event-hot encoding in the corresponding target time-series $y \in \mathbb{R}^n$. Additionally, a small area of interest around the peak with a width of τ may also be encoded with one. In the context of this work, y is referred to as surrogate signal. The majority of machine learning models M employed for this task are implemented as sequence-to-sequence models, which learn the following mapping Eq. 1:

$$M: X \mapsto y \tag{1}$$

Given that an ideal model M is unlikely to exist, the approximation of the target surrogate signal y is defined as $\hat{y} \coloneqq M(X)$. Subsequent to model inference, the estimated time-

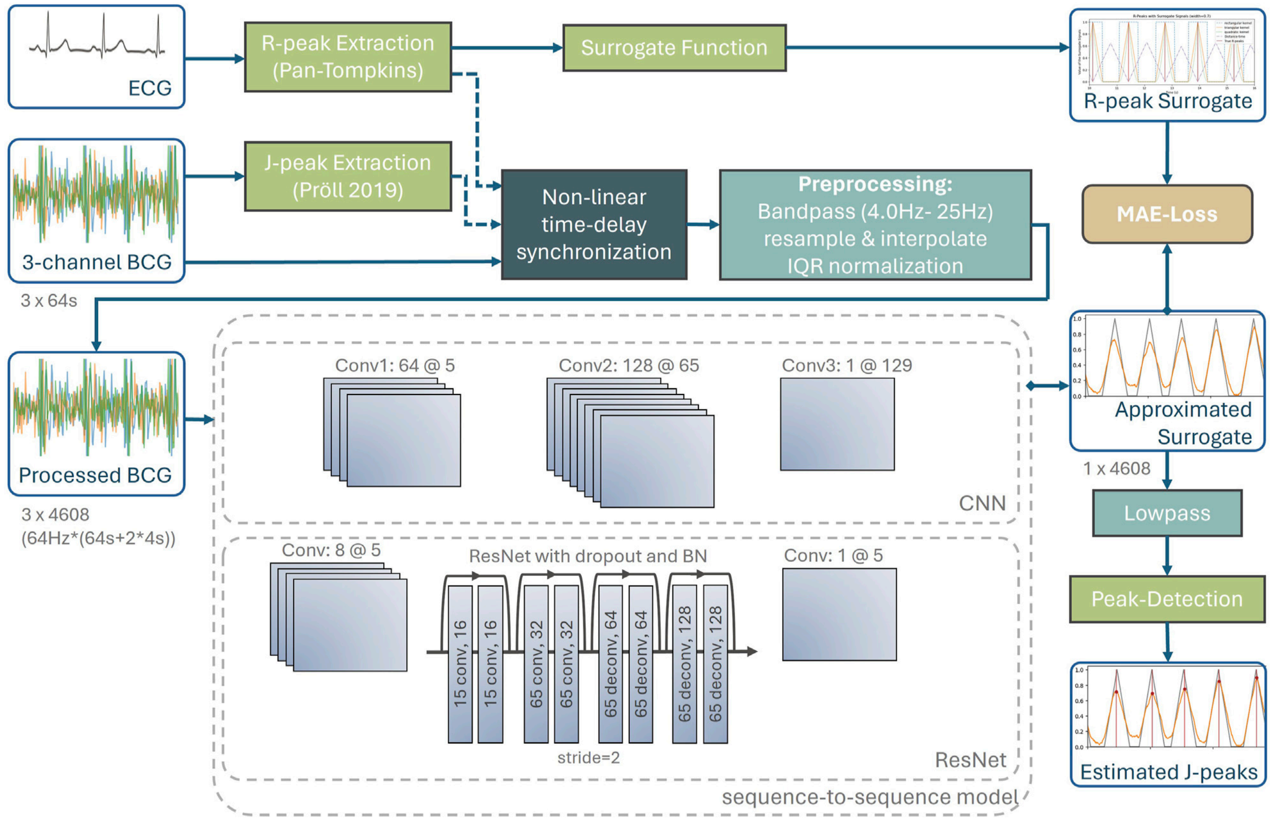


FIGURE 2
Complete processing pipeline of the proposed method, including synchronization with ground truth heartbeats accessed from the ECG, preprocessing, neural networks as well as post-processing to extract the J-peaks from the approximated surrogate signal.

series $\hat{y}$ is frequently subjected to post-processing in order to enhance the clarity and centering of the characteristic peak. This can be achieved by utilizing a low-pass filter or a moving average.

Finally, a classical peak detection algorithm is applied, with temporal and magnitude thresholds that have been optimized for the purpose of extracting heartbeats. Furthermore, the algorithm may also employ adaptive thresholds. Formally, the set of J-peaks $\hat{P} = \{p \in \mathbb{R}\}$ is extracted from the approximated surrogate signal $\hat{y}$ using a peak detection algorithm. The individual timestamps of the heartbeats, denoted by p, constitute the elements of the set.

We hypothesize that the state-of-the-art method of binary event-hot encoding of heartbeats may not be optimal for J-peak detection, resulting in imprecise event detection. The surrogate approximation $\hat{y}$ may be skewed after the low-pass post-processing, which could lead to inaccurate peak detection. Consequently, it is postulated that the encoding of heartbeats, designated as “kernels” in this paper, exerts a significant influence on the efficacy of subsequent J-peak detection.

To the best of our knowledge, no existing literature addresses the optimal kernel for the encoding of J-peaks with the aim of improving the precision of J-peak detection. A more general literature review, not limited to BCG data, revealed a single similar approach to event extraction from time-series data using deep learning (Azib et al., 2023). This work provides a theoretical framework for event detection in time-series for interval-based events, which was validated on fraud events. In this paper, we propose multiple encodings of heartbeats for generating the surrogate signal and empirically evaluate them with the aim of optimizing the J-peak detection $\hat{P}$ by aiding the model to learn $M: X \mapsto \hat{y}$. The method will be introduced in detail in Section 4.2.

4 Materials and methods

4.1 Data acquisition

The dataset comprises both the BCG and ECG, which were collected from 11 participants over a period of approximately 8 h during 17 nights of sleep. The data were collected as part of the Virtual Sleep Lab project, as detailed in Kranzinger et al. (2023). The electrocardiogram (ECG) was recorded using the *BrainAmp* Standard Amplifier (Brain Products GmbH, Germany), a laboratory-standard device known for high-quality recordings. The BCG was measured using an inertial measurement unit (IMU) with a 16-bit resolution (0.06 mg/LSB), and was mounted within the mattress centrally underneath the expected position of the subjects’ chests. The accelerations in three dimensions were sampled at a rate of 1,000 Hz. Subsequently, the signal was interpolated with a cubic spline and resampled at 64 Hz. In total, more than 140 h were measured, with an average interbeat interval (IBI) of 0.92 s.

4.2 Proposed method

For BCG, no ground truth J-peaks annotations exist. Therefore, the reference heartbeats from the simultaneous ECG measurement are accessed. The evaluated event detectors are quantified by comparing the estimated J-peaks with the related reference R-peaks using multiple evaluation metrics. Figure 2 depicts the complete processing pipeline of the proposed method. Within this section, each part of the pipeline, from high-precision data synchronization to the final extraction of the estimated J-peaks, is explained in detail.

4.2.1 High-precision data synchronization

A supervised machine learning setting requires ground truth heartbeat events that can be detected from ECG or the less accurate photoplethysmogram (PPG). According to the literature, the RJ-intervals, i.e., the time delay between R-peak and J-peak, vary typically between 180 ms and 240 ms. They may depend on certain factors, such as respiration, however, their changes are slow (Casanella et al., 2012).

In supervised machine learning setups for the J-peak detection utilizing R-peaks extracted from a synchronized ECG as target events, this would lead to varying RJ-intervals along the measurement. A practical approach would be to assume a constant RJ-interval per measurement as correct time-delay. However, the variation of 60 ms of the RJ-interval might lead to a suboptimal heart peak detection precision.

In this work, this issue is addressed by applying a non-linear time-delay synchronization for event-based time-series data (Schranz et al., 2024) between J-peaks and their corresponding R-peaks such that the slowly varying RJ-intervals can be approximated to zero for all J-peaks across the measurement.

As a first preprocessing step, a highly accurate time delay estimation between ECG and BCG is performed using the *nearest-advocate* package (Schranz et al., 2024). As this algorithm requires event-based time series data, the R-peaks were extracted from the ECG and the J-peaks from the BCG using a digital signal processing approach (Pröll et al., 2019). This algorithm was used because signal processing methods tend to be more robust on a new dataset, although there are likely to be more precise methods. The *nearest-advocate* package was also used to reduce non-linearities caused by non-linear clock drifts in the measurement systems and physiological variations that cause changes in RJ intervals.

The resulting dataset therefore has a three-dimensional BCG and corresponding R-peak events that are temporally aligned with the target J-peaks. This initial preprocessing step will make subsequent machine learning models more invariant to changing RJ intervals.

4.2.2 Preprocessing

Windows with a duration of 64-s are sampled from the subjects. Each BCG consists of three channels representing the x, y, and z-axis of the IMU. A bandpass filter with cutoff frequencies of 4.0 Hz and 25 Hz was applied to each of the three dimensions of the raw BCG signal. According to the standardization approach of (Gomez-Clapers et al., 2014), the high-pass cutoff-frequency should be lower, such as 1.5 Hz, but our hyperparameter optimization has shown that the pipeline yield improved results if signals below 4.0 Hz are omitted. The bandpass-filtered signal is then normalized using the interquartile range, which is less sensitive to outliers than standard z-score normalization.

The *nearest-advocate* time-delay estimation was then applied again within the range of −2 to +2 s to ensure a proper signal quality and synchronicity for BCG and ECG. Windows with a time-delay of 0.1 s or more between R-peaks and preliminary J-peaks were omitted from the training dataset. Although this discarded approximately 32% of the windows, no systematic bias was introduced because the synchronization between ECG and BCG is independent of signal quality and only the latter affects the quality of subsequent model training. All windows were used for the validation dataset.

4.2.3 Heartbeat encoding of the target R-peak

The core of the proposed approach is the special encoding of heartbeats by a surrogate signal, which is depicted in Figure 3. The reference heartbeats as extracted from an ECG are illustrated in red vertical lines, with four surrogate signals with different encodings. The surrogate signal is the target function that is learned by the neural network. The purpose of the surrogate signal is that the subsequent peak detection is more accurate on the surrogate approximation of the deep learning model.

Therefore, within the scope of the paper, three different kernel shapes, i.e., quadratic, triangular and rectangular, will be empirically evaluated with the aim of finding the most suited kernel for aiding the model to learn $\hat{y}$. All of these kernels share the properties of being symmetric around the reference heartbeat in the center at a maximum. Note that the rectangular encoding reflects the binary masking of the heartbeat with additional area of interest. Additionally, the distance-time encoding as proposed by (Vijayarangan et al., 2020) for the similar field of R-peak detection in ECG was evaluated. A surrogate signal generated by distance-time encoding has the property, that for any timestamp in y the value represents the distance to the closest heartbeat.

4.2.4 Deep learning models

Two network architectures are evaluated, both implementing a sequence-to-sequence approach, that estimates an equidistant time-series y referred to as surrogate signal.

4.2.4.1 Convolutional neural network

To approximate the surrogate signal, a convolutional neural network (CNN) with three layers of 64, 128, and one channel each is used. The kernels of 5, 65, and 129 are set to become increasingly wider. The model uses the ReLU activation function as a nonlinear mapping between layers and a batch size of 32. The training uses the Adam optimizer with a learning rate of 0.0001 for 40 epochs.

4.2.4.2 Residual Network

Additionally, a Residual Network (ResNet) is applied, as several works in the literature have used a variant of the related ResNet or U-Net architectures (Cathelain et al., 2020; Zhou et al., 2021; Liu et al., 2022). To do this, the initial convolutional layer has 8 channels with a kernel width of 5. Then a ResNet with two convolutional and two deconvolutional residual blocks, each with a step size of 2 was applied. Batch normalization and a 40% dropout were applied between each residual block. Finally, a single-channel

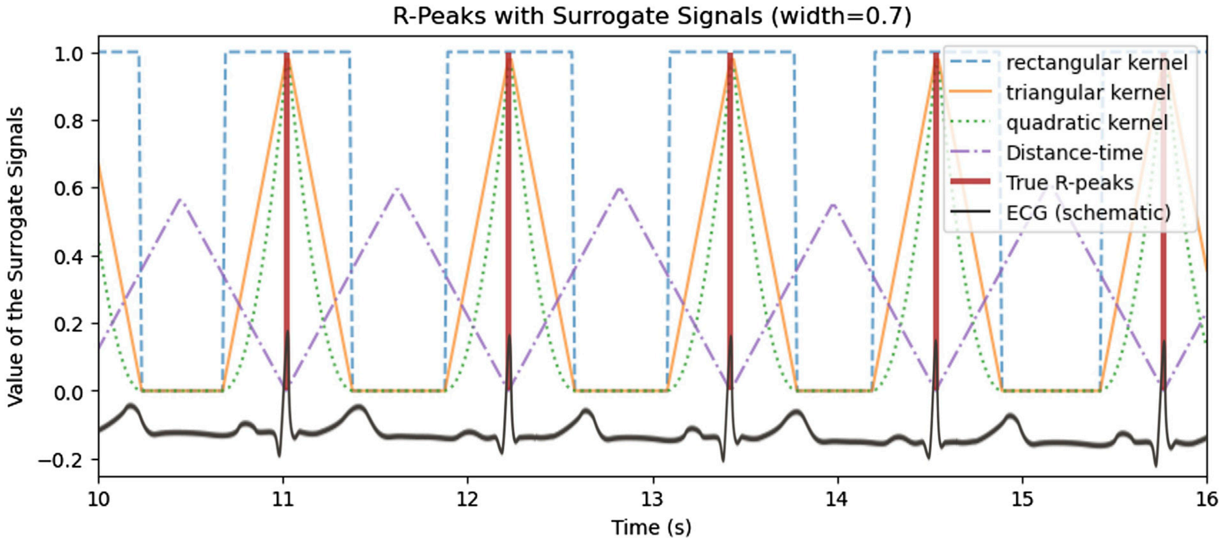


FIGURE 3
Encoding of the reference heartbeat events in (vertical red lines) with multiple surrogate signals with different encodings.

convolutional layer with a kernel width of 5 was applied. All other properties are the same as for the CNN.

4.2.5 Post-processing

Since the output of the network is an approximation of the surrogate signal $\hat{y}$, post-processing is necessary. For this purpose, the model output was smoothed with a second-order low-pass filter with a cut-off frequency of 7.5 Hz. Finally, a peak detection was performed using the *scipy*-package, with the following parameters: distance = 20, height = 0.01, and prominence = 0.1.

4.3 Evaluation

The comprehensive method evaluations in (Pröll et al., 2021) only target the accuracy of estimated mean heart rates within 8-s windows. However, the implicit aggregation of heartbeats to mean heart rate limits the applicability for further analyses. For example, the calculation of heart rate variability metrics relies on interbeat intervals (IBI) and reflects a person's physiological state and health (Shaffer and Ginsberg, 2017). In addition, most sleep stage classification algorithms rely on IBIs as input, i.e., interpolation of the temporal differences between successive heartbeats (Kranzinger et al., 2023).

Since the temporal detection of heartbeats is also important for the subsequent analysis of heartbeats, the detected J-peaks are evaluated using complementary criteria. The following metrics are used for comparison:

1. **HR MAE**: The estimation of heart rates within the full 64-s windows, with deviations reported as mean absolute error (MAE).
2. **HR MAE 8 s**: Estimation of heart rates within a reduced window of 8 s to establish comparability to the proposed methods in of Pröll et al. (2021).
3. **NAd_sym (ms)**: The Nearest-Advocate criterion (Schranz et al., 2024). This quantity is designed to measure the synchronicity between a pair of event-based time series. The resulting value after time-delay correction reflects the average distance between each detected J-peak, and its nearest reference R-peak. The algorithm is applied symmetrically and results are provided in milliseconds (ms). This measure considers only the temporal deviation of detected heart peaks.
4. **IBI MAE (ms)**: MAE between the interbeat interval (IBI) of the detected J-peaks and the reference IBI. Since the precision of the detection is high, the results are given in milliseconds (ms).

In the cross-validation procedure, the individual windows were grouped by subject in order to obtain an unbiased estimator for new subjects. The hyperparameters of the pre-processing, the model, the kernel width, and the post-processing were optimized using a grid search approach.

5 Results

Figure 4 shows the intermediate and final results of the proposed J-peak detection pipeline for an exemplary 10-s window. The 3-channel BCG measured by an acceleration sensor from which the heartbeats are to be detected is shown in the lower plot. In both plots, the surrogate signal is illustrated in gray. The heart beats are encoded with triangular kernels, with a center at the exact temporal position of the peak. Heartbeats that are closer together than the kernel width cause interference with super-position, as shown around second 11 in the plot.

The approximation of the surrogate signal (orange) is very similar to the target for clean BCG signals. For noisy episodes in BCG, the approximation remains higher for areas between heart beats, indicating a higher uncertainty of the deep learning

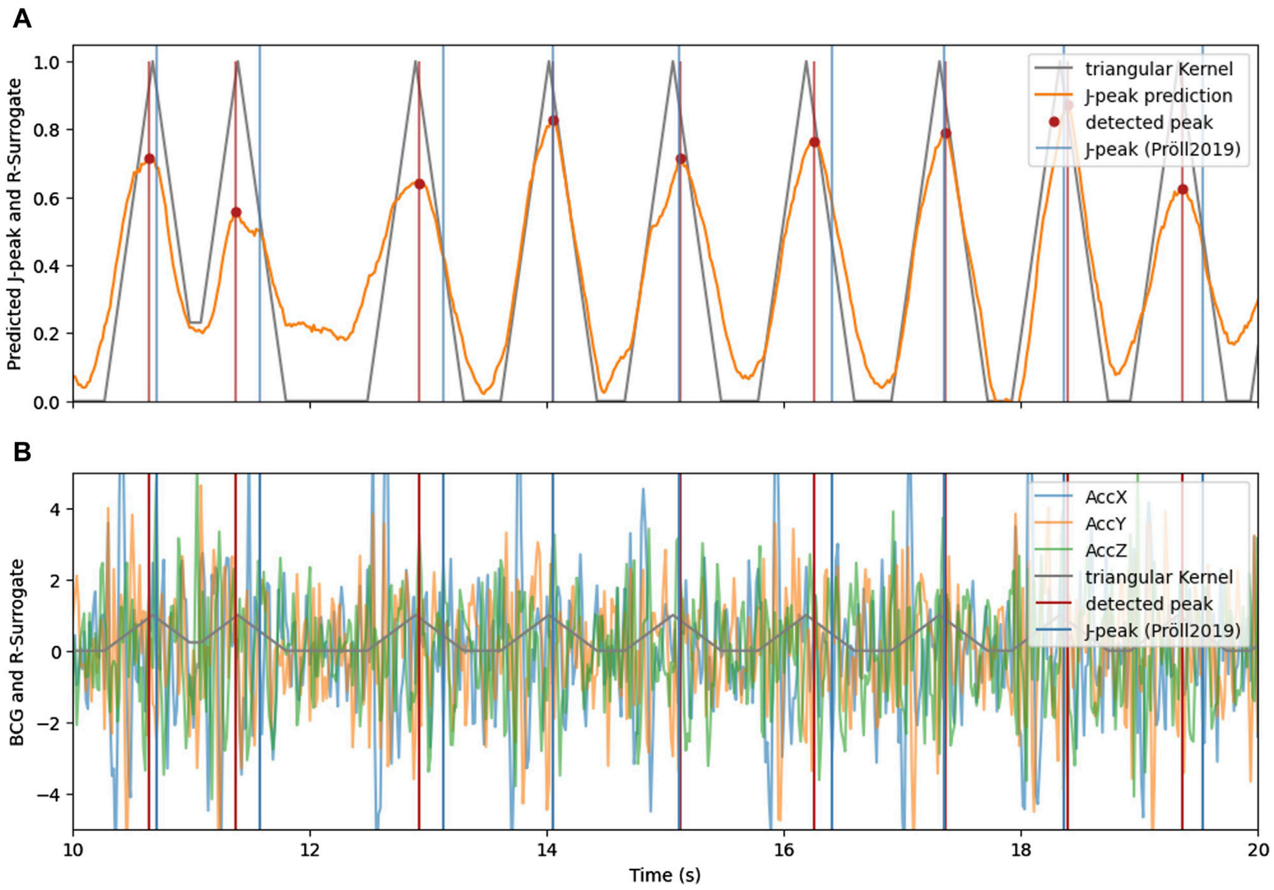


FIGURE 4
(A) The triangular signal (gray) with its approximation by the neural network (orange) and detected heartbeats using peak detection. **(B)** 3-axis BCG with indicated triangular surrogate signal (gray) and the J-peaks as detected by (Pröll et al., 2019) (blue) and with the proposed method and CNN-architecture (red).

TABLE 1 Results for each method, with reported mean and standard deviation across subjects. For each evaluation metric, the best result is indicated in bold.

Method/Model	HR MAE	HR MAE 8 s	NAd_sym (ms)	IBI MAE (ms)
Pino et al. (2017)	3.01± 2.3	3.83± 2.5	65.6± 18	57.6± 29
Choe and Cho (2017)	4.81± 3.9	5.74± 4.2	69.6± 21	101± 66
Brüser et al. (2011)	20.6± 4.7	22.5± 4.7	95.3± 15	215± 73
Pröll et al. (2019)	2.32± 1.5	3.18± 1.6	79± 15	78.9± 22
Pröll et al. (2021)	—	3.18± 0.54	—	—
CNN rectangular kernel	1.46± 0.83	2.00± 0.89	57.6± 7.7	44.2± 7.5
CNN triangular kernel	**1.1± 0.71**	1.52± 0.77	52.4± 10	40± 8.2
CNN quadratic kernel	1.31± 0.74	1.74± 0.81	53.8± 10	39.8± 8.9
CNN Distance Time	1.31± 0.88	1.73± 0.89	53± 9.7	41.4± 8.5
ResNet triangular kernel	1.22± 0.63	**1.38± 0.64**	**48.8± 8**	**27.9± 7**

estimation. However, even in the noisy area between the 10th and 13th second, the proposed method is able to accurately detect the heart beats.

Table 1 summarizes the results of the experiments. The four measures provided per method are described in Section 4.3. We evaluated our proposed methods against several existing methods,

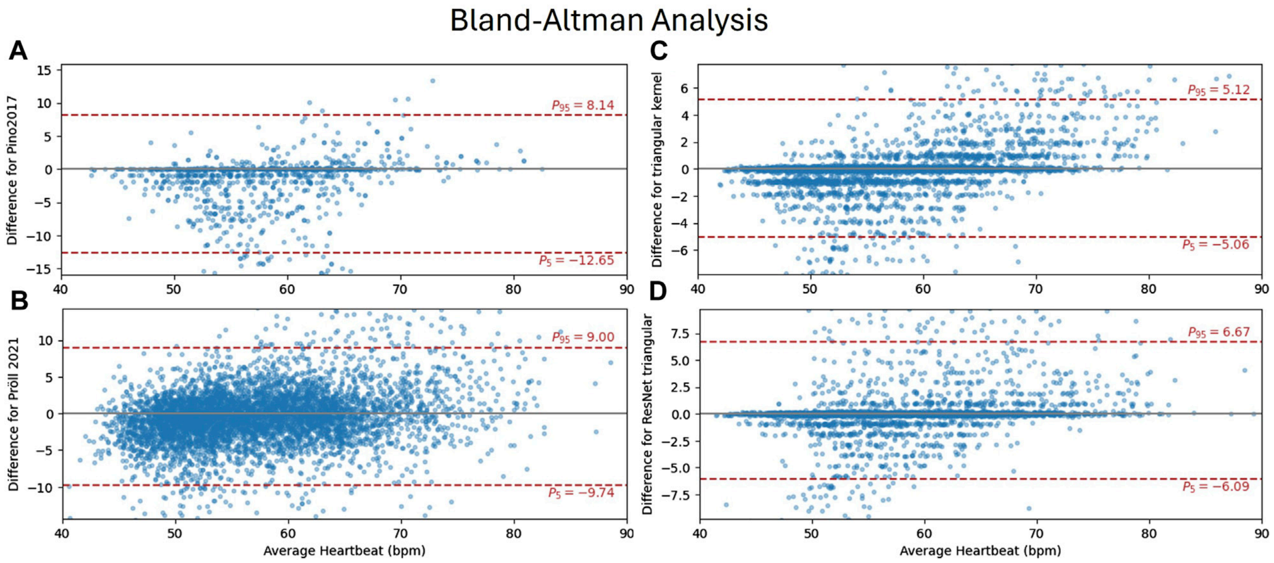


FIGURE 5
Bland-Altman analysis comparing ground truth heart rates (accessed from ECG) and the methods of (Pino et al., 2017) **(A)** (Pröll et al., 2021), **(B)** and the proposed CNN **(C)** and ResNet **(D)** with triangular kernel each within 8-s windows. The *y*-axis shows the residual of the estimate $hr_{ECG} - hr_{\hat{B}CG}$ in beats per minute (bpm), with limits of agreement (LoA) measuring their deviation.

including a state-of-the-art deep learning method of (Pröll et al., 2021). The classical digital signal processing methods were used as implemented in (Pröll et al., 2019) and with default parameters, to facilitate comparability. For Pröll et al. (2021), the best-performing model architecture on their dataset, the Modified CNN-GRUx2, was used and trained on our dataset for 75 epochs. Since this model estimates the mean heartbeat within 8-second windows, only this measure was reported. As this model estimates the mean heartbeat within 8-s windows, only this measure was reported.

The proposed method with a CNN has been evaluated with various kernels (rectangular, triangular, quadratic) as well as the distance-time encoding as proposed by (Vijayarangan et al., 2020) for generating a surrogate signal for ECG. In addition, the most suitable kernel was also evaluated with a ResNet architecture.

The results in Table 1 show the high performance of the proposed method for both for the heart rate estimation [*HR MAE* and *HR MAE 8s*) as well as the precision of the heart peak detection (*NAd_sym (ms)* and *IBI MAE (ms)*]. In particular, the triangular kernel yielded excellent results, with the quadratic kernel and the distance time modelling of the heart peak events being on par.

Figure 5 shows a Bland-Altman analysis for four selected algorithms in subplots a) to d), with their 95%-limit of agreement (LoA) in red (Pino et al., 2017). (a), accurately estimates heart rates for a high percentage of windows, as indicated by scatters close to zero (Pröll et al., 2021). (b) is the only method where the quantification of errors resulting from an integer number of heartbeats being incorrectly detected is not visible. This is because this method estimates heart rates directly as a regression task. The proposed methods on the right side (c and d) show very similar distributions, with more outliers for the ResNet (d). This results in a wider LoA, although the MAE of heart rates is lower. For both proposed methods, it can be seen, that multiple outliers are caused by false positives (not detected) events for heart rates around 50 bpm and false negatives (missed events) for higher heart rates.

6 Discussion

6.1 Method comparison

For each method, the accuracy of heartbeat estimation is better for the full 64-s window than for the reduced 8-s window. The difference of the respective estimations is small, given a reduced interval, by a factor of 8. This can be explained by the implications of falsely detected peaks: Any false positive or false negative peak detection will result in an incorrect number of events within the time range. As the estimated heart rate is calculated as the mean interval between beats, a wider interval is more robust against a missing or incorrectly detected beats. However, the wider interval increases the likelihood of one or more false peaks. Therefore, the difference between *HR MAE* and *HR MAE 8s* is quite small.

For the existing approaches (Pröll et al., 2019; Pröll et al., 2021), the accuracy in terms of heart rates is equivalent. Since Pröll et al. (2021) estimate heart rates for an 8-second window, only this measure can be reported. This method is characterized by a very small standard deviation across subjects, which may indicate an advantage of direct estimation of the target measure. The significant advantage of Pröll's deep learning method (Pröll et al. 2021) over his classical digital signal processing method, as reported in Pröll et al. (2019), could not be replicated on this dataset within these experiments. This could be explained by an insufficient number of training samples or a more challenging raw BCG signal. There could be explained by a too small number of training samples or a more challenging raw BCG signal.

In contrast, Pino et al. (2017) provides the most accurate detection of J-peaks in terms of *NAd_sym (ms)* and *IBI MAE (ms)*, but has a higher standard deviation and higher error in heart rate estimation. This suggests that the preprocessing method of Pino et al. (2017) may allow for more accurate event detection, but carries a higher risk of false positives or negatives. Furthermore, the transferability to other subjects may be more limited.

The method of Brüser et al. (2011) did not produce the expected results with an MAE of heart rates greater than 20 bpm. We believe that the default parameters were not successful for the given data set. The Bland-Altman analysis (not reported in this work) suggests a plausible range of deviations for heart rates of 50pm, increasing for higher heart rates.

The proposed methods excel for each metric. Regarding the evaluation of heartbeat encodings, the triangular kernel was the most successful for each metric. This indicates a more accurate heartbeat estimation as well as a higher precision of heart peak detection. The hyperparameter optimization suggests a kernel-width of 0.8 s for the triangular kernel and 1.2 s for the quadratic kernel. Furthermore, the inter-subject standard deviation is significantly lower, except for Pröll et al. (2021) concerning the *HR MAE 8s* measure. As expected, the rectangular (binary) encoding of heartbeats yielded solid, but inferior results in comparison to continuous surrogate signals with a single optimum at the heartbeat timestamp.

The evaluation of the more complex ResNet resulted in a higher precision for the J-peak detection, with a mean precision of less than 50 ms. It has been reported in the literature that the subsequent use of a recurrent network such as a GRU or LSTM (long short-term memory) improves the results. Developing the network architecture with a recurrent network or multi-head attention layer is a point for further development.

6.2 Kernel evaluation

The results demonstrate that the quality of heartbeat estimation depends significantly on the kernel type utilized to generate the surrogate signal. Therefore, the surrogate signal should be easily learnable by the sequence-to-sequence model and facilitate a precise subsequent peak detection during post-processing. It is hypothesized that the surrogate signal should be continuous and exhibit distinct, well-defined peaks in order to accommodate both properties. Empirical validation has demonstrated that the binary (rectangular) kernel, which yields non-continuous surrogate signal encodings without distinct maxima, is inferior for peak detection compared to other approaches. A visual comparison of the kernels is shown in Figure 3.

Furthermore, the authors hypothesize that the width of the kernel is of importance: On one hand, the width of the kernels should be broad enough to support the target heartbeats also under imprecise measurement or an imperfect dataset. In particular, if the R-peaks are utilized as ground truth heartbeat events, the kernel width must cover also varying RJ-intervals. On the other hand, the width of the kernels, i.e., the support within the surrogate signal, should be constrained, such that points in time not close to peaks have a default value such as zero. This is not the case for Distance Time encoding, where each value represents the time difference to the nearest peak. Moreover, another hypothesis is that the kernel shape should be symmetric in order to improve the learning of the corresponding peak within the signal. However, this was not directly evaluated, rather than indirectly by inference provided by larger kernel width.

Additionally, interferences of two adjacent kernels occur, if the respective peaks are closer together than the kernel width. This issue was particularly evident with the quadratic kernel, which showed optimal performance with a kernel width of 1.2 s, which is greater than the average interbeat interval. Note that the peak of the quadratic kernel is twice as sharp as that of the triangular kernel of the same width in terms of the first derivation. It is still not completely clear why the triangular kernel is superior to the quadratic kernel. The authors hypothesize that there is a trade-off in the kernel width between precision of the peak and inference with adjacent kernel shapes. This does not only involve the already optimized kernel width, but also its shape. It is acknowledged that further research is required to evaluate this trade-off systematically in order to identify the optimal kernel shape for heartbeat encoding.

6.3 Intra-subject variability

In the literature, both inter- and intra-subject variability are cited as a major challenge in the analysis of BCG signals (Choe and Cho, 2017; Sadek and Abdulrazak, 2021). In all the results above, the cross-validation was grouped by participants to access the inter-subject variability.

To analyze the within-subject variability only, the effects of classical cross-validation without grouping by subject were analyzed. It was found that the validation error is only about 10% (instead of 80%) higher than the training error. This indicates that the extraction of R-peaks is subject to high interpersonal variability and could be generalized very well across time intervals from the same subject. It is therefore expected that an increase in the number of subjects from the current 11 (with a total of 17 nights) will significantly improve the quality of the model. Alternatively, this work can support the development of data augmentation methods to improve model performance in an original measurement without additional subjects.

6.4 Limitations and further work

A limitation of this work is that the limitation of the dataset that was acquired from only eleven participants and over 17 nights. Section 6.3 suggests a much very high intra-subject generalization, however, the inter-subject generalization is significantly lower. In future work, a more comprehensive as well as open dataset will be utilized to get more robust results and to establish a more rigorous method comparison. Furthermore, data augmentation will be employed with the aim to improve the inter-subject generalization. Regarding the dataset, the anthropometries of the subjects should be critically reviewed, especially considering diversity and fit to potential target user groups.

Another limitation of the used dataset was the need for non-linear time synchronization due to temporal issues in the ballistocardiogram data acquisition. To solve this issue, a non-linear time synchronization as suggested in Schranz et al. (2024) was conducted between R-peaks from ECG and preliminarily extracted J-peaks from BCG using the method of Pröll et al. (2019). As the RJ-interval is non-zero and changing slowly (Casanella et al., 2012), the non-linear time synchronization may have compensated the varying time-delay between R-peak and subsequent J-peak. Therefore, this preprocessing step that was employed with the intention to compensate a measurement issue might have improved the suitability of using R-peaks as ground truth events for training a supervised neural network for J-peak extraction. The varying RJ-intervals are typically considered as constant in current literature, or this property is noted as open issue. In order to answer this research question, a very precisely synchronized dataset is required, which further increases the interest in continuing the current work on a larger and open dataset with a dedicated research focus on the preprocessing of BCG data.

Another discussion point for future research is the superiority of the triangular kernel over the quadratic kernel. Here, more experiments should be conducted to systematically evaluate the trade-off mentioned in section 6.2, in order to identify the optimal kernel shape for heartbeat encoding.

7 Conclusion

In this work, a method for improved heartbeat detection in BCG is proposed. This method uses various kernel shapes to generate surrogate signals that encode the discrete heartbeat events. Using deep learning models in a sequence-to-sequence setting, this surrogate signal is approximated, allowing a more precise J-peaks extraction in the subsequent peak detection. To the best of our knowledge, this is the first time temporal events are encoded with kernels to enable an improved event detection using a regression-based sequence-to-sequence model. Moreover, this work conducted the first comparison of various event encodings for event detection using deep learning.

The evaluation of different kernel shapes showed, that the simple triangular kernel provided the best surrogate signal to extract J-Peaks with a high precision. Using the proposed method, the MAE of the estimated heart rate was 1.1 s within 64-s and 1.52 s for an 8-s window, halving the precision of the best evaluated existing approach. Compared to a CNN architecture, ResNet architecture improved the accuracy of heartbeat detection, with a mean accuracy of less than 50 ms.

The findings may provide a foundation for enhanced health monitoring during sleep, including comprehensive heart rate variability analysis and sleep stage classification. This research further substantiates the potential of ballistocardiogram sensor technology for unobtrusive and cost-effective health monitoring.

There are several options for future development of the proposed method. Bland-Altman analysis provides quantified estimates. Optimizing the mean HR as an additional target measure with a hybrid loss in the deep learning model training could further improve the heart rate estimation. The overall estimation accuracy could be further improved by adding a recurrent layer after the CNN respectively ResNet architecture layers. A larger or openly available dataset could be used to perform a rigorous comparison of methods, including additional deep learning approaches. This could reduce the high inter-subject variability of the current evaluation. Furthermore, the use of data augmentation methods is well suited to address both intra- and inter-subject variability.

In conclusion, the proposed triangular and quadratic kernels for generating a surrogate signal to be approximated is a novelty and showed significant improvements for J-peak detection in BCG compared to existing solutions. This undermines our initial hypothesis that the design of the surrogate signals for target measures has a significant impact on the quality of the output. This approach can also provide a general solution for applying deep learning models, especially in the sequence-to-sequence setting, for event detection in univariate or multivariate time series data.

Data availability statement

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

Ethics statement

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

CS: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. CH: Conceptualization, Methodology, Software, Validation, Writing–review and editing. SM: Data curation, Writing–review and editing. DH: Data curation, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was thankfully supported by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology under Contract No. 2021-0.641.557.

Conflict of interest

Author DH was employed by Das Gesundheitshaus GmbH.

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.

References

Azib, M., Renard, B., Garnier, P., Génot, V., and André, N. (2023). Event detection in time series: universal deep learning approach.

Brownlow, J. A., Miller, K. E., and Gehrman, P. R. (2020). Insomnia and cognitive performance. *Sleep. Med. Clin.* 15, 71–76. doi:10.1016/j.jsmc.2019.10.002

Brüser, C., Stadlthanner, K., Waele, S. D., and Leonhardt, S. (2011). Adaptive beat-to-beat heart rate estimation in ballistocardiograms. *IEEE Trans. Inf. Technol. Biomed.* 15, 778–786. doi:10.1109/TITB.2011.2128337

Casanella, R., Gomez-Clapers, J., and Pallas-Areny, R. (2012) "On time interval measurements using BCG," in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York. 5034–5037. doi:10.1109/EMBC.2012.6347124

Cathelain, G., Rivet, B., Achard, S., Bergounioux, J., and Jouen, F. (2020) "U-net neural network for heartbeat detection in ballistocardiography," in International Conference of the IEEE Engineering in Medicine and Biology. New York. doi:10.1109/EMBC44109.2020.9176687

Choe, S.-T., and Cho, W. (2017). *Simplified real-time heartbeat detection in ballistocardiography using a dispersion-maximum method.* Biomedical Research-tokyo.

Craven, J., McCartney, D., Desbrow, B., Sabapathy, S., Bellinger, P., Roberts, L., et al. (2022). Effects of acute sleep loss on physical performance: a systematic and meta-analytical review. *Sports Med. Auckl. N.Z.* 52, 2669–2690. doi:10.1007/s40279-022-01706-y

Garbarino, S., Lanteri, P., Bragazzi, N. L., Magnavita, N., and Scoditti, E. (2021). Role of sleep deprivation in immune-related disease risk and outcomes. *Commun. Biol.* 4, 1304–1317. doi:10.1038/s42003-021-02825-4

Giovangrandi, L., Inan, O. T., Wiard, R. M., Etemadi, M., and Kovacs, G. T. (2011) "Ballistocardiography–A method worth revisiting," in Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2011, December 23, 2014. New york. 4279–4282. doi:10.1109/IEMBS.2011.6091062

Gomez-Clapers, J., Serra-Rocamora, A., Casanella, R., and Pallas-Areny, R. (2014). Towards the standardization of ballistocardiography systems for J-peak timing measurement. *Measurement* 58, 310–316. doi:10.1016/j.measurement.2014.09.003

Heise, D., and Skubic, M. (2010) "Monitoring pulse and respiration with a non-invasive hydraulic bed sensor," in Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2010, New York. 2119–2123. doi:10.1109/IEMBS.2010.5627219

Itani, O., Jike, M., Watanabe, N., and Kaneita, Y. (2017). Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. *Sleep. Med.* 32, 246–256. doi:10.1016/j.sleep.2016.08.006

Jakowski, S., Kiel, A., Kullik, L., and Erlacher, D. (2023). "Sleep to heal and restore: the role of sleep in the recovery and regeneration process," in *The importance of recovery for physical and mental health* (Routledge).

Kranzinger, C., Bernhart, S., Kremser, W., Venek, V., Rieser, H., Mayr, S., et al. (2023). Classification of human motion data based on inertial measurement units in sports: a scoping review. *Appl. Sci. 2023* 13 (15), 8684. doi:10.3390/APP13158684

Liu, Y., Lyu, Y., He, Z., Yang, Y., Li, J., Pang, Z., et al. (2022). ResNet-BiLSTM: a multiscale deep learning model for heartbeat detection using ballistocardiogram signals. *J. Healthc. Eng.* 2022, 6388445. doi:10.1155/2022/6388445

Nguyen, T., Qin, X., Dinh, A., and Bui, F. (2019). Low resource complexity R-peak detection based on triangle template matching and moving average filter. *Sensors* 19, 3997. doi:10.3390/s19183997

Pan, J., and Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng. BME-* 32, 230–236. doi:10.1109/TBME.1985.325532

Pino, E. J., Chávez, J. A. P., and Aqueveque, P. (2017). "BCG algorithm for unobtrusive heart rate monitoring," in *2017 IEEE healthcare innovations and point of care technologies (HI-poct)*, 180–183. doi:10.1109/HIC.2017.8227614

Pröll, S. M., Hofbauer, S., Kolbitsch, C., Schubert, R., and Fritscher, K. D. (2019) "Ejection wave segmentation for contact-free heart rate estimation from ballistocardiographic signals," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. New York, 3571–3576. doi:10.1109/EMBC.2019.8857731

Pröll, S. M., Tappeiner, E., Hofbauer, S., Kolbitsch, C., Schubert, R., and Fritscher, K. D. (2021). Heart rate estimation from ballistocardiographic signals using deep learning. *Physiol. Meas.* 42, 075005. doi:10.1088/1361-6579/ac10aa

Sadek, I., and Abdulrazak, B. (2021). A comparison of three heart rate detection algorithms over ballistocardiogram signals. *Biomed. Signal Process. Control* 70, 103017. doi:10.1016/j.bspc.2021.103017

Schranz, C., Mayr, S., Bernhart, S., and Halmich, C. (2024). Nearest advocate: a novel event-based time delay estimation algorithm for multi-sensor time-series data synchronization. *EURASIP J. Adv. Signal Process.* 46, 46. doi:10.1186/s13634-024-01143-1

Shaffer, F., and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Front. Public Health* 5, 258. doi:10.3389/fpubh.2017.00258

Tomaso, C. C., Johnson, A. B., and Nelson, T. D. (2021). The effect of sleep deprivation and restriction on mood, emotion, and emotion regulation: three meta-analyses in one. *Sleep* 44, zsaa289. doi:10.1093/sleep/zsaa289

Vijayarangan, S. R. V., Murugesan, B. S. P. P., Joseph, J., and Sivaprakasam, M. (2020) "RPnet: a Deep Learning approach for robust R Peak detection in noisy ECG," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). New York. 345–348. doi:10.1109/EMBC44109.2020.9176084

Yun, D., Lee, H. C., Jung, C. W., Kwon, S., Lee, S. R., Kim, K., et al. (2022). Robust R-peak detection in an electrocardiogram with stationary wavelet transformation and separable convolution. *Sci. Rep. 2022* 12 (1)), 19638–19710. doi:10.1038/s41598-022-19495-9

Zhou, T., Men, S., Liang, J., Yu, B., Zhang, H., and Luo, X. (2021). 1D U-Net++: an effective method for ballistocardiogram J-peak detection. *J. Mech. Med. Biol.* 21. doi:10.1142/S0219519421400583

frontiers | Frontiers in Network Physiology

TYPE Original Research
PUBLISHED 10 October 2024
DOI 10.3389/fnetp.2024.1467180

OPEN ACCESS

EDITED BY
Marta Varela,
Imperial College London, United Kingdom

REVIEWED BY
Alejandro Alcaine,
Universidad San Jorge, Spain
Marianna Meo,
Boston Scientific, Netherlands

*CORRESPONDENCE
Giorgia M. Bosi,
✉ g.bosi@ucl.ac.uk

RECEIVED 19 July 2024
ACCEPTED 24 September 2024
PUBLISHED 10 October 2024

CITATION
Lee MT, Martorana V, Md RI, Sivera R, Cook AC, Menezes L, Burriesci G, Torii R and Bosi GM (2024) On preserving anatomical detail in statistical shape analysis for clustering: focus on left atrial appendage morphology.
Front. Netw. Physiol. 4:1467180.
doi: 10.3389/fnetp.2024.1467180

On preserving anatomical detail in statistical shape analysis for clustering: focus on left atrial appendage morphology

Matthew T. Lee[1], Vincenzo Martorana[2], Rafizul Islam Md[3], Raphael Sivera[4], Andrew C. Cook[4], Leon Menezes[5], Gaetano Burriesci[1,6], Ryo Torii[1] and Giorgia M. Bosi[1]*

[1]UCL Mechanical Engineering, University College of London, London, United Kingdom, [2]Department of Economics Management, and Statistics, University of Palermo, Palermo, Italy, [3]UCL Medical Physics and Biomedical Engineering, University College of London, London, United Kingdom, [4]UCL Institute of Cardiovascular Science, University College of London, London, United Kingdom, [5]Biomedical Research Centre, National Institute for Health and Care Research, University College London Hospitals, London, United Kingdom, [6]Bioengineering Group, Ri.MED Foundation, Palermo, Italy

Introduction: Statistical shape analysis (SSA) with clustering is often used to objectively define and categorise anatomical shape variations. However, studies until now have often focused on simplified anatomical reconstructions, despite the complexity of studied anatomies. This work aims to provide insights on the anatomical detail preservation required for SSA of highly diverse and complex anatomies, with particular focus on the left atrial appendage (LAA). This anatomical region is clinically relevant as the location of almost all left atrial thrombi forming during atrial fibrillation (AF). Moreover, its highly patient-specific complex architecture makes its clinical classification especially subjective.

Methods: Preliminary LAA meshes were automatically detected after robust image selection and wider left atrial segmentation. Following registration, four additional LAA mesh datasets were created as reductions of the preliminary dataset, with surface reconstruction based on reduced sample point densities. Utilising SSA model parameters determined to optimally represent the preliminary dataset, SSA model performance for the four simplified datasets was calculated. A representative simplified dataset was selected, and clustering analysis and performance were evaluated (compared to clinical labels) between the original trabeculated LAA anatomy and the representative simplification.

Results: As expected, simplified anatomies have better SSA evaluation scores (compactness, specificity and generalisation), corresponding to simpler LAA shape representation. However, oversimplification of shapes may noticeably affect 3D model output due to differences in geometric correspondence. Furthermore, even minor simplification may affect LAA shape clustering, where the adjusted mutual information (AMI) score of the clustered trabeculated dataset was 0.67, in comparison to 0.12 for the simplified dataset.

Discussion: This study suggests that greater anatomical preservation for complex and diverse LAA morphologies, currently neglected, may be more useful for shape categorisation via clustering analyses.

KEYWORDS

statistical shape analysis, hierarchical clustering, left atrial appendage (LAA), atrial fibrillation, principal component analysis -PCA, clustering performance evaluation, segmentation (image processing)

1 Introduction

Shape is mathematically defined as "all the geometrical information that remains when location, scale and rotational effects are filtered out from an object" (Kendall, 1977). Shape analysis refers to a wide variety of mathematical/computational methods that may be used to identify the geometrical similarities and differences within a cohort of shapes. In recent years, there has been an adoption of statistical shape analysis (SSA) applications to human organs and vessels; this type of analysis is considered to be a step up from clinical morphometry due to greater objectivity and/or the identification and quantification of subtle geometrical information (Goparaju et al., 2022; Cerrolaza et al., 2019). Of the many such studied anatomies, the left atrial appendage (LAA), a natural closed-ended outgrowth of the left atrium (Figure 1A), stands out for its morphological complexity (in terms of both macro-shape and anatomical intricacy) and high diversity among different subjects.

The LAA is considered the origin of up to 91% of all left atrial thrombi during atrial fibrillation (AF) (Blackshear and Odell, 1996), the most common cardiac arrhythmia, affecting 59 million people worldwide and with increasing prevalence in older patients (about 20%–33% of risk above 45 years of age) (Linz et al., 2024). LAA shape category for thrombosis risk assessment is typically determined through clinical classification systems. The most used classification system defines 4 LAA types–chicken wing, windsock, cauliflower and cactus (Wang et al., 2010; Korhonen et al., 2015) (in debatable order of lower to greater thrombosis risk (Musotto et al., 2022; Bosi et al., 2018)) – that may be determined through morphometric measurements of LAA length, bending angle and number of lobes. However, this categorisation is commonly subject to clinical disagreement, with a study revealing consensus among three expert clinicians to be only reached in 28.9% of cases (Wu et al., 2019). Instead, as labelled in Figure 1B, more recent clinical (Yaghi et al., 2020) and SSA (Juhl et al., 2024; Ahmad et al., 2024) studies suggest that LAA categorisation may be primarily approached as chicken wing-like (characterized by high length and bending angle), and non-chicken wing-like.

Conventional LAA anatomical nomenclature (Barbero and Ho, 2017) is also displayed in Figure 1B for these two shapes: divided into ostium, neck, primary and secondary lobes, and trabeculae. The ostium refers to the entry-point for blood flow, dividing the left atrium from the LAA. The neck refers to the main body volume above the ostium, which connects to both the primary lobe and tip,

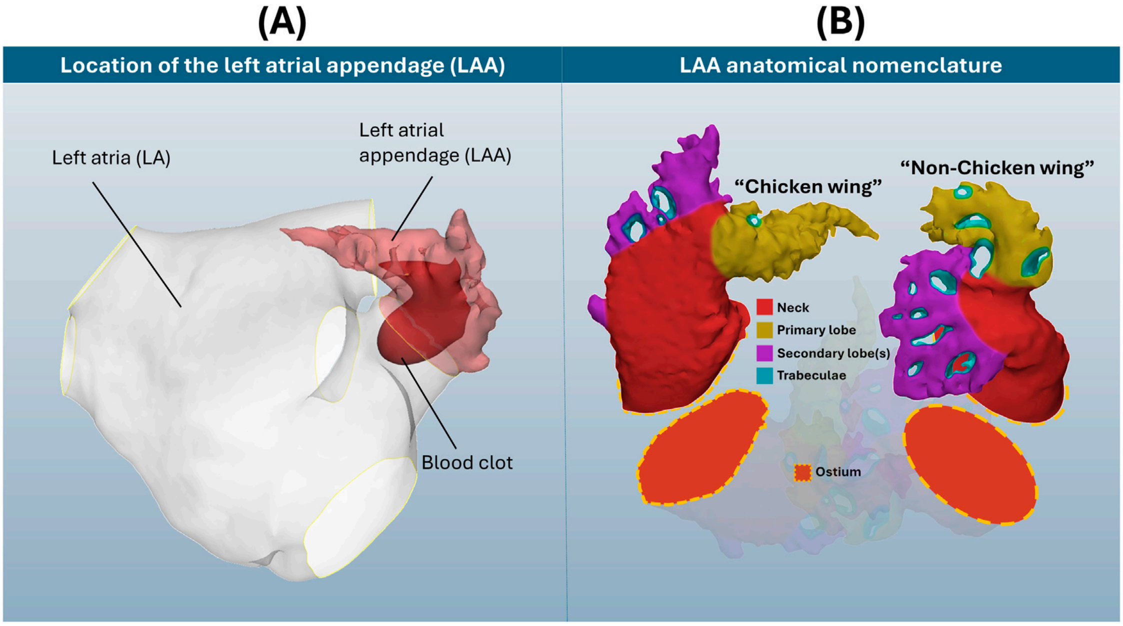


FIGURE 1
(A) Location of the LAA on the left atria, with blood clot representation. **(B)** Visual display of two selected LAA cases, with anatomical nomenclature of ostium, neck, primary and secondary lobes and trabeculae. Note how these example LAA anatomies differ considerably in both shape and detail, which does not include the full breadth of LAA morphological variation.

as well as secondary lobes along the LAA length. Trabeculae, appearing as holes that pass fully through the LAA blood pool, are devoid of blood flow due to pectinate muscle fibres connecting opposing walls of the appendage chamber. As seen in Figure 1B, LAA anatomies may differ considerably in both their macro-shape and intricate anatomical detail, i.e., trabeculae.

The inclusion of intricate anatomical details, such as LAA trabeculae, may further improve thrombosis risk assessment of LAA shape. In a normally functioning human heart, blood passes through the complex anatomy of the LAA in atrial diastole and washes out thoroughly during atrial systole. In AF conditions, the presence of these fine LAA morphological features has a much greater impact on the fluid mechanics–with greater thrombosis risk around trabeculae and towards the tips of lobes (Musotto et al., 2022). Furthermore, a recent computational study of LAA morphological parameters (Martorana et al.) suggests that the quantification of trabeculae may also be useful for shape analysis.

To better evaluate LAA shape than current clinical classification systems, studies have suggested various approaches towards in-depth LAA morphological understandings. Multivariate morphometric LAA shape analyses, to which haemodynamic measurements may also be combined (Pons et al., 2022), are useful to represent thrombosis risk with respect to simple shape measurements. More in-depth approaches, i.e., LAA SSA, have the additional advantage of preserving LAA anatomical variation in 3D mesh formats and outputting novel LAA categorical shapes (Goparaju et al., 2022; Juhl et al., 2024; Ahmad et al., 2024). SSA is based on the geometric correspondence of entire *shapes* (Kendall, 1977), where similarly shaped objects have greater correspondence (and *vice versa*), that is defined by the particular SSA implementation. LAA SSA representation for categorisation has been defined both explicitly with point correspondence (Goparaju et al., 2022; Juhl et al., 2024) and with implicit techniques (Goparaju et al., 2022; Ahmad et al., 2024). Building upon these SSA frameworks, such studies may then propose a computational categorisation of their LAA shape representations. This categorisation may be defined by hard (Ahmad et al., 2024; Goparaju et al., 2018) and soft (Juhl et al., 2024; Slipsager et al., 2019) clustering approaches, as well as non-clustering dimensionality reduction (Goparaju et al., 2022).

Despite multiple advances in LAA SSA (Goparaju et al., 2022; Juhl et al., 2024; Ahmad et al., 2024; Goparaju et al., 2018; Slipsager et al., 2019; Bhalodia et al., 2010; Bieging et al., 2021; Adams et al., 2023; Adams et al., 2022; Cates et al., 2015), no study has yet investigated the impact of intricate LAA morphological features such as trabeculae, surface roughness and tertiary lobe structure on LAA shape category definition. As key morphological components for the assessment of thrombosis risk, this study proposes that these features may also provide morphological information suitable for LAA shape categorisation (focussing on LAA SSA for clustering analysis). Therefore, this study compares LAA shape categorisation determined via hard clustering of LAA SSA models from fully trabeculated versus simplified datasets, suggesting that intricate anatomical detail (that includes trabeculations) provides additional analytical value for clustering LAA shape. This study does not aim to develop a new LAA classification scheme, but rather focus on the importance of preserving these anatomical details for clustering purposes.

2 Materials and methods

2.1 Image and mesh processing

85 clinical computerised tomography (CT) scans were used with informed consent by University College London Hospital (UCLH), consisting of non-AF patients examined for moderate coronary disease. The average participant age was 61.5 years, with 48 of the 85 of male sex. As this dataset is composed of control cases, not associated with thromboembolic risk, this study focuses on anatomical detail. Images are 512 × 512 pixels, with a pixel spacing of 0.488 mm × 0.488 mm, and a slice thickness of 0.625 mm acquired with the GE Discovery STE scanner. The manual segmentation protocol of full left atria was adapted from previous studies (Bosi et al., 2018; Capelli et al., 2012) to include measurements of contrast-to-noise ratio (CNR) and signal-to-noise ratio (SNR), following clinically recommended protocols (Marques et al., 2018), to ensure image (and hence later LAA shape) viability (Figure 2A). To summarise this process briefly, following calculation of CNR and SNR, 85 segmentation masks were generated in Mimics 24.0 (Materialise, Belgium) from the dye contrast threshold. These masks were manually processed by a segmentation expert to select only left atrial structures, including the LAA, pulmonary vein trunks and a mitral plane. After segmentation, each of the 85 left atria was evaluated by an expert cardiac anatomist to focus on chicken wing and non-chicken wing labels only. 21 LAAs were categorised as chicken wing and the remaining 64 as non-chicken wing.

Then, the full left atria, as surface models, were meshed using triangular elements of 0.5 mm edge length for subsequent LAA definition. To keep the process as objective as possible and preserve all anatomical details, the following approaches were taken. To ensure an objective definition of LAA ostial planes (conventionally defined through subjective manual assessment (Hołda et al., 2017)), a fully automatic LAA detection algorithm (Martorana et al.) was applied to all 85 segmented anatomies (Figure 2B). Briefly, this LAA detection method is based of distance analysis of computationally skeletonised left atria to automatically identify the LAA ostial plane, thus allowing LAA detection (Martorana et al.). To ensure normalisation across all detected LAAs, each mesh was then scaled to the same arbitrary volume (6,000 mm^3, close to the average mesh volume). Global registration of the detected LAAs was performed via the Super4PCS algorithm (Mellado et al., 2014) to a single case, followed by local iterative closest point (ICP) (Rusinkiewicz and Levoy, 2001) and multiview registration (Pulli, 1999) across the full dataset (Figure 2C). Local ICP and multiview registration were repeated until all possible pairs fell within alignment distance. For 2000 sample points describing each anatomy chosen at each ICP iteration, the chosen minimal starting distance was 10 mm, reduced iteratively so that 80% of the samples would lie at a distance lower than 0.5 mm. Up to this point, no LAA structural definition was lost (i.e., shapes are fully inclusive of objectively defined LAA ostia, full surface structure, bending and anatomical lobes and trabeculae), ensuring that LAA shapes match 'all the geometrical information that remains when location, scale and rotational effects are filtered out from an object, as per Kendall's definition of shape (Kendall, 1977).

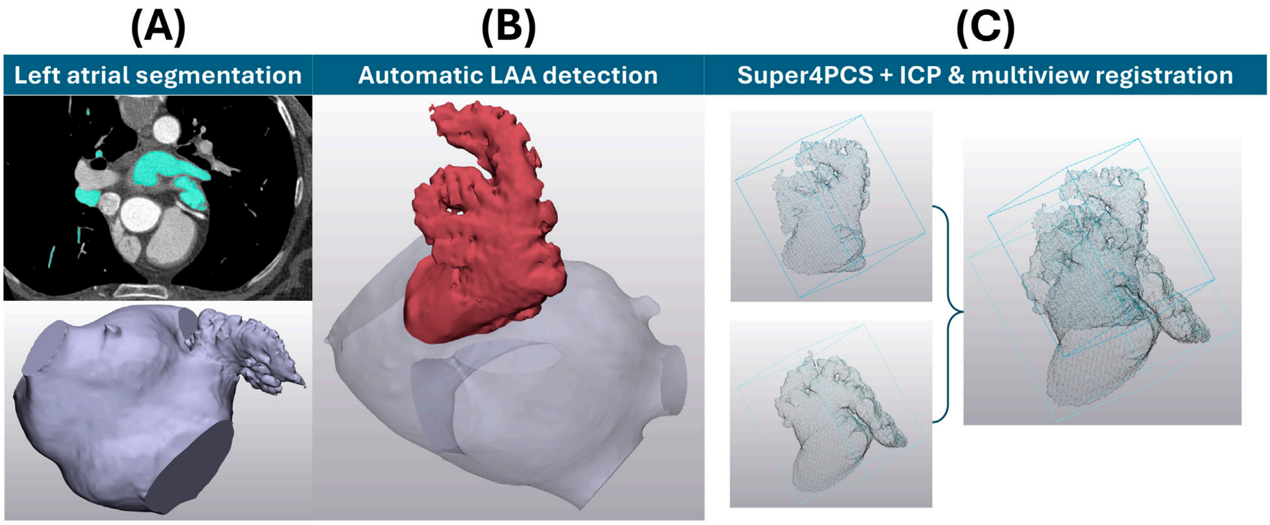


FIGURE 2
(A) LAA mesh acquisition and processing prior to SSA and clustering. The upper far left shows an example slice of the CT image stack to achieve the lower left atrial segmentation. **(B)** The LAA position determined through a fully automatic detection algorithm (Martorana et al.). **(C)** Two examples of LAA point clouds before and after alignment through Super4PCS registration, followed by ICP & multiview registration of all possible pairs.

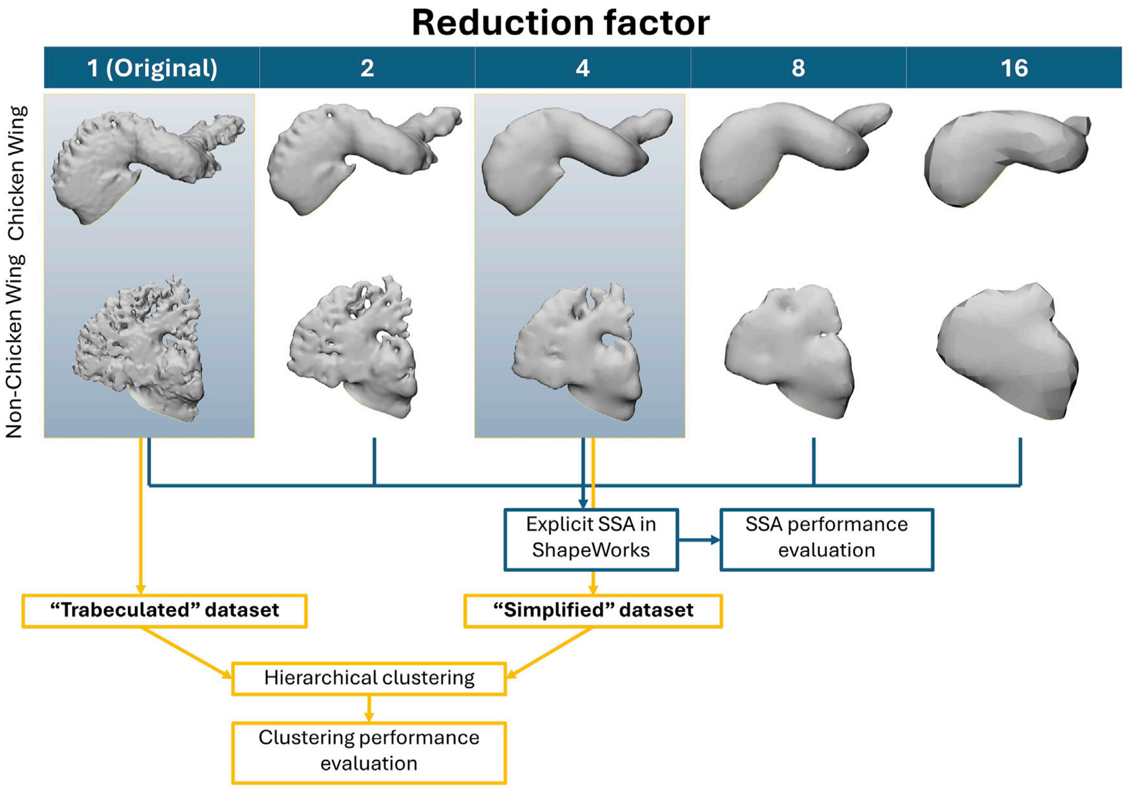


FIGURE 3
The simplified meshes (left to right) for two examples of LAA chicken wing and non-chicken wing morphologies from original trabeculated reconstruction, until full sample reduction. Note the visual loss in LAA trabeculae by 4-times sample reduction, and visual lobar definition loss by 8-times sample reduction. The data flow for the subsequent SSA and clustering methodology is also displayed in the bottom half of the figure–with SSA of all five datasets to determine SSA performance with greater sample reduction, followed by clustering comparison between the trabeculated dataset and one simplified dataset (4-times sample reduction).

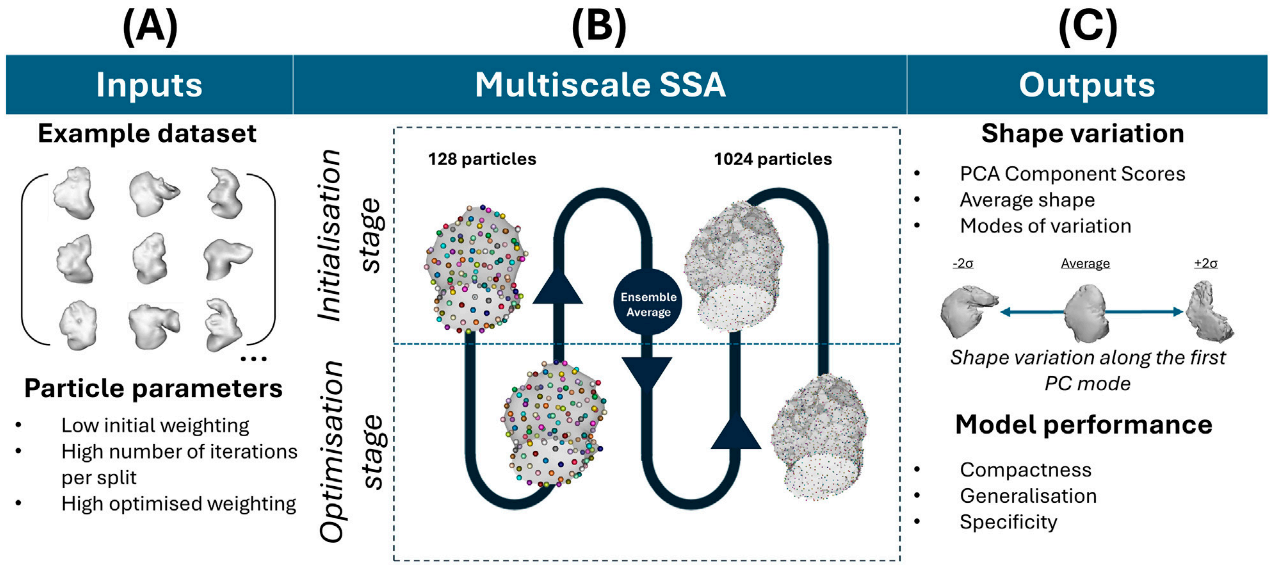


FIGURE 4
SSA workflow in ShapeWorks. **(A)** Refers to the input dataset and applied SSA parameters. **(B)** Refers to the SSA process, which is multiscale in the initialisation and optimisation of particle placements, with increasing particles' number **(C)** Refers to the outputs of the SSA (i.e., the PCA component scores after particle optimisation, the average shape and its variations) and the model performance evaluation metrics.

2.2 Simplified dataset generation

Based on the surface mesh generated for the 85 LAAs, simplified datasets of the registered LAA meshes were generated in MeshLab (Cignoni et al., 2008). Poisson surface reconstruction creates watertight surfaces from point sets with oriented surface normals, with set reconstruction depths corresponding to effective voxel resolutions (Kazhdan and Hoppe, 2013). To simplify the intricate meshes, a reduction factor of 2-times, 4-times, 8-times and 16-times was first applied to the point sets of LAA meshes in the original trabeculated dataset, with preservation of the original surface normals. To sequentially reduce intricate features such as trabeculae for surface reconstruction, the minimum sampling density was set as the reduction factor for each simplified dataset. To ensure less reconstruction bias due to the reduced number of points, the surface reconstruction depth d (which corresponds to solving on a voxel grid whose resolution is no larger than $(2^d)^3$ (Kazhdan and Hoppe, 2013)) was specified for each simplification as equal to 8, 7, 6 and 5. The simplified variations of the intricate dataset are shown in Figure 3: LAA surface reconstruction with 4-times reduction results in fully removed trabeculae; further reductions may lead to greater loss in lobar definition.

2.3 Statistical shape analysis

LAA SSA was applied with the explicit method in ShapeWorks software, the most commonly studied "off-the-shelf" software for LAA shape analysis (Goparaju et al., 2022; Goparaju et al., 2018; Bhalodia et al., 2010; Bieging et al., 2021; Adams et al., 2023; Adams et al., 2022). All analyses were run on an AMD Ryzen 9 7950X3D 16-Core Processor, 4201 Mhz, 16 Core(s), 32 Logical Processor(s). The workflow for the SSA is laid out in Figure 4 and described below. The SSA model was run with 1,024 particles in multiscale from 128 (so that the initialisation and optimisation of particle position is rerun for each particle split), and principal component analysis (PCA) of the final particle correspondences was computed. Parameter selection (featuring a low initial weighting of particle position with a very high iteration number per particle split, and a high final optimised weighting (Cates et al., 2017)) was iteratively adjusted to balance SSA model evaluation metrics of compactness, generalisation and specificity (Davies, 2002) as implemented by ShapeWorks (Shape Model Evaluation). Briefly, compactness score $C(n_m)$, the degree to which a model has captured the morphological variation within a dataset, is defined as the sum of the eigenvalues λ_i up to the selected number of PC modes n_m, summarised as: $C(n_m) = \sum_{i=1}^{n_m}\lambda_i$. Generalisation score $\hat{G}(n_m)$, a measure of a SSA model's ability to represent unseen shapes from a given dataset, may be quantified with the approximation error (Euclidean distance, in mm) between any held-out shape instance x_j and its corresponding SSA model reconstruction $\tilde{x}_j$, summarised as $\hat{G}(n_m) = \frac{1}{n_s}\sum_{j=1}^{n_s}\|x_j - \tilde{x}_j\|$, where n_s is the number of samples. Specificity score $\hat{S}(n_m)$, a measure of the plausibility of SSA model-generated shapes, may be computed as the approximation error (Euclidean distance, in mm) between any randomly sampled shape y_A and its nearest training sample x_i, summarised as $\hat{S}(n_m) \doteq \frac{1}{M}\sum_{A=1}^{M}\min_i \|y_A - x_i\|$ where M is the number of random samples taken. Final parameters were chosen to increase compactness i.e., the morphological variation captured by SSA, as desirable for clustering, despite lowered specificity and generalisation.

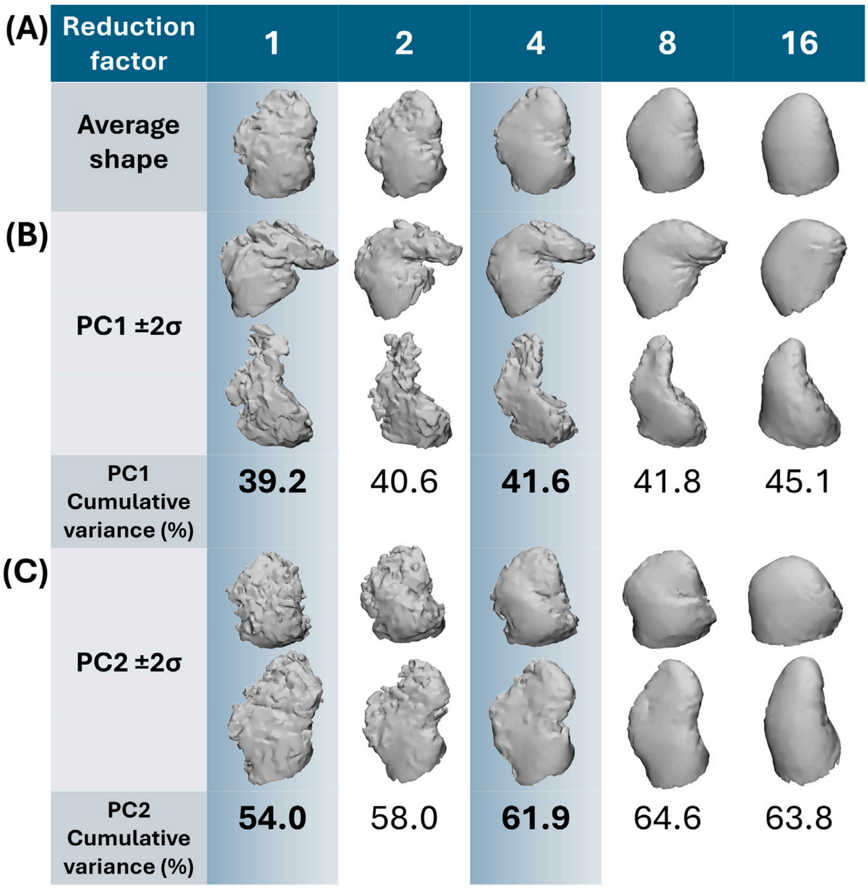


FIGURE 5
Shape variation captured by the first and second PC. In **(A)** the average shape with increasing reduction factor is presented. In **(B)** moving between 2 standard deviations on PC1 away from the average (±2σ) corresponds to chicken wing-like and non-chicken wing-like shape; with greater cumulative variance captured with increasing reduction with simpler shapes. In **(C)** moving between 2 standard deviations on PC2 away from the average (±2σ) corresponds more to secondary lobe size. Highlighted in blue are the two datasets (fully trabeculated and 4-times reduction) used for clustering comparisons.

2.4 Hierarchical clustering

For clarity, clustering analyses are only presented between the original trabeculated LAA surface versus the 4-times reduced dataset. 4-times reduction was chosen as it presents a clear reduction of fine anatomical detail loss, i.e., loss of trabeculae, but largely preserves secondary lobe structure. These two datasets are referred to as the “trabeculated dataset” versus the “simplified dataset” in the results section. Agglomerative hierarchical clustering was applied with MATLAB functions. Complete linkage and correlation distance were chosen; the former to ensure more compact clustering (Ezugwu et al., 2022) and the latter so that anti-correlated objects (i.e., chicken wing-like and non-chicken wing-like shapes) are as far apart as possible (van Dongen and Enright, 2012). The number of PCs accounting for 85% of the total variance (Cangelosi and Goriely, 2007) in the trabeculated dataset was retained for subsequent hierarchical clustering analysis, and the optimal number of clusters was calculated with the silhouette metric (Rousseeuw, 1987), to determine the cut-off value on the dendrograms. Clustering performance evaluation was performed with respect to the previously defined clinical labels, using the adjusted mutual information (AMI) score (Vinh et al., 2010) as the assessment metric. AMI is a measure of similarity (mutual information (MI)) between two labels of the same data, adjusted for chance. For two clusterings U and V:

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{\text{average}(\mathrm{H}(\mathrm{U}), \mathrm{H}(\mathrm{V})) - E(MI(U,V))}$$

3 Results

3.1 Statistical shape analysis

SSA took between 27.8 and 31.3 min to run for each dataset, regardless of anatomical intricacy. The results are presented in terms of visual geometric correspondence (Figure 5) and model evaluation score differences between the trabeculated and simplified datasets with increasing number of PCs (Figure 6).

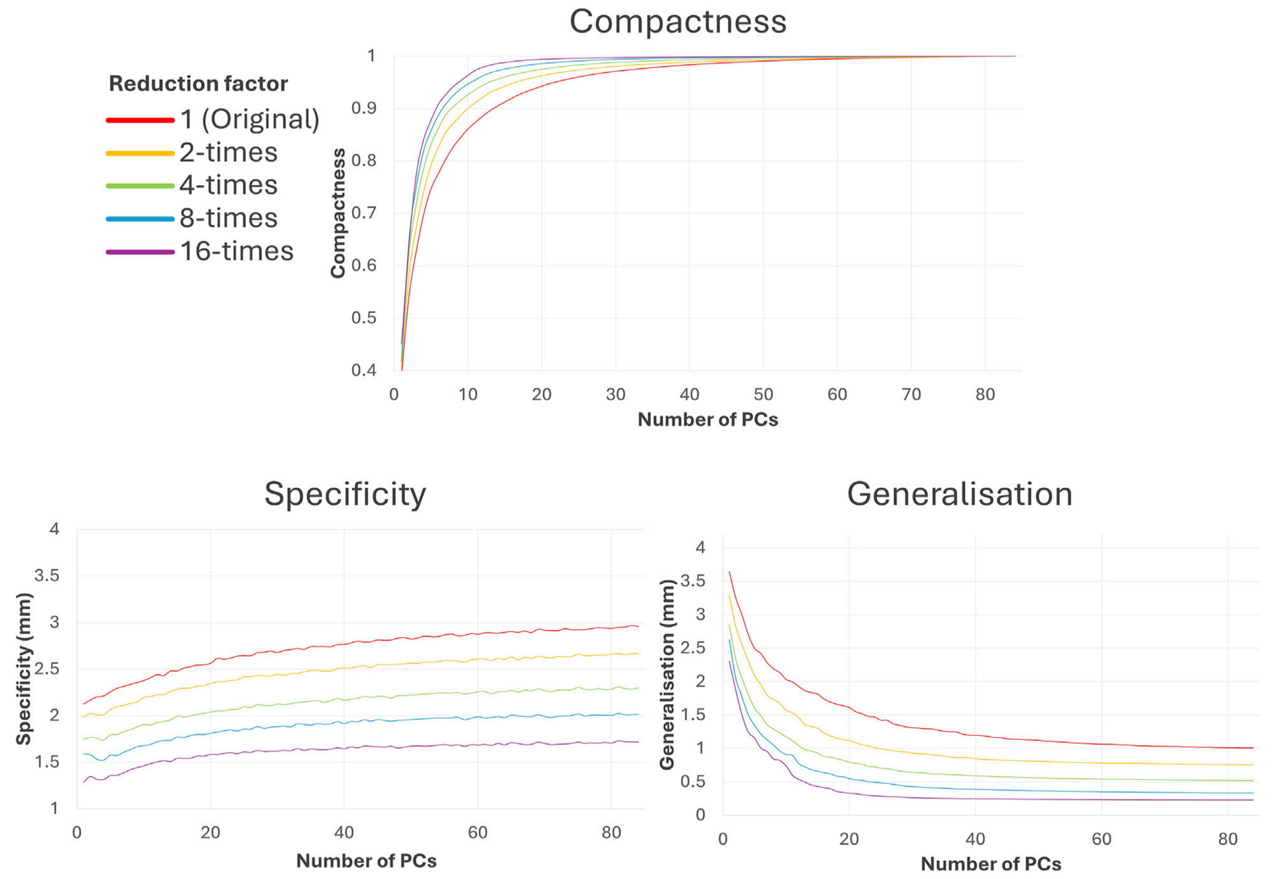


FIGURE 6
Difference in SSA model evaluation scores compared to the trabeculated dataset with increasing number of PCs. As shown, increasing shape simplification (with increasing reduction factor) increases the amount of morphological variance captured at lower PCs (compactness), decreases the Euclidean distance between a sample shape and its closest training sample (specificity) and improves unseen shape representation (generalisation).

3.1.1 Geometric correspondence & PCA

For both trabeculated and simplified datasets, most morphological variation (captured by PC1) is between chicken wing-like and non-chicken wing-like shape changes, which matches the observations of previous studies. As presented in Figure 5, moving along the PC1 axis corresponds with shapes more/less similar to the chicken-wing morphology. Moving down PC2 corresponds with smaller/larger secondary lobes. As may be expected, the anatomical detail present in SSA output shapes follows the degree of input shape simplification, with the increase of reduction factor corresponding to a loss in trabecular, surface and lobar definition matching the input datasets. For example, secondary lobes and trabeculae are no longer present by 8-times and 16-times reduction; and even primary lobe morphology is affected.

3.1.2 Shape model evaluation

As may be expected, utilising simpler input shapes translates to easier shape model evaluation. Increasing the reduction factor improves the associated compactness, specificity and generalisation in SSA, as seen in Figure 6. Greater compactness is preserved at lower PCs with increasing reduction factor, which also means that compactness score plateaus earlier. This implies that with simplified datasets, more morphological variation is captured for less PCs. The difference between compactness scores with reduction factor is non-linear; and increasing reduction factor has less effect following 4-times reduction. Specificity error decreases with increasing shape reduction and increases with the number of PCs, implying that more plausible shapes corresponding to each dataset may be generated with more simplified shapes. There is a roughly linear decrease in specificity with increasing reduction factor. Generalisation error (decreasing with the number of PCs) similarly decreases with increasing shape reduction and plateaus earlier, implying that the unseen shapes are better predicted with more simplified datasets. There is a slight non-linear decrease with increasing reduction factor, where greater reduction corresponds with less generalisation decrease.

3.2 Hierarchical clusters

Hierarchical clustering results are presented between the original "trabeculated" dataset, and the representative "simplified" dataset of 4-times reduction, with dendrogram results in Figure 7 and visualisation of the data distribution in Figure 8. 10 PCs were found to account for 86.1% of the cumulative variance for the trabeculated dataset, with the optimal number of clusters determined as 2 from a silhouette

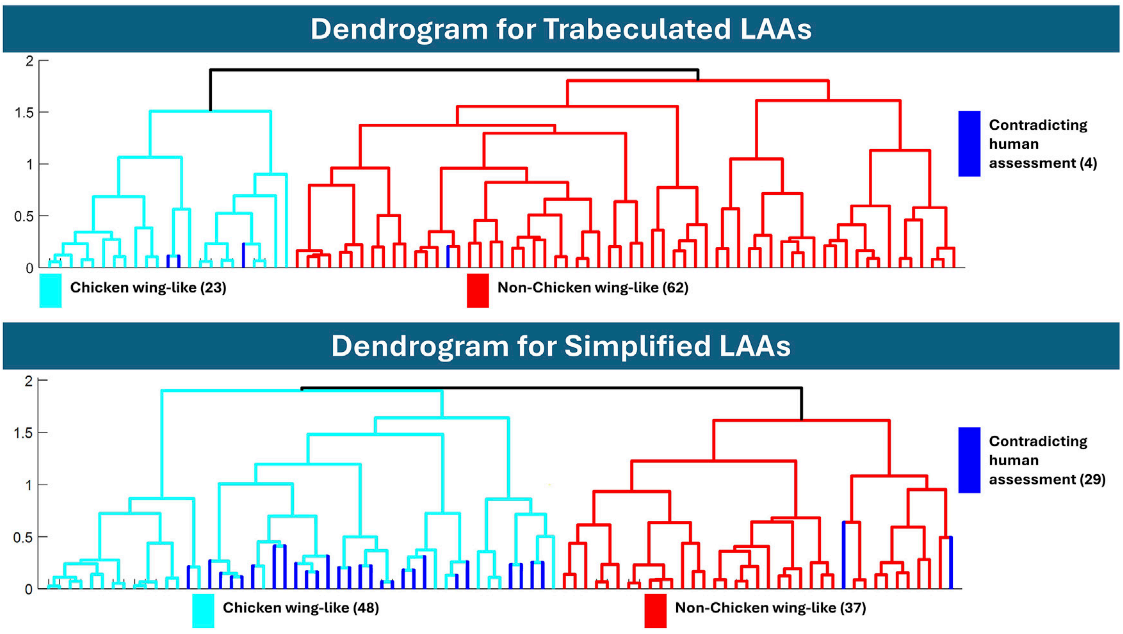


FIGURE 7
Dendrograms after hierarchical clustering of the trabeculated and simplified datasets. The dendrogram of trabeculated LAA morphologies indicated 23 as chicken wing-like, while the dendrogram of simplified morphology indicated 48. If categorised by a human expert, 21 LAAs are defined as chicken wing, suggesting that the trabeculated dendrogram is closer to human assessment.

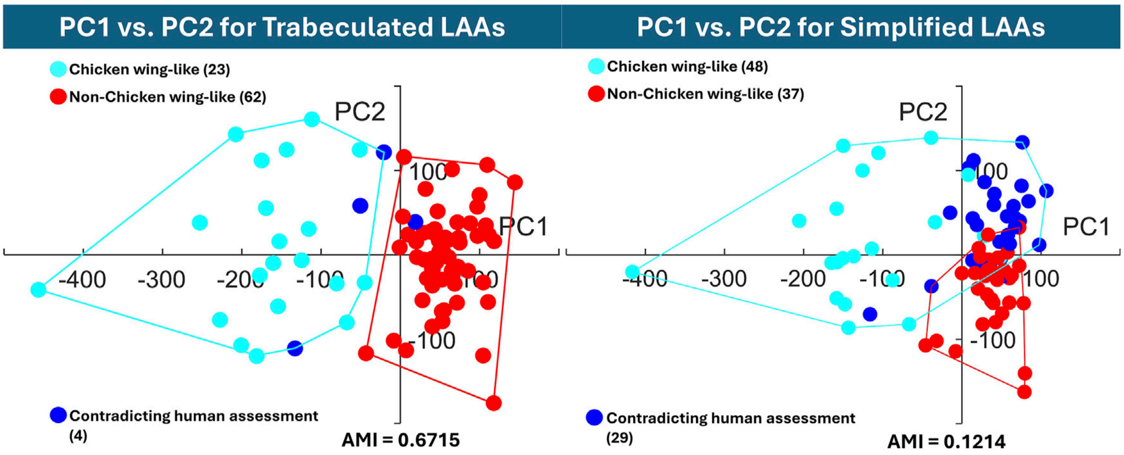


FIGURE 8
The hierarchical cluster assignments are displayed on the trabeculated PCA distribution (PC1 on the horizontal axis against PC2 on the vertical axis), with AMI according to earlier clinical labels. The graph of the trabeculated dataset shows clear cluster separation between chicken wing (PC1 in the negative direction) and non-chicken wing cases, while the simplified dataset displays high overlap.

score of 0.7948. Following the increase in shape model compactness with reduction factor, 10 PCs instead accounted for 92.6% of the cumulative variance for the simplified dataset, with the optimal number of clusters again determined as 2 from a silhouette score of 0.7458. For both datasets, dendrograms with the 2 optimal clusters are presented in Figure 7 and are highlighted on the trabeculated PCA distribution (showing PC1 against PC2) in Figure 8. Figure 8 also records the AMI score of each dataset to the clinical labels.

3.2.1 Dendrogram analysis

Comparing hierarchical clustering of fully trabeculated versus simplified morphologies, the dendrogram for the trabeculated dataset is closer to the current gold standard, i.e., human expert assessment, with 23 LAA morphologies being categorised into a chicken wing-like cluster (with four differences to clinical labels). While computed for 85% cumulative variance, the same clustering is achieved with 90% and 95% cumulative variance. In contrast, 48 LAA morphologies were categorised into the chicken wing-like cluster for the simplified dataset dendrogram (with 29 differences from clinical labels).

3.2.2 Cluster performance evaluation and data distribution

To quantitatively evaluate clustering performance, the AMI score was calculated for both the trabeculated and simplified clusters. With an AMI of 0.6715, the clustering of the trabeculated SSA model PCs is much closer to human assessment than the clustering of simplified SSA model PCs with a score of 0.1214. To visually present the clustering performance, the obtained hierarchical clusters are highlighted on their original PCA distributions for two axes (PC1 against PC2) in Figure 8. As shown, there is clearer cluster separation for the trabeculated dataset, where the chicken wing-like cluster is more dispersed than the non-chicken wing-like cluster. In contrast, the simplified dataset presents a strong overlap relative to human assessment. This overlap is mainly in the positive PC1 and PC2 directions, corresponding to non-chicken wing-like shapes and to smaller secondary lobes respectively, as presented in Figure 5C.

4 Discussion

4.1 Principal findings

With selected parameters for SSA and clustering, results suggest that LAA shape categorisation via hierarchical clustering performs better with preservation of full anatomical details (the "trabeculated dataset") than with trabecular detail loss (called the "simplified dataset"). While greater LAA anatomical simplification directly corresponds with better SSA model evaluation scores for compactness, specificity and generalisation (Figure 6), it was hypothesised that the loss of trabecular detail affects the preservation of morphological variation pertinent for LAA shape categorisation (Figures 3 and 5).

Between the trabeculated and simplified datasets, the improvement to SSA evaluation with reduction at the 10 PCs used for subsequent clustering is as follows: +0.065 compactness, −0.47 mm specificity and −0.86 mm generalisation (Figure 6). This is expected as the shape simplification process has led to a decrease in anatomical trabeculae and lobar definition that would have accounted for greater morphological difference between shapes. This implies that increasing anatomical simplification increases both the SSA model's ability to plausibly generate LAA shapes within simplified datasets and how well the model may generally represent unseen LAA shapes. However, as greater reduction by 8-times and 16-times visually affects even LAA lobar structure (Figure 3), it is thought that the greater anatomical simplification affects the geometric correspondence between shapes (Figure 5). Therefore, reduction by 4-times was selected as the simplified dataset for subsequent clustering comparisons. For visual comparison between PC1 and PC2 for these two datasets (Figure 5), PC1 captures chicken wing-like and non-chicken wing-like bending angle. PC2 instead describes LAA shapes with smaller or larger secondary lobes.

In contrast, increasing LAA reduction in SSA lowered clustering performance. The simplified model clusters, with a low AMI score of 0.1214, are mainly overlapping in the +PC1 and +PC2 quadrant (Figure 8), with 29 shapes being assigned differently to human assessment. This suggests that while + PC2 is associated with smaller secondary lobes, the inclusion of secondary lobe detail, e.g., trabeculae, better separates chicken wing-like shapes. On the trabeculated model clusters of Figure 8, the higher AMI score of 0.6715 corresponds with good cluster separation on the trabeculated PCA distribution, with only four shapes assigned differently to human assessment. This clustering is also more stable, with the same clusters being achieved for 90% and 95% cumulative variance. Therefore, these results may justify the preservation of intricate anatomical details, particularly LAA trabeculae, for shape categorisation with hierarchical clustering, despite improvements to pure SSA evaluation scores. In terms of computation time, SSA was less affected by the anatomical differences between datasets rather than the parameters chosen, taking between 27.8 and 31.3 min to run on the same AMD Ryzen 9 7950X3D 16-Core Processor, 4201 Mhz, 16 Core(s), 32 Logical Processor(s).

4.2 Broader research context

4.2.1 Clinical LAA shape categorisation schemes

Despite its popularity, conventional LAA classification (into four shape classes, chicken wing, cactus, cauliflower and windsock) is highly subjective, with a clinical study suggesting full shape category agreement between three observers was only reached in 28.9% of 2,264 cases (Wu et al., 2019). Other studies suggest the presence of 2–8 LAA classes depending on additional study aims. Some studies with only 2 shape classes separate LAAs into lower versus greater risk, based on the number of lobes (He et al., 2020) or with/without chicken wing-like bending (Yaghi et al., 2020). A clinical study suggests that LAA morphologies are instead combinations of up to 8 qualitative lobe shapes, preferring visual lobe classification instead of general shape categorisation (Beutler et al., 2014). With special focus on quantitative anatomical measurements not just of the LAA but of adjacent structures and the body, LAA clinical classification may even extend to 7 shape categories with 6 subtypes (Li et al., 2015). These studies highlight the sheer diversity of LAA shape complexity even without consideration of finer anatomical details, and the need for an objective shape categorisation from clustering analysis of SSA models, as employed here. As our study currently focusses on chicken wing-like and non-chicken wing-like shape categorisation, this is more similar to the simplified clinical categorisation with/without chicken wing-like bending (Yaghi et al., 2020), but without needing human intervention.

4.2.2 Applications of anatomical detail in LAA meshes

While clinical categorisation schemes are useful for simplified understandings of the connection between LAA morphology and thrombosis risk, the subjectivity of such classifications (Wu et al., 2019) may subsequently lead to inaccurate risk stratification. Furthermore, clinical categorisation typically does not consider the impact of intricate anatomical details, which may be difficult to measure manually.

A more in-depth comprehension of the LAA shape-haemodynamic relationship requires 3D LAA meshes, which provide 3D anatomical variation that is useful for computational modelling. While many studies do not consider intricate anatomical details, studies that do consider such impact (Musotto et al., 2022) suggest that trabeculae play an important role in LAA haemodynamics, by reducing LAA blood washout.

4.2.3 Other LAA SSA studies

Previous LAA SSA studies aim to objectively define LAA shape categories beyond current clinical capabilities, although no study to date is built from LAA morphology with full anatomical detail preservation. Explicit LAA SSA is typically based on the point distribution model (PDM) (Cootes et al., 1995), where correspondence between shapes is defined by the automatic placement of points across surfaces. The most studied optimisation scheme for LAA explicit correspondence is the entropy scheme used in ShapeWorks (Cates et al., 2017) (applied on both the LAA only (Goparaju et al., 2022; Goparaju et al., 2018) and for the conjoint left atria with LAA (Bieging et al., 2021; Adams et al., 2023; Adams et al., 2022; Cates et al., 2015)), where increasing particle correspondence may be iteratively initialised and optimised with regularisation parameters. Alternatively, explicit LAA SSA studies may determine initial point correspondence through Markov Random Field regularisation (Juhl et al., 2024; Slipsager et al., 2019) of the correspondence vector fields between source and target shapes (Paulsen et al., 2003). LAA SSA may also be applied implicitly on both the LAA only (Goparaju et al., 2022; Ahmad et al., 2024; Goparaju et al., 2018) and for the conjoint left atria with LAA (Corrado et al., 2020). Implicit approaches typically rely on the optimisation of deformations in a Riemannian space to warp shapes into others (Bône et al., 2018; Hartman et al., 2023). Established frameworks, such as Deformetrica (Bône et al., 2018), have been used (Goparaju et al., 2022; Goparaju et al., 2018), and recent works have also experimented with dedicated frameworks (Hartman et al., 2023) applied specifically to the LAA (Ahmad et al., 2024). However, to our knowledge, such methods do not allow the high complexity of the LAA surfaces to be considered. Of all the studies mentioned, the most recent advances in LAA SSA (Juhl et al., 2024; Ahmad et al., 2024) have focused mainly on chicken wing and non-chicken wing shape classification, proposing that more in-depth shape categorisation may fit within this overarching division.

Lower LAA morphological complexity may be a consequence of lower image input resolution (Cates et al., 2015), or that images have been intentionally "downsampled" to reduce noise (Juhl et al., 2024) e.g., for deep learning segmentation (Juhl et al., 2024; Ahmad et al., 2024). As discussed earlier, inclusion of fine LAA morphological detail not only improves thrombosis risk assessment of AF patients (Musotto et al., 2022) (the primary reason for LAA shape analysis) but may also be discriminatory for shape categorisation (Martorana et al.). Therefore, previous SSA studies may be limited in clinical applicability.

4.2.4 Computational categorisation methods in LAA SSA

Current shape categorisation methods in LAA SSA may utilise hard and soft clustering approaches, as well as non-clustering dimensionality reduction. Hard clustering on LAA SSA has been approached with k-means (Goparaju et al., 2018) and hierarchical clustering with additional multidimensional scaling (Ahmad et al., 2024), in comparison to our study focussing on hierarchical clustering only. A hard clustering approach may be more useful for the analysis discussed in this study, where categorisation between chicken wings and non-chicken wings should present less overlap. Soft clustering of LAA SSA, where overlap may be considered, has been approached with Gaussian Mixture Modelling (Juhl et al., 2024; Slipsager et al., 2019). Alternatively, another study suggests the use of t-stochastic Nearest Neighbour Embedding for their LAA SSA (Goparaju et al., 2022), which may be useful to display trends not visible with clustering methodologies.

To our knowledge, no other LAA SSA studies have presented the numerical efficacy of their shape categorisation with respect to human evaluation, so this is difficult to compare to other studies. In this work, AMI was chosen to evaluate cluster performance over rand-index scoring as unequal cluster sizes were expected (van der Hoef and Warrens, 2019), with only 21 of the 85 segmented LAAs having been expertly classified as chicken wing morphology earlier. Furthermore, as an adjustment of the regular mutual information metric, chance clustering assignments are accounted for.

4.3 Strengths and limitations of study

The applicability of the proposed LAA SSA model and clustering is limited by the analysed number of anatomies in the original dataset. This is particularly important for highly diverse anatomies such as the LAA, where it is highly likely for morphologies to demonstrate categorical variance beyond subjective clinical classification, even without considering fine anatomical details. In comparison with other LAA works, the number of LAAs utilised in our study (85 in total) lies between other studies, which can vary from 20 (Ahmad et al., 2024) to 130 (Goparaju et al., 2022). However, no other SSA study to our knowledge has preserved our level of LAA anatomical detail, which is the basis for this study.

Some limitations are related to operator-dependent steps in our workflow. Firstly, the manual left atrial segmentation (prior to fully automatic LAA detection) requires user definition of contrast threshold (aided by the additional mathematical CNR measurement protocol) and human effort and time to ensure segmentation is not affected by unwanted imaging artefacts. The second operator-dependent step is the clinical classification used to obtain the clinical labels to which clustering is compared in AMI scoring; clinical subjectivity was minimised in this study by focusing clinical labels to chicken wing versus non-chicken which is known to present the greatest morphological difference of bending angle (Yaghi et al., 2020). Two of the aforementioned LAA SSA studies have aimed to tackle the segmentation problem via deep learning (Juhl et al., 2024; Ahmad et al., 2024); however, as already stated, these works do not fully capture the same level of anatomical detail, presenting very smooth meshes, i.e., without trabeculae. Furthermore, the fully automatic LAA detection of the ostial plane utilised in our study may be further advantageous over both these studies that either cut the shape where it is narrowest (Juhl et al., 2024) (which describes an anatomical region generally different from the ostium definition) or perform manual clipping of left atrial meshes (Ahmad et al., 2024).

Finally, it should be noted that while a pixel spacing of 0.488 mm from CT is high for conventional clinical scans, even higher resolutions exist for alternative *ex-vivo* imaging-based studies e.g., microCT, synchrotron-based or photon-counting CT imaging. This study indicates that clustering of anatomies acquired with smaller pixel spacing performs significantly better than lower resolutions, which suggests that even higher resolution scan data could improve the results further. To increase the reliability and statistical significance of this work, it would be beneficial to incorporate more LAA morphologies in the SSA performed; however, it was not possible to include datasets acquired from publicly accessible databases (Atria Segmentation Challenge 2018; Karim et al., 2018) as they either did not match the imaging modality and/or the required resolution.

5 Conclusion and future works

SSA studies for clustering analysis of highly diverse anatomies, particularly the human LAA, may suffer from analytical disparities and therefore clinical relevance due to major differences in anatomical detail preservation. Following robust image and mesh processing, this study applies SSA and clustering analysis to 5 LAA datasets (each composed of 85 shapes), sequentially reduced in anatomical detail. While evaluation scores of SSA metrics of compactness, specificity and generalisation suggest lower resolutions may improve LAA shape representation of such simplified anatomies, it should also be recognised this better representation may not correlate with improved LAA shape categorisation. The cluster performance scores suggests that clustering for LAA shape categorisation benefits from greater preservation of anatomical detail (beyond the level conventionally preserved in LAA SSA). Future work could improve upon binary categorisation (i.e., chicken wing-like vs. non-chicken wing-like) by adjusting the dendrogram cut-off thus leading to smaller morphological sub-groups. In preserving trabeculae, this study advances towards connecting SSA anatomical detail to thrombosis risk categorisation.

Data availability statement

The datasets presented in this article are not readily available because of concerns regarding participant/patient anonymity. Requests to access the datasets should be directed to the corresponding author.

Ethics statement

This study was carried out in accordance with the recommendations of the South East Research Ethics Research Committee, Ayelsford, Kent, United Kingdom. All patients/participants participate in this study in accordance with the Declaration of Helsinki. The protocol was approved by the South East Research Ethics Research Committee, Ayelsford, Kent, United Kingdom. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

ML: Conceptualization, Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. VM: Data curation, Methodology, Writing–original draft, Writing–review and editing. RM: Data curation, Methodology, Writing–original draft, Writing–review and editing. RS: Investigation, Methodology, Writing–original draft, Writing–review and editing. AC: Data curation, Writing–original draft, Writing–review and editing. LM: Data curation, Writing–original draft, Writing–review and editing. GaB: Supervision, Writing–original draft, Writing–review and editing. RT: Supervision, Writing–original draft, Writing–review and editing. GiB: Conceptualization, Funding acquisition, Supervision, Writing–original draft, Writing–review and editing, Data curation, Investigation, Methodology.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors gratefully acknowledge the support from the Royal Academy of Engineering Fellowship (RF/201920/19/221).

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 author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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

Atria segmentation Challenge (2018). Atria segmentation Challenge — cardiac atlas project Available at: https://www.cardiacatlas.org/atriaseg2018-challenge/.

Adams, J., Khan, N., Morris, A., and Elhabian, S. (2022). "Spatiotemporal cardiac statistical shape modeling: a data-driven approach," in *Statistical atlases and computational models of the heart regular and CMRxMotion Challenge papers*. Editors O. Camara,

E. Puyol-Antón, C. Qin, M. Sermesant, A. Suinesiaputra, and S. Wang (Cham: Springer Nature Switzerland), 143–156. (Lecture Notes in Computer Science).

Adams, J., Khan, N., Morris, A., and Elhabian, S. (2023). Learning spatiotemporal statistical shape models for non-linear dynamic anatomies. *Front. Bioeng. Biotechnol.* 11, 1086234. doi:10.3389/fbioe.2023.1086234

Ahmad, Z., Yin, M., Sukurdeep, Y., Rotenberg, N., Kholmovski, E., and Trayanova, N. A. (2024). Elastic shape analysis computations for clustering left atrial appendage geometries of atrial fibrillation patients. *arXiv*. doi:10.48550/arXiv.2403.08685

Barbero, U., and Ho, S. Y. (2017). Anatomy of the atria: a road map to the left atrial appendage. *Herzschrittmacherther Elektrophysiol* 28 (4), 347–354. doi:10.1007/s00399-017-0535-x

Beutler, D. S., Gerkin, R., and Loli, A. (2014). The morphology of left atrial appendage lobes: a novel characteristic naming scheme derived through three-dimensional cardiac computed tomography. *World J. cardiovasc. Surg.* 04, 17–24. doi:10.4236/wjcs.2014.43004

Bhalodia, R., Subramanian, A., Morris, A., Cates, J., Whitaker, R., Kholmovski, E., et al. (2010). Does alignment in statistical shape modeling of left atrium appendage impact stroke prediction? *Comput. Cardiol.* 46, 46. doi:10.22489/cinc.2019.200

Bieging, E. T., Morris, A., Chang, L., Dagher, L., Marrouche, N. F., and Cates, J. (2021). Statistical shape analysis of the left atrial appendage predicts stroke in atrial fibrillation. *Int. J. Cardiovasc Imaging* 37 (8), 2521–2527. doi:10.1007/s10554-021-02262-8

Blackshear, J. L., and Odell, J. A. (1996). Appendage obliteration to reduce stroke in cardiac surgical patients with atrial fibrillation. *Ann. Thorac. Surg.* 61 (2), 755–759. doi:10.1016/0003-4975(95)00887-X

Bône, A., Louis, M., Martin, B., and Durrleman, S. (2018). "Deformetrica 4: an open-source software for statistical shape analysis," in *Shape in medical imaging*. Editors M. Reuter, C. Wachinger, H. Lombaert, B. Paniagua, M. Lüthi, and B. Egger (Cham: Springer International Publishing), 11167, 3–13. Lecture Notes in Computer Science. doi:10.1007/978-3-030-04747-4_1

Bosi, G. M., Cook, A., Rai, R., Menezes, L. J., Schievano, S., Torii, R., et al. (2018). Computational fluid dynamic analysis of the left atrial appendage to predict thrombosis risk. *Front. Cardiovasc Med.* 5, 34. doi:10.3389/fcvm.2018.00034

Cangelosi, R., and Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* 2 (1), 2. doi:10.1186/1745-6150-2-2

Capelli, C., Bosi, G. M., Cerri, E., Nordmeyer, J., Odenwald, T., Bonhoeffer, P., et al. (2012). Patient-specific simulations of transcatheter aortic valve stent implantation. *Med. Biol. Eng. Comput.* 50 (2), 183–192. doi:10.1007/s11517-012-0864-1

Cates, J., Bieging, E., Morris, A., Gardner, G., Akoum, N., Kholmovski, E., et al. (2015). Computational shape models characterize shape change of the left atrium in atrial fibrillation. *Clin. Med. Insights Cardiol.* 8 (Suppl. 1), 99–109. doi:10.4137/CMC.S15710

Cates, J., Elhabian, S., and Whitaker, R. (2017). "Chapter 10 - ShapeWorks: particle-based shape correspondence and visualization software," in *Statistical shape and deformation analysis*. Editors G. Zheng, S. Li, and G. Székely (Academic Press), 257–298.

Cerrolaza, J. J., Picazo, M. L., Humbert, L., Sato, Y., Rueckert, D., Ballester, M. Á. G., et al. (2019). Computational anatomy for multi-organ analysis in medical imaging: a review. *Med. Image Anal.* 56, 44–67. doi:10.1016/j.media.2019.04.002

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., and Ranzuglia, G. (2008). "MeshLab: an open-source mesh processing tool," in Eurographics Italian chapter conference (The Eurographics Association), 8.

Cootes, T. F., Taylor, C., Cooper, D., and Graham, J. (1995) "Training models of shape from sets of examples," in *Proc BMVC92*. Springer-Verlag.

Corrado, C., Razeghi, O., Roney, C., Coveney, S., Williams, S., Sim, I., et al. (2020). Quantifying atrial anatomy uncertainty from clinical data and its impact on electro-physiology simulation predictions. *Med. Image Anal.* 61, 101626. doi:10.1016/j.media.2019.101626

Davies, R. H. (2002). *Learning shape: optimal models for analysing natural variability*. United Kingdom: The University of Manchester.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., et al. (2022). A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng. Appl. Artif. Intell.* 110, 104743. doi:10.1016/j.engappai.2022.104743

Goparaju, A., Csecs, I., Morris, A., Kholmovski, E., Marrouche, N., Whitaker, R., et al. (2018). On the evaluation and validation of off-the-shelf statistical shape modeling tools: a clinical application. *Shape Med. Imaging (2018)* 11167, 14–27. doi:10.1007/978-3-030-04747-4_2

Goparaju, A., Iyer, K., Bône, A., Hu, N., Henninger, H. B., Anderson, A. E., et al. (2022). Benchmarking off-the-shelf statistical shape modeling tools in clinical applications. *Med. Image Anal.* 76, 102271. doi:10.1016/j.media.2021.102271

Hartman, E., Sukurdeep, Y., Klassen, E., Charon, N., and Bauer, M. (2023). Elastic shape analysis of surfaces with second-order sobolev metrics: a comprehensive numerical framework. *Int. J. Comput. Vis.* 131 (5), 1183–1209. doi:10.1007/s11263-022-01743-0

He, J., Fu, Z., Yang, L., Liu, W., Tian, Y., Liu, Q., et al. (2020). The predictive value of a concise classification of left atrial appendage morphology to thrombosis in non-valvular atrial fibrillation patients. *Clin. Cardiol.* 43 (7), 789–795. doi:10.1002/clc.23381

Hołda, M. K., Koziej, M., Hołda, J., Tyrak, K., Piątek, K., Bolechała, F., et al. (2017). Anatomic characteristics of the mitral isthmus region: the left atrial appendage isthmus as a possible ablation target. *Ann. Anat.* 210, 103–111. doi:10.1016/j.aanat.2016.11.011

Juhl, K. A., Slipsager, J., de Backer, O., Kofoed, K., Camara, O., and Paulsen, R. (2024). Signed distance field based segmentation and statistical shape modelling of the left atrial appendage. *arXiv*. doi:10.48550/arXiv.2402.07708

Karim, R., Blake, L. E., Inoue, J., Tao, Q., Jia, S., Housden, R. J., et al. (2018). Algorithms for left atrial wall segmentation and thickness – evaluation on an open-source CT and MRI image database. *Med. Image Anal.* 50, 36–53. doi:10.1016/j.media.2018.08.004

Kazhdan, M., and Hoppe, H. (2013). Screened Poisson surface reconstruction. *ACM Trans. Graph* 32, 1–13. doi:10.1145/2487228.2487237

Kendall, D. G. (1977). The diffusion of shape. *Adv. Appl. Probab.* 9 (3), 428–430. doi:10.2307/1426091

Korhonen, M., Muuronen, A., Arponen, O., Mustonen, P., Hedman, M., Jäkälä, P., et al. (2015). Left atrial appendage morphology in patients with suspected cardiogenic stroke without known atrial fibrillation. *PLoS One* 10 (3), e0118822. doi:10.1371/journal.pone.0118822

Li, C. Y., Gao, B. L., Liu, X. W., Fan, Q. Y., Zhang, X. J., Liu, G. C., et al. (2015). Quantitative evaluation of the substantially variable morphology and function of the left atrial appendage and its relation with adjacent structures. *PLoS One* 10 (7), e0126818. doi:10.1371/journal.pone.0126818

Linz, D., Gawalko, M., Betz, K., Hendriks, J. M., Lip, G. Y. H., Vinter, N., et al. (2024). Atrial fibrillation: epidemiology, screening and digital health. *Lancet Reg. Health Eur.* 37, 100786. doi:10.1016/j.lanepe.2023.100786

Marques, H., de Araújo Gonçalves, P., Ferreira, A. M., Cruz, R., Lopes, J., dos Santos, R., et al. (2018). Cardiac computed tomography prior to atrial fibrillation ablation: effects of technological advances and protocol optimization. *Rev. Port. Cardiol. English Ed.* 37 (11), 873–883. doi:10.1016/j.repc.2018.03.011

Martorana, V., Lee, M.T.-En, Rafizul, I., Menezes, L. J., Coronnello, C., Burriesci, G., et al. An unsupervised method to detect the left atrial appendage and extract its features.

Mellado, N., Aiger, D., and Mitra, N. J. (2014). Super 4PCS fast global pointcloud registration via smart indexing. *Comput. Graph. Forum* 33 (5), 205–215. doi:10.1111/cgf.12446

Musotto, G., Monteleone, A., Vella, D., Di Leonardo, S., Viola, A., Pitarresi, G., et al. (2022). The role of patient-specific morphological features of the left atrial appendage on the thromboembolic risk under atrial fibrillation. *Front. Cardiovasc Med.* 9, 894187. doi:10.3389/fcvm.2022.894187

Paulsen, R. R., and Hilger, K. B. (2003). "Shape modelling using Markov random field restoration of point correspondences," in *Information processing in medical imaging*. Editors C. Taylor and J. A. Noble (Berlin, Heidelberg: Springer), 1–12.

Pons, M. I., Mill, J., Fernandez-Quilez, A., Olivares, A. L., Silva, E., de Potter, T., et al. (2022). Joint analysis of morphological parameters and *in silico* haemodynamics of the left atrial appendage for thrombogenic risk assessment. *J. Interv. Cardiol.* 2022, 9125224. doi:10.1155/2022/9125224

Pulli, K. (1999). "Multiview registration for large data sets," in Second international conference on 3-D digital imaging and modeling (cat NoPR00062), 160–168.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Rusinkiewicz, S., and Levoy, M. (2001). "Efficient variants of the ICP algorithm," in Proceedings third international conference on 3-D digital imaging and modeling, Quebec, QC, 28 May 2001 - 01 June 2001. 145–152.

Shape model evaluation - ShapeWorks Available at: https://sciinstitute.github.io/ShapeWorks/new/ssm-eval.html.

Slipsager, J. M., Juhl, K. A., Sigvardsen, P. E., Kofoed, K. F., De Backer, O., Olivares, A. L., et al. (2019). "Statistical shape clustering of left atrial appendages," in *Statistical atlases and computational models of the heart atrial segmentation and LV quantification challenges*. Editors M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, and A. Young (Cham: Springer International Publishing), 32–39. (Lecture Notes in Computer Science).

van der Hoef, H., and Warrens, M. J. (2019). Understanding information theoretic measures for comparing clusterings. *Behaviormetrika* 46 (2), 353–370. doi:10.1007/s41237-018-0075-7

van Dongen, S., and Enright, A. J. (2012). Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv*. doi:10.48550/arXiv.1208.3145

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn Res.* 11, 2837–2854. doi:10.5555/1756006.1953024

Wang, Y., Di Biase, L., Horton, R. P., Nguyen, T., Morhanty, P., and Natale, A. (2010). Left atrial appendage studied by computed tomography to help planning for appendage closure device placement. *J. Cardiovasc. Electrophysiol.* 21 (9), 973–982. doi:10.1111/j.1540-8167.2010.01814.x

Wu, L., Liang, E., Fan, S., Zheng, L., Du, Z., Liu, S., et al. (2019). Relation of left atrial appendage morphology determined by computed tomography to prior stroke or to increased risk of stroke in patients with atrial fibrillation. *Am. J. Cardiol.* 123 (8), 1283–1286. doi:10.1016/j.amjcard.2019.01.024

Yaghi, S., Chang, A. D., Akiki, R., Collins, S., Novack, T., Hemendinger, M., et al. (2020). The left atrial appendage morphology is associated with embolic stroke subtypes using a simple classification system: a proof of concept study. *J. Cardiovasc Comput. Tomogr.* 14 (1), 27–33. doi:10.1016/j.jcct.2019.04.005

TYPE Brief Research Report
PUBLISHED 14 February 2025
DOI 10.3389/fnetp.2025.1488349

Check for updates

OPEN ACCESS

EDITED BY
Alessio Gizzi,
Campus Bio-Medico University, Italy

REVIEWED BY
Marianna Meo,
Boston Scientific, Netherlands
Yaopeng Ma,
Bar-Ilan University, Israel

*CORRESPONDENCE
Christoforos Galazis,
✉ c.galazis20@imperial.ac.uk

RECEIVED 29 August 2024
ACCEPTED 20 January 2025
PUBLISHED 14 February 2025

CITATION
Galazis C, Chiu C-E, Arichi T, Bharath AA and Varela M (2025) PINNing cerebral blood flow: analysis of perfusion MRI in infants using physics-informed neural networks.
Front. Netw. Physiol. 5:1488349.
doi: 10.3389/fnetp.2025.1488349

PINNing cerebral blood flow: analysis of perfusion MRI in infants using physics-informed neural networks

Christoforos Galazis[1,2]*, Ching-En Chiu[2,3], Tomoki Arichi[4], Anil A. Bharath[5,6] and Marta Varela[2,7]

[1]Department of Computing, Imperial College London, London, United Kingdom, [2]National Heart and Lung Institute, Imperial College London, London, United Kingdom, [3]Department of Electrical Engineering, Imperial College London, London, United Kingdom, [4]Centre for the Developing Brain, King's College London, London, United Kingdom, [5]Imperial Global Singapore, CREATE Tower, Singapore, Singapore, [6]Department of Bioengineering, Imperial College London, London, United Kingdom, [7]Cardiovascular and Genomics Research Institute, City St George's University of London, London, United Kingdom

Arterial spin labelling (ASL) magnetic resonance imaging (MRI) enables cerebral perfusion measurement, which is crucial in detecting and managing neurological issues in infants born prematurely or after perinatal complications. However, cerebral blood flow (CBF) estimation in infants using ASL remains challenging due to the complex interplay of network physiology, involving dynamic interactions between cardiac output and cerebral perfusion, as well as issues with parameter uncertainty and data noise. We propose a new spatial uncertainty-based physics-informed neural network (PINN), SUPINN, to estimate CBF and other parameters from infant ASL data. SUPINN employs a multi-branch architecture to concurrently estimate regional and global model parameters across multiple voxels. It computes regional spatial uncertainties to weigh the signal. SUPINN can reliably estimate CBF (relative error −0.3 ± 71.7), bolus arrival time (AT) (30.5 ± 257.8), and blood longitudinal relaxation time (T_{1b}) (−4.4 ± 28.9), surpassing parameter estimates performed using least squares or standard PINNs. Furthermore, SUPINN produces physiologically plausible spatially smooth CBF and AT maps. Our study demonstrates the successful modification of PINNs for accurate multi-parameter perfusion estimation from noisy and limited ASL data in infants. Frameworks like SUPINN have the potential to advance our understanding of the complex cardio-brain network physiology, aiding in the detection and management of diseases. Source code is provided at: https://github.com/cgalaz01/supinn.

KEYWORDS

physics-informed neural networks, cardiac-brain network physiology, neuroimaging, arterial spin labelling, cerebral blood perfusion

1 Introduction

Arterial spin labelling (ASL) is a non-invasive magnetic resonance imaging (MRI) technique that measures cerebral blood flow (CBF) without exogenous contrast agents (Lindner et al., 2023). CBF maps can be computed on a voxel-by-voxel basis by fitting mathematical models of haemodynamics based on ordinary differential equations (ODEs) (Alsop et al., 2015). These models help capture the complex temporal dynamics of blood flow, which are essential for understanding the intricate cardiac-brain network physiology. This understanding may aid in diagnosing and managing various conditions, such as some forms of dementia and stroke (Rossi et al., 2022; Tahsili-Fahadan and Geocadin, 2017).

The bidirectional cardiac-brain network physiology operates as an intricate system where the heart and brain continuously influence each other (Candia-Rivera et al., 2024), a topic that has garnered research interest for some time (Bashan et al., 2012). The heart supplies oxygenated blood to the brain, affecting cerebral perfusion and pulsatile flow (Silverman and Petersen, 2020; Jammal Salameh et al., 2024), while the brain regulates cardiac function through the two autonomic nervous systems, the sympathetic and parasympathetic (Gordan et al., 2015). This network incorporates feedback loops such as cerebral autoregulation and neurovascular coupling to maintain optimal function (Claassen et al., 2021).

In infants, particularly those with conditions like congenital heart disease (CHD) or preterm birth, this network is especially vulnerable due to immature autoregulation and developmental sensitivity (De Silvestro et al., 2024; Claassen et al., 2021). These factors can result in altered cerebral haemodynamics, leading to issues such as delayed brain maturation, an increased risk of cerebral white matter injury, and potentially adverse long-term neurodevelopmental outcomes (McQuillen et al., 2010). Preterm neonates are often admitted to hospital to receive external physiological support whilst their bodies mature, of which brain perfusion must be sufficient during this period.

The infant demographic thus benefits from non-invasive CBF monitoring techniques like ASL (Counsell et al., 2019). ASL can provide insights into the complex physiological interplay between the heart and brain, guiding interventions to support optimal brain development and overall cardiovascular health (McQuillen et al., 2010; Castle-Kirszbaum et al., 2022).

A thorough understanding of this cardiac-brain network is crucial for managing infant health. Specifically, it is essential for optimising neuroprotection strategies, improving surgical and medical management, and enhancing the long-term neurodevelopmental prospects of these infants (De Silvestro et al., 2023). However, further research is needed to fully understand the independent effects and mechanisms of cardio-cerebral coupling (Castle-Kirszbaum et al., 2022; Meng et al., 2015), particularly in the developing infant brain (Baik-Schneditz et al., 2021). Achieving this understanding in infants will require the development of even more accurate CBF monitoring techniques than those currently available.

Computing voxel-by-voxel CBF maps is achieved by fitting mathematical models of haemodynamics based on ODEs (Alsop et al., 2015). Many of these perfusion model ODEs assume very simplified physiology (e.g., plug blood flow to the brain, single magnetisation compartments in the brain) and can therefore be solved analytically (Buxton et al., 1998; Alsop et al., 2015). It is often further assumed that the perfusion model parameters are perfectly known. In these conditions, CBF is estimated from a single perfusion-weighted image (PWI). These assumptions do not apply to CBF estimates in pathological conditions or groups with heterogeneous physiological properties, such as infants.

Imaging infants, particularly those born preterm, presents further challenges due to lower signal-to-noise ratio (SNR). This is attributed to lower baseline CBF and longer arrival times (AT) of the magnetically labelled bolus (Dubois et al., 2021; Varela et al., 2015). Additionally, the need for higher spatial resolution in smaller infant brains further reduces SNR (Dubois et al., 2021). Motion during scanning is also common in infants, further degrading image quality and leading to artifacts (Dubois et al., 2021; Varela et al., 2015).

Unfortunately, voxel-by-voxel ASL analysis is susceptible to spatial inconsistencies, amplified by the lower SNR noise in infant perfusion weighted image (PWI) signals (Krishnapriyan et al., 2021; Wang et al., 2022). Haemodynamic models are challenging to parameterise in the infant population due to dramatic physiological changes in the first weeks of life, during which most physiological parameters differ substantially from adult values. This is true of haemodynamic variables such as CBF, and also tissue composition, reflected in MR relaxation time constants such as T_1 and T_2. This is further complicated by the limited availability of data in this demographic (De Silvestro et al., 2023).

In adult ASL, CBF estimation is commonly performed at a single time point following labelling (Detre et al., 2012). This relies on several assumptions about haemodynamics and MR parameters that do not usually hold for infants. Given the complexity of the cerebral blood flow network in infants, past ASL studies in infants have therefore acquired PWIs at multiple time points following labelling to enable the simultaneous estimation of haemodynamic parameters beyond CBF, such as AT (Varela et al., 2015). Past studies estimated CBF and other parameters using methods such as least squares fitting (LSF) using the analytical solution to the perfusion ODE (Varela et al., 2015). However, due to the complexity of haemodynamic models, most model parameters need to be estimated separately. The lack of methods capable of simultaneously estimating both local and global parameters presents a significant challenge.

CBF has been estimated from infant ASL data using optimisers like LSF (Varela et al., 2015) and Bayesian estimation (Pinto et al., 2023), where adult models are fitted to the PWI signal. These voxel-by-voxel approaches often struggle with the very noisy PWIs typical of infant data, especially when estimating several parameters at once. Recently, neural network (NN)-based techniques for parameter estimation have become increasingly popular. NNs have demonstrated a remarkable ability to make accurate predictions even from noisy and corrupt data (Tian et al., 2020; Hernandez-Garcia et al., 2022). However, such performance typically requires vast amounts of training data (Tian et al., 2020), which are currently not available for infants (Korom et al., 2022; Hernandez-Garcia et al., 2022; De Silvestro et al., 2023).

Physics-informed neural networks (PINNs) (Karniadakis et al., 2021), an emerging branch of machine learning, integrate physical laws (expressed as differential equations, DEs) into machine learning models. This approach improves a network's predictive capabilities even with limited and noisy data, as the DE agreement terms effectively act as a strong regulariser (Karniadakis et al., 2021). PINNs can simultaneously solve DEs (forward problem) and estimate system parameters (inverse problem) from sparse experimental data. This makes them well-suited for biomedical applications (Ghalambaz et al., 2024), evident by their increased usage in fields such as cardiovascular (Moradi et al., 2023; Herrero Martin et al., 2022; Sahli Costabal et al., 2020; van Herten et al., 2022;

Kissas et al., 2020) and brain (Sarabian et al., 2022; Kamali et al., 2023; de Vries et al., 2023; Min et al., 2023) research.

In cardiovascular studies, PINNs have been successfully applied to predict electrophysiological tissue properties from action potential recordings (Herrero Martin et al., 2022) and to diagnose atrial fibrillation by estimating electrical activation maps (Sahli Costabal et al., 2020). Additionally, PINNs have been used to quantify myocardial perfusion using MR imaging (van Herten et al., 2022) and to predict arterial pressure by analysing MRI data of blood velocity and wall displacement (Kissas et al., 2020). However, while PINNs are typically robust to noise, they suffer from the spatial inconsistencies associated with voxel-by-voxel fitting. PINNs' performance is notoriously variable, especially in inverse mode (Bajaj et al., 2023).

A significant challenge in PINN development is that they are often tested using synthetic data, which may not be a robust benchmark for performance on experimentally-acquired data. This is because few biomedical problems described by differential equations have known analytical solutions. Consequently, applications like CBF estimation using ASL data present rare opportunities to test PINNs' performance directly on experimental data and compare it to established parameter estimation methods such as LSF. Such real-world applications are crucial for validating and improving PINN methodologies in biomedical research.

This study introduces and evaluates PINNs as a tool for reliably estimating haemodynamic parameters from noisy infant ASL images. We propose a novel PINN framework, named Spatial Uncertainty PINN (SUPINN), which incorporates two key noise-mitigating improvements: 1) Regional: We assume neighbouring voxels share similar local parameters (e.g., CBF and AT) and therefore similar time courses. We thus propose weighting the confidence in each measurement by its spatial variability. 2) Global: For global parameters (e.g., T_{1b}), which are identical across all voxels within a subject, our multi-branch SUPINN learns from multiple voxels simultaneously to estimate a shared global parameter. Our method is particularly suited for imaging data acquired with limited and noisy samples over a given time period.

2 Methods

Our source code is publicly available at: https://github.com/cgalaz01/supinn.

2.1 Dataset

ASL brain MRI studies were conducted on seven infants aged 32–78 weeks postmenstrual age. An additional five infants were scanned but excluded due to significant motion artifacts or because they awoke during the scan, rendering the data unusable. The final cohort included three infants with no pathology, one with periventricular leukomalacia, one with basal ganglia and white matter atrophy along with mild ventriculomegaly, one with agenesis of the corpus callosum, brain atrophy, and mild ventriculomegaly, and one with mild ventriculomegaly. Although this study does not include infants with known cardiac impairment, it is sufficient as our focus at this stage is on evaluating PINNs within the available diverse cohort.

All images were acquired in a Philips 3T Achieva scanner using an 8-element head coil under ethical approval following informed parental consent (REC: 09/H0707/83). PWIs were acquired on a single mid-brain transverse plane at 12 time points (every 300 ms) following a single pulsed labelling event (Petersen et al., 2006), at a spatial resolution of $3.04 \times 3.04 \times 5.5\ mm^3$. The 300 ms time interval between PWI acquisitions was deemed suitable for this demographic (Varela et al., 2015), as it provides a practical balance between SNR and temporal perfusion signal sampling. For a representative PWI time series and accompanying signal plot, refer to Figure 1.

To improve the SNR, the acquisition was repeated multiple times, with the number of repeats ranging from 30 to 90 depending on the remaining scanning session duration and the subject's ability to remain still. Images identified as having motion artefacts were excluded from the averaging process based on manual inspection. Notably, no signal filtering was applied in this study to further reduce noise.

In all subjects, our analysis focused on a manually segmented region of interest that includes the thalami and basal ganglia (Figure 2). This deep grey matter region shows better SNR and fewer partial volume effects than cortical grey matter.

2.2 Mathematical model for ASL

The relationship between the PWI signal, $S(t)$, and CBF can be expressed as the temporal convolution between an arterial input function, $AIF(t)$, and a tissue response function, $R(t)$: $S = AIF * R$ (Buxton et al., 1998). AIF is a top-hat function, here with a known duration $\tau = 900\,ms$, that arrives at each voxel at a variable $t = AT$, and $R(t)$ is dominated by magnetisation relaxation over venous outflow. As in Alsop et al. (2015), we assume that the longitudinal magnetisation relaxation of the blood is well described by T_{1b} throughout.

We neglect the effect of the repeated excitation pulses on apparent T_{1b} and assume that all PWI scaling constants are known, as in Varela et al. (2015). Then:

$$S(t) = \begin{cases} 0 & \text{if } t < AT \\ CBF \times (t - AT) \times e^{\frac{-t}{T_{1b}}} & \text{if } AT \le t < AT + \tau \\ CBF \times \tau \times e^{\frac{-t}{T_{1b}}} & \text{if } AT + \tau \le t \end{cases} \tag{1}$$

This model can be differentiated to yield an ODE defined in 3 branches:

$$\frac{dS}{dt} = \begin{cases} 0 & \text{if } t < AT \\ CBF \times e^{\frac{-t}{T_{1b}}} \times \left(1 - \dfrac{t - AT}{T_{1b}}\right) & \text{if } AT \le t < AT + \tau \\ -CBF \times e^{\frac{-t}{T_{1b}}} \times \dfrac{\tau}{T_{1b}} & \text{if } AT + \tau \le t \end{cases} \tag{2}$$

The three branches in Equations 1, 2 depict three distinct signal evolution phases: the periods before, during, and after the arrival of labelled blood at each voxel. We found that approximating the discontinuous three-branched ODE in Equation 2 using a NN leads to poor convergence properties. To circumvent this issue, we

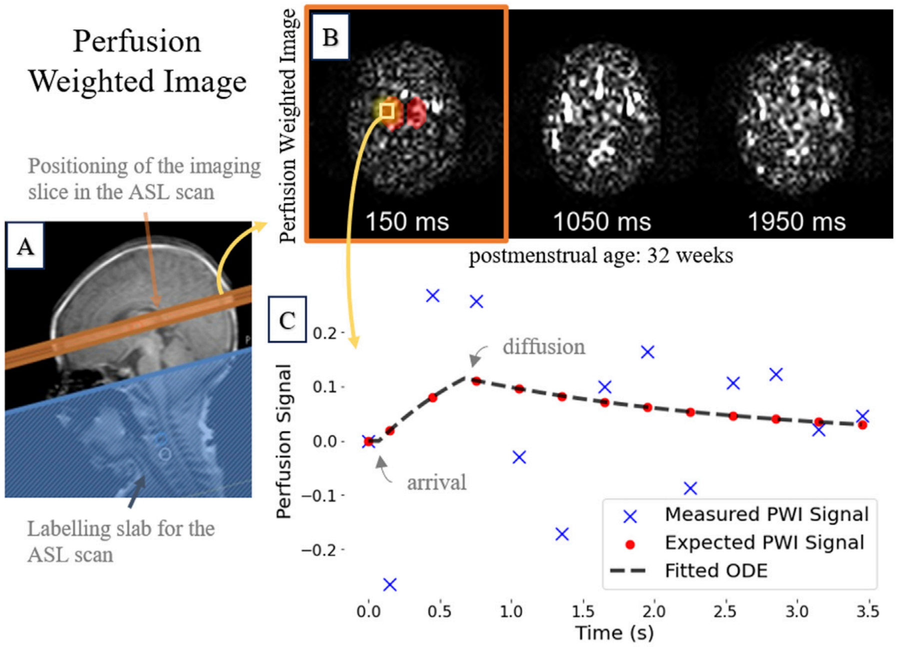


FIGURE 1
A representative 32-week postmenstrual case showing: **(A)** T_2-weighted image highlighting the ASL imaging slice (orange); **(B)** Subsampled perfusion-weighted image time series; and **(C)** The measured perfusion signal of a single voxel over the entire duration, along with the corresponding ground-truth analytical model (see Equation 2).

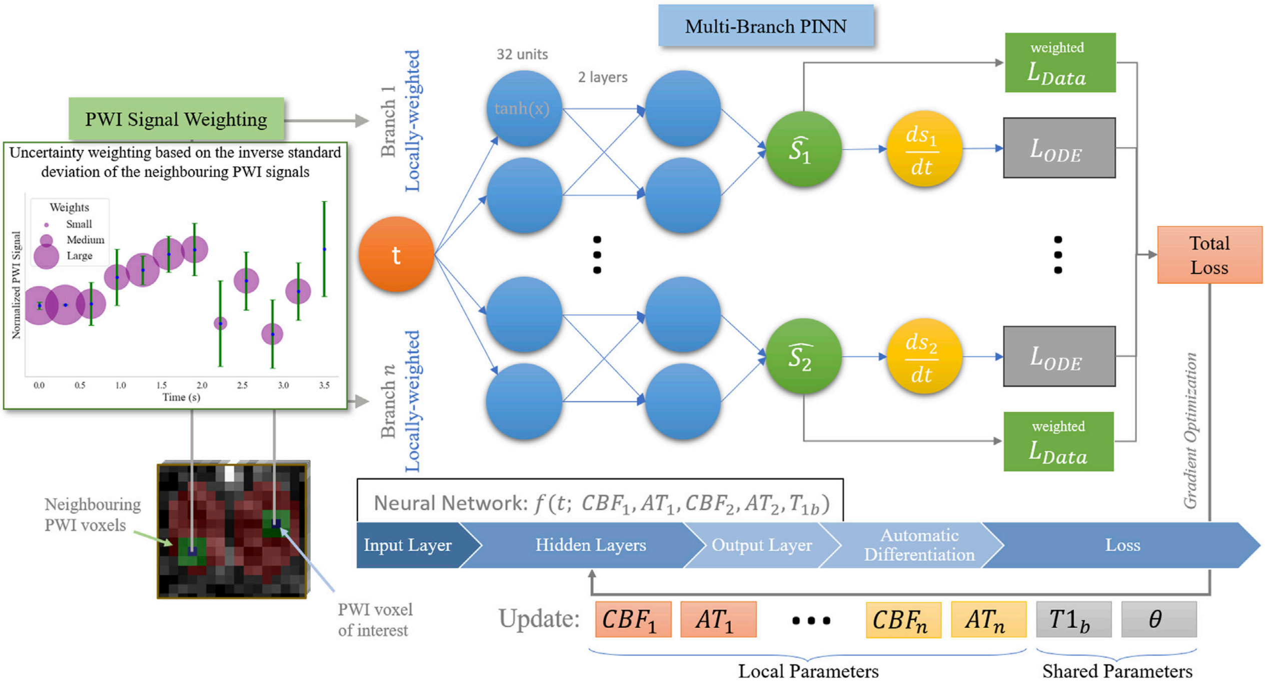


FIGURE 2
Overview of our proposed SUPINN model, depicted here in a two-branch variant for illustration purposes, but adaptable to larger configurations. This study employs a three-branch model based on empirical findings.

combine the three phases using smoothing hyperbolic tangent functions (see Supplementary Table S1).

2.3 Ground truth estimation

An auxiliary MRI scan was used to estimate ground-truth T_{1b} in each subject (Varela et al., 2011). Then, a robust LSF was performed using the analytical haemodynamic model in Equation 2 to estimate ground-truth CBF and AT on a voxel-by-voxel basis.

Most biomedical problems described by DEs do not have an analytical solution and can only be solved numerically. For these, the accuracy of parameter identification methods is typically estimated using *in silico* data, which do not capture the complexities of experimental measurements. The existence of an analytical ASL haemodynamic model (Equation 2) presents a unique opportunity to test on experimental data the accuracy of model parameter estimation methods such as PINNs.

2.4 Loss function and training scheme

PINNs are optimised to learn a solution that both matches the data and satisfies known cardiac-brain network physiology principles. They minimise the combined loss function defined as: $\mathcal{L} = \mathcal{L}_{ODE} + \gamma\mathcal{L}_{data}$. Due to the high noise in the data, $\mathcal{L}_{data}$ is weighted using an empirically set coefficient $\gamma = 0.005$. Initial conditions, $S(t = 0) = 0$, are enforced by rescaling $S(t)$ using a hyperbolic tangent function (Lu et al., 2021).

$\mathcal{L}_{ODE}$ measures the agreement with Equation 2. This loss is calculated by evaluating the residual of the differential equation at a set of collocation points (N_O) using the network's predictions and taking the mean squared error:

$$\mathcal{L}_{ODE} = \frac{1}{N_O}\sum_{i}^{N_O}\left(\frac{d\hat{s}}{dt}(t_i) - f(t_i, \hat{s}(t_i))\right)^2 \qquad (3)$$

$\mathcal{L}_{data}$ is the data loss, which measures the mean squared error between the network's PWI estimation and the values measured across the 12 time points (N_D) acquired in each voxel:

$$\mathcal{L}_{data} = \frac{1}{N_D}\sum_{i}^{N_D}\left(w_{t_i} \times \|\hat{S}(t_i) - S(t_i)\|^2\right), \qquad (4)$$

where $w = 1$ is the weight of each PWI time point. w is used in SUPINN with details available in Section 2.6.

When optimising the PINNs' weights, we propose a three-tier hierarchical optimisation scheme (see Supplementary Table S2). We initially optimise the PINNs in forward mode, focusing on aligning the network approximately with the underlying ODE without estimating specific parameters. We then solve the ODE in inverse mode to estimate the local parameters CBF and AT, and the global parameter T_{1b}. We finalise by fine-tuning the parameter estimation.

2.5 PINN architecture

PINNs are implemented using DeepXDE v1.11 (Lu et al., 2021) and TensorFlow v2.15 (Abadi et al., 2016). As a baseline PINN architecture (Raissi et al., 2019; Karniadakis et al., 2021), we use a fully connected neural network with hyperbolic tangent activation functions and two hidden layers, each consisting of 32 units. It includes one input unit for time t and one output unit for the PWI signal $S(t)$.

2.6 SUPINN architecture

The baseline PINN models the signal from each voxel separately, ignoring the spatial relationships between the different sets of measurements. We expect, however, that neighbouring voxels have similar CBF and AT values, with deviations primarily due to noise. To incorporate this information in the model, we propose a spatial uncertainty PINN, SUPINN (Figure 2). SUPINN inversely weighs the contribution of each PWI time point, w (see Equation 4), by their uncertainty levels. The uncertainty is estimated by calculating the standard deviation of the PWI signal in immediate neighbouring voxels within the region of interest at a given time point: $w_t = 1/\sqrt{\frac{\sum(S(t_i)-\mu_{t_i})^2}{8}}$, where w is the weight at time point t. The weights for each voxel across time are then scaled such that the highest uncertainty corresponds to a weight of $w = 0.1$ and the smallest uncertainty to $w = 1$. The weights in data loss $\mathcal{L}_{data}$ (Equation 4) are updated accordingly.

SUPINN uses a multi-branch architecture to reliably estimate global (subject-specific) parameters, such as T_{1b} by pooling information from more than one voxel. It simultaneously estimates voxel-specific parameters *CBF* and *AT*. The subnetworks' graphs are merged, allowing information sharing through backpropagation.

Each SUPINN branch employs the baseline PINN architecture described in Section 2.5. We have experimentally found that using a three-branch SUPINN for this task results in an optimal balance between estimation accuracy and computational efficiency. Increasing the number of branches leads to minimal decreases in estimation error with exponentially larger computation times (see Supplementary Figure S1). In addition to the voxel of interest, two additional voxels are randomly selected within the whole region of interest that was manually delineated for the remaining branches. This delineated sampling region has an average width of $52.55 \pm 7.74\ mm$ and height of $39.09 \pm 6.59\ mm$. While voxel-specific CBF and AT parameters are estimated independently in each branch, T_{1b} is shared across the selected voxels. The loss function, $\mathcal{L}$, for this architecture is the sum of the data agreement and ODE agreement losses (Equations 3, 4) for each branch: $\mathcal{L} = \sum_i^{N=3}\mathcal{L}_{i,ODE} + \mathcal{L}_{i,data}$.

2.7 Experimental setup

We compared SUPINN against several benchmarks: a standard PINN (Section 2.5), a robust LSF method (Varela et al., 2015), and a modified LSF (LSF-multi) that averages parameter estimations from three selected voxels. As we have limited data, evaluation against deep NN is not currently possible. All computations were performed on a 3XS Intel Core i7 CPU. The average execution times per voxel were

TABLE 1 Summary of the convergence rate, relative error and Laplacian variance for CBF, AT and T_{1b}, and mean squared error of the predicted solution. A model's quality is indicated by a low standard deviation and a mean error close to 0.

Model	Convergence rate (%)	Relative error (%)			Laplacian variance		Mean squared error
		CBF	AT	T_{1b}	CBF	AT	PWI signal $(\times 10^{-3})$
LSF	62.6	390.7 ± 1306.7	53.8 ± 510.7	−43.1 ± 32.2	29.1 ± 11.8	3.1 ± 2.7	26.9 ± 22.7
LSF-multi	96.4	549.7 ± 1272.0	121.9 ± 467.0	−31.4 ± 29.9	12.4 ± 5.7	1.2 ± 1.0	38.3 ± 31.4
PINN	99.9	96.0 ± 475.8	68.6 ± 283.9	8.6 ± 35.9	0.5 ± 0.4	0.5 ± 0.8	1.1 ± 1.3
SUPINN	***100.0***	***−0.3 ± 71.7***	***30.5 ± 257.8***	***−4.4 ± 28.9***	***0.4 ± 0.4***	***0.1 ± 0.1***	***0.7 ± 0.8***

approximately 0.05 s for LSF/LSF-multi, 31 s for PINN, and 40 s for SUPINN. Given an average voxel size in the region of interest of 110 ± 46 voxels, this corresponds to average total execution times per case of 5.5 s for LSF/LSF-multi, 56.8 min for PINN, and 73.3 min for SUPINN. We note that substantial improvement in training time can be obtained on PINN/SUPINN if trained on a GPU.

Evaluation metrics include the mean and standard deviation of the relative error (RE), computed as $(predicted - target)/target \times 100$ for each parameter. When a method led to CBF estimates that increasingly diverged from ground truth CBF by more than one order of magnitude after $50K$ iterations, it was deemed not to have converged. These failed estimates were not taken into account when assessing the quantitative performance of each method. We compute a method's convergence rate as $|total - failed|/total \times 100$. The spatial smoothness of CBF and AT was assessed using the mean and standard deviation of the Laplacian variance across subjects (Pertuz et al., 2013), where lower variance signifies greater spatial parameter homogeneity. We also estimate the mean squared error (MSE) between the prediction and ground truth PWI signal (forward mode).

3 Results

Our proposed SUPINN architecture, designed to address variable data noise levels and simultaneously estimate local and global parameters, showed excellent performance on infant ASL data (see Table 1). SUPINN showed improvements in both PWI signal (forward) and parameter (inverse) estimations compared to the standard PINN and LSF/LSF-multi methods at the cost of increased computational time.

SUPINN led to more accurate parameter estimates, especially for CBF. Specifically, SUPINN achieved a RE of −0.3 ± 71.7 for CBF, 30.5 ± 257.8 for AT, and −4.4 ± 28.9 for T_{1b}. Additionally, the predicted PWI signal closely matched the ground truth, as evidenced by the smallest MSE of 0.4 ± 0.8, as shown in Table 1. Finally, both the base PINN and SUPINN achieved high parameter convergence rates, with rates of 99.9% and 100%, respectively.

We typically observe higher noise levels in the PWI signal of younger infants. Despite this challenge, Supplementary Figure S2 shows that SUPINN consistently achieved lower RE in CBF across all subjects compared to other methods despite low SNR. Additionally, SUPINN achieved the most accurate estimates of AT and T_{1b} in the majority of cases. Notably, SUPINN also demonstrated resilience in estimating parameters for infants with neurological disorders (indicated with an asterisk in the figure).

The robustness of our model is further demonstrated in Supplementary Figure S3, where we evaluated its performance on synthetic signals. White Gaussian noise was added to each synthetically generated PWI signal to simulate stationary noise, as motion artefacts are expected to be manually removed during the averaging process. The standard deviation progressively increased in increments of 0.1, up to a maximum of 0.5. Despite increasing the standard deviation of the noise, SUPINN maintained stable parameter estimations, especially for CBF and AT. This highlights the model's ability to handle noisy data effectively. In comparison, the baseline PINN also exhibited resilience in estimating AT and T_{1b}, but its CBF estimations deteriorated progressively as the noise level increased. On the other hand, the LSF method showed the greatest sensitivity to noise, with parameter estimations degrading noticeably even with a small amount of added noise.

Figure 3 illustrates the spatial maps of the CBF and AT predictions for a representative infant. The SUPINN estimates, shown in the first column, exhibit higher spatial consistency for both CBF and AT compared to other methods. This consistency is quantified by the lowest Laplacian variance achieved, as detailed in Table 1. Specifically, SUPINN attained a Laplacian variance of 0.4 ± 0.4 for CBF and 0.1 ± 0.1 for AT across all cases, indicating smoother and more reliable spatial predictions.

The average normalised CBF in the region of interest, as estimated by SUPINN, showed a general increase with age, which aligns with expectations. The youngest infant, with a postmenstrual age of 38 weeks, had a CBF of 0.12 ± 0.11, while the oldest, at 78 weeks, had a CBF of 0.56 ± 0.29. The CBF values for all subjects are presented in Supplementary Figure S4. However, due to the limited number of cases and the high variability in modelling this demographic, drawing definitive conclusions about the effects of pathology compared to healthy subjects remains challenging. For instance, within the same age group, a subject aged 49 weeks exhibited a CBF drop of approximately 0.12 compared to other infants in the same age range. On the other hand, an infant aged 32 weeks with pathology had a CBF value similar to that of a healthy infant aged 34 weeks. On the other hand, normalised AT values were similar across subjects and ranged from 0.32 to 0.49 s, with the oldest subject exhibiting the lowest value.

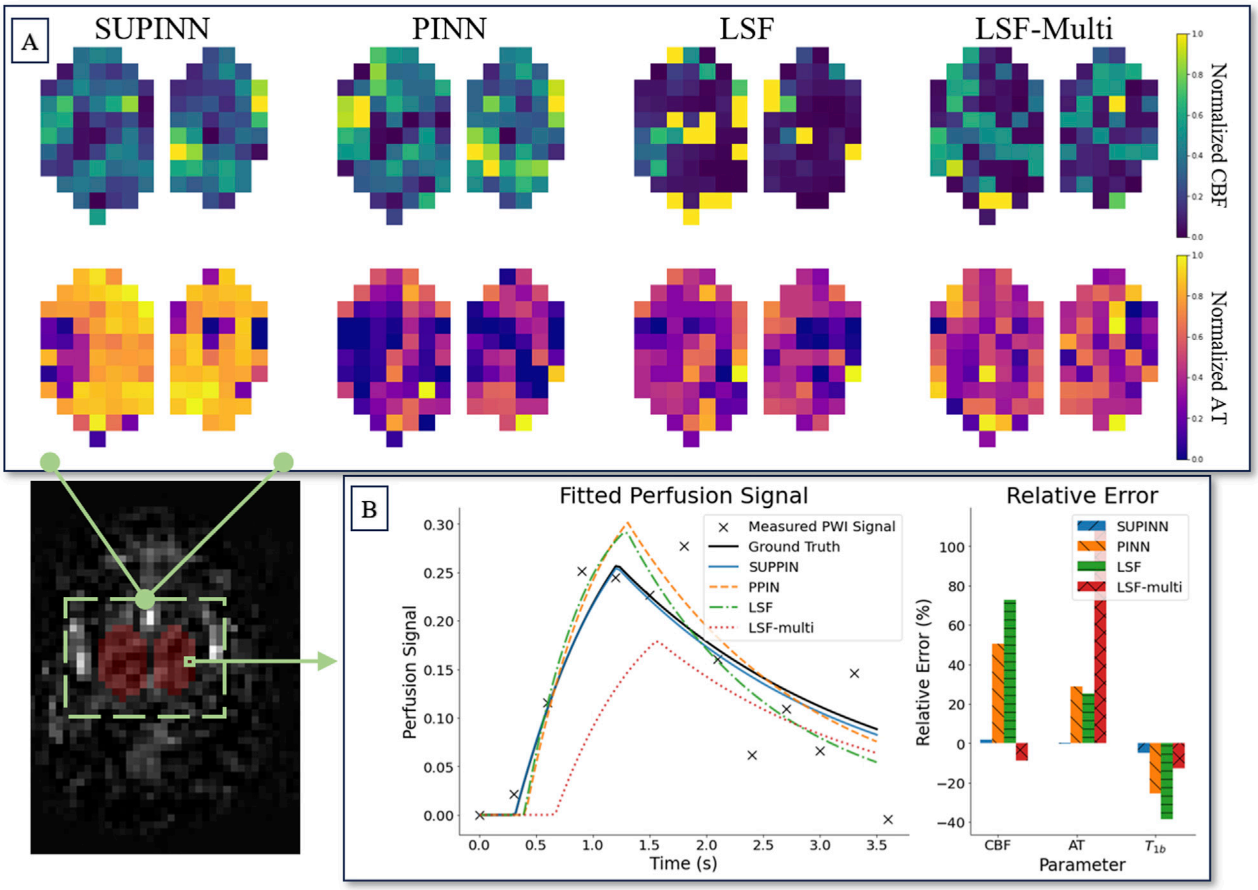


FIGURE 3
(A) Shows spatial maps of parameter estimation in deep grey matter for a subject aged 32 weeks. Each row corresponds to the normalised CBF (top) and AT (bottom) parameters. The columns display the estimation results from four methods (left to right): SUPINN, PINN, LSF, and LSF-multi. **(B)** Depicts the parameter relative error of the models for a single voxel.

4 Discussion

We introduce SUPINN, a novel multi-branch PINN technique for estimating parameters from noisy data. By solving ODEs over neighbouring regions with similar properties and estimating uncertainty through voxel comparisons, SUPINN simultaneously estimates local and global parameters with high accuracy. We test it on the challenging task of estimating haemodynamic parameters from extremely noisy infant multi-delay ASL data, where it outperforms both standard PINNs and LSF.

SUPINN's strong performance is also underpinned by our three-tier optimisation regime, use of hard initial conditions and the replacement of non-differentiable transitions in the baseline model (Equation 2) by a smoothly interpolated version. These enhancements are crucial for accurately capturing the complex cerebral haemodynamics in infants, in whom subtle alterations in perfusion can have implications for brain development.

LSF is widely used for parameter identification from various medical images, including ASL. It performs reliably when estimating a small number of parameters, particularly multiplicative factors or temporal intervals (such as CBF or AT in Equation 2). Following the literature (Varela et al., 2015; Hernandez-Garcia et al., 2022), we used robust LSF to estimate ground-truth CBF and AT when separate ground-truth measurements of T_{1b} were available. LSF is nevertheless extremely unreliable when estimating exponents such as T_{1b} in conjunction with CBF and AT.

PINNs have several advantages over LSF other than improved overall performance. Evidently from SUPINN, they offer a framework for more flexibly combining data from different brain and, in the future, cardiac regions. Contrary to standard PINNs, SUPINN is able to handle data with high noise to further improve performance. SUPINN leads to spatially smoother CBF and AT maps within the same brain region, aligning more closely with physiological expectations. Moreover, PINNs can be applied to ODEs with no known analytical solutions, opening up the possibility of using more sophisticated and personalised perfusion ODEs.

Recent advancements in PINN architectures, such as those described by Zou et al. (2025), Pilar and Wahlström (2024), Zou et al. (2024), further improve their utility by facilitating uncertainty quantification, particularly under conditions of heavy noise. Additionally, efforts are made towards adapting PINNs for model personalisation (Chen et al., 2021), which is useful especially when there could also be uncertainty in the assumptions used to derive the model itself. These capabilities are especially valuable when modelling the infant demographic, where data can be highly variable and noisy. Changes affecting the perfusion signal curve must be incorporated into the ODE parameters, and we expect these operator-controlled changes to result in less uncertainty than physiological unknowns. However, motion artefacts remain a challenge, requiring manual inspection and removal before averaging the PWI signal. Recent efforts have used deep learning techniques to reduce artefacts and improve overall SNR (Hales et al., 2020; Hernandez-Garcia et al., 2022).

Although SUPINN achieves spatially smoother CBF and AT maps, we employed a relatively simple sampling strategy - random sampling. This was due to the use of a single PWI plane and its lower resolution, which limited the practicality of alternative sampling approaches. In the future, we plan to acquire multiple PWI planes across the infant brain, enabling the implementation of spatially dependent sampling methods.

Since SUPINN, and to an extent LSF-multi, relies on sampling within the designated grey matter region, segmentation inaccuracies are expected to degrade overall performance due to the inclusion of lower SNR points in the branches. Such degradation for PINNs and LSF will only be observed for points outside the true region, while the true points remain unchanged. This issue could be mitigated for SUPINN by increasing the number of branches, under the assumption that the proportion of mislabelled voxels would be small, at the cost of computational time.

The multi-delay ASL data are well-suited for testing parameter identification methods, as the existence of an analytical solution allows for easy application of LSF. SUPINN's performance can therefore be directly evaluated on real, noisy clinical data. This is in contrast to most PINN studies, which are typically evaluated on synthetic data with known noise distributions. Although our current dataset does not include cases of CHD in infants, the techniques developed here are likely to be applicable to such cases, given the similar challenges in analysing cerebral haemodynamics. Furthermore, it is encouraging to see that other research efforts have successfully utilised PINNs to estimate CBF (Ishida et al., 2024; Rotkopf et al., 2024; de Vries et al., 2023), reinforcing the potential of these methods in addressing similar challenges.

Future work will expand the evaluation to include a larger infant cohort of both healthy and CHD cases to validate the robustness and generalizability of SUPINN. This will enable us to assess the efficacy of the improved CBF estimation specifically in the context of CHD and explore its relation to the disease. Optimising voxel selection strategies and exploring alternative PINN architectures, such as graph-based approaches, can further improve performance by better representing spatial relationships critical in various clinical scenarios, including CHD.

SUPINN's applicability extends to other problems where ODEs are solved over neighbouring regions with similar parameters. SUPINN can, for example, contribute to estimating quantitative MRI properties (such as T_1 or T_2) by simultaneously solving the Bloch equations in neighbouring voxels within the same tissue (Zimmermann et al., 2024).

This paper proposes SUPINN, a PINN method able to handle noisy data by leveraging spatial information. We demonstrate its potential to improve the characterisation of haemodynamics using infant ASL. With further refinement and validation, SUPINN can become a valuable clinical tool, providing precise and accurate physiological data for diagnosis, monitoring, and treatment planning in various clinical contexts, including potential applications in infants with CHD.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Restricted access. Requests to access these datasets should be directed to Christoforos Galazis, c.galazis20@imperial.ac.uk.

Ethics statement

The studies involving humans were approved by the Hammersmith and Queen Charlotte's and Chelsea Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

CG: Methodology, Software, Writing–original draft. C-EC: Software, Writing–review and editing. TA: Data curation, Writing–review and editing. AB: Supervision, Writing–review and editing. MV: Conceptualization, Data curation, Formal analysis, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the UK Research and Innovation (UKRI) Centres of Doctoral Training (CDT) in Artificial Intelligence for Healthcare (AI4H http://ai4health.io) (Grant No. EP/S023283/1), St George's Hospital Charity, the National Institute for Health and Care Research (NIHR) Imperial Biomedical Research Centre (BRC), and the British Heart Foundation Centre of Research Excellence at Imperial College London (RE/18/4/34215).

Acknowledgments

We acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232).

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/fnetp.2025.1488349/full#supplementary-material

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467

Alsop, D. C., Detre, J. A., Golay, X., Günther, M., Hendrikse, J., Hernandez-Garcia, L., et al. (2015). Recommended implementation of arterial spin-labeled perfusion mri for clinical applications: a consensus of the ismrm perfusion study group and the european consortium for asl in dementia. *Magnetic Reson. Med.* 73, 102–116. doi:10.1002/mrm.25197

Baik-Schneditz, N., Schwaberger, B., Mileder, L., Höller, N., Avian, A., Urlesberger, B., et al. (2021). Cardiac output and cerebral oxygenation in term neonates during neonatal transition. *Children* 8, 439. doi:10.3390/children8060439

Bajaj, C., McLennan, L., Andeen, T., and Roy, A. (2023). Recipes for when physics fails: recovering robust learning of physics informed neural networks. *Mach. Learn. Sci. Technol.* 4, 015013. doi:10.1088/2632-2153/acb416

Bashan, A., Bartsch, R. P., Kantelhardt, J. W., Havlin, S., and Ivanov, P. C. (2012). Network physiology reveals relations between network topology and physiological function. *Nat. Commun.* 3, 702. doi:10.1038/ncomms1705

Buxton, R. B., Frank, L. R., Wong, E. C., Siewert, B., Warach, S., and Edelman, R. R. (1998). A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magnetic Reson. Med.* 40, 383–396. doi:10.1002/mrm.1910400308

Candia-Rivera, D., Chavez, M., and de Vico Fallani, F. (2024). Measures of the coupling between fluctuating brain network organization and heartbeat dynamics. *Netw. Neurosci.* 8, 557–575. doi:10.1162/netn_a_00369

Castle-Kirszbaum, M., Parkin, W. G., Goldschlager, T., and Lewis, P. M. (2022). Cardiac output and cerebral blood flow: a systematic review of cardio-cerebral coupling. *J. Neurosurg. Anesthesiol.* 34, 352–363. doi:10.1097/ANA.0000000000000768

Chen, Z., Liu, Y., and Sun, H. (2021). Physics-informed learning of governing equations from scarce data. *Nat. Commun.* 12, 6136. doi:10.1038/s41467-021-26434-1

Claassen, J. A., Thijssen, D. H., Panerai, R. B., and Faraci, F. M. (2021). Regulation of cerebral blood flow in humans: physiology and clinical implications of autoregulation. *Physiol. Rev.* 101, 1487–1559. doi:10.1152/physrev.00022.2020

Counsell, S. J., Arichi, T., Arulkumaran, S., and Rutherford, M. A. (2019). Fetal and neonatal neuroimaging. *Handb. Clin. neurology* 162, 67–103. doi:10.1016/B978-0-444-64029-1.00004-7

De Silvestro, A., Natalucci, G., Feldmann, M., Hagmann, C., Nguyen, T. D., Coraj, S., et al. (2024). Effects of hemodynamic alterations and oxygen saturation on cerebral perfusion in congenital heart disease. *Pediatr. Res.* 96, 990–998. doi:10.1038/s41390-024-03106-6

De Silvestro, A. A., Kellenberger, C. J., Gosteli, M., O'Gorman, R., and Knirsch, W. (2023). Postnatal cerebral hemodynamics in infants with severe congenital heart disease: a scoping review. *Pediatr. Res.* 94, 931–943. doi:10.1038/s41390-023-02543-z

Detre, J. A., Rao, H., Wang, D. J., Chen, Y. F., and Wang, Z. (2012). Applications of arterial spin labeled mri in the brain. *J. Magnetic Reson. Imaging* 35, 1026–1037. doi:10.1002/jmri.23581

de Vries, L., van Herten, R. L., Hoving, J. W., Išgum, I., Emmer, B. J., Majoie, C. B., et al. (2023). Spatio-temporal physics-informed learning: a novel approach to ct perfusion analysis in acute ischemic stroke. *Med. image Anal.* 90, 102971. doi:10.1016/j.media.2023.102971

Dubois, J., Alison, M., Counsell, S. J., Hertz-Pannier, L., Hüppi, P. S., and Benders, M. J. (2021). Mri of the neonatal brain: a review of methodological challenges and neuroscientific advances. *J. Magnetic Reson. Imaging* 53, 1318–1343. doi:10.1002/jmri.27192

Ghalambaz, M., Sheremet, M. A., Khan, M. A., Raizah, Z., and Shafi, J. (2024). Physics-informed neural networks (pinns): application categories, trends and impact. *Int. J. Numer. Methods Heat and Fluid Flow* 34, 3131–3165. doi:10.1108/hff-09-2023-0568

Gordan, R., Gwathmey, J. K., and Xie, L.-H. (2015). Autonomic and endocrine control of cardiovascular function. *World J. Cardiol.* 7, 204–214. doi:10.4330/wjc.v7.i4.204

Hales, P. W., Pfeuffer, J., and A Clark, C. (2020). Combined denoising and suppression of transient artifacts in arterial spin labeling mri using deep learning. *J. Magnetic Reson. Imaging* 52, 1413–1426. doi:10.1002/jmri.27255

Hernandez-Garcia, L., Aramendía-Vidaurreta, V., Bolar, D. S., Dai, W., Fernández-Seara, M. A., Guo, J., et al. (2022). Recent technical developments in asl: a review of the state of the art. *Magnetic Reson. Med.* 88, 2021–2042. doi:10.1002/mrm.29381

Herrero Martin, C., Oved, A., Chowdhury, R. A., Ullmann, E., Peters, N. S., Bharath, A. A., et al. (2022). Ep-pinns: cardiac electrophysiology characterisation using physics-informed neural networks. *Front. Cardiovasc. Med.* 8, 768419. doi:10.3389/fcvm.2021.768419

Ishida, S., Fujiwara, Y., Takei, N., Kimura, H., and Tsujikawa, T. (2024). Comparison between supervised and physics-informed unsupervised deep neural networks for estimating cerebral perfusion using multi-delay arterial spin labeling mri. *NMR Biomed.* 37, e5177. doi:10.1002/nbm.5177

Jammal Salameh, L., Bitzenhofer, S. H., Hanganu-Opatz, I. L., Dutschmann, M., and Egger, V. (2024). Blood pressure pulsations modulate central neuronal activity via mechanosensitive ion channels. *Science* 383, eadk8511. doi:10.1126/science.adk8511

Kamali, A., Sarabian, M., and Laksari, K. (2023). Elasticity imaging using physics-informed neural networks: spatial discovery of elastic modulus and Poisson's ratio. *Acta Biomater.* 155, 400–409. doi:10.1016/j.actbio.2022.11.024

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nat. Rev. Phys.* 3, 422–440. doi:10.1038/s42254-021-00314-5

Kissas, G., Yang, Y., Hwuang, E., Witschey, W. R., Detre, J. A., and Perdikaris, P. (2020). Machine learning in cardiovascular flows modeling: predicting arterial blood pressure from non-invasive 4d flow mri data using physics-informed neural networks. *Comput. Methods Appl. Mech. Eng.* 358, 112623. doi:10.1016/j.cma.2019.112623

Korom, M., Camacho, M. C., Filippi, C. A., Licandro, R., Moore, L. A., Dufford, A., et al. (2022). Dear reviewers: responses to common reviewer critiques about infant neuroimaging studies. *Dev. Cogn. Neurosci.* 53, 101055. doi:10.1016/j.dcn.2021.101055

Krishnapriyan, A., Gholami, A., Zhe, S., Kirby, R., and Mahoney, M. W. (2021). Characterizing possible failure modes in physics-informed neural networks. *Adv. Neural Inf. Process. Syst.* 34, 26548–26560. doi:10.5555/3540261.3542294

Lindner, T., Bolar, D. S., Achten, E., Barkhof, F., Bastos-Leite, A. J., Detre, J. A., et al. (2023). Current state and guidance on arterial spin labeling perfusion mri in clinical neuroimaging. *Magnetic Reson. Med.* 89, 2024–2047. doi:10.1002/mrm.29572

Lu, L., Meng, X., Mao, Z., and Karniadakis, G. E. (2021). DeepXDE: a deep learning library for solving differential equations. *SIAM Rev.* 63, 208–228. doi:10.1137/19M1274067

McQuillen, P. S., Goff, D. A., and Licht, D. J. (2010). Effects of congenital heart disease on brain development. *Prog. Pediatr. Cardiol.* 29, 79–85. doi:10.1016/j.ppedcard.2010.06.011

Meng, L., Hou, W., Chui, J., Han, R., and Gelb, A. W. (2015). Cardiac output and cerebral blood flow: the integrated regulation of brain perfusion in adult humans. *Anesthesiology* 123, 1198–1208. doi:10.1097/ALN.0000000000000872

Min, Z., Baum, Z. M., Saeed, S. U., Emberton, M., Barratt, D. C., Taylor, Z. A., et al. (2023). "Non-rigid medical image registration using physics-informed neural networks," in *International conference on information processing in medical imaging* (Springer), 601–613.

Moradi, H., Al-Hourani, A., Concilia, G., Khoshmanesh, F., Nezami, F. R., Needham, S., et al. (2023). Recent developments in modeling, imaging, and monitoring of cardiovascular diseases using machine learning. *Biophys. Rev.* 15, 19–33. doi:10.1007/s12551-022-01040-7

Pertuz, S., Puig, D., and Garcia, M. A. (2013). Analysis of focus measure operators for shape-from-focus. *Pattern Recognit.* 46, 1415–1432. doi:10.1016/j.patcog.2012.11.011

Petersen, E., Zimine, I., Ho, Y. L., and Golay, X. (2006). Non-invasive measurement of perfusion: a critical review of arterial spin labelling techniques. *Br. J. radiology* 79, 688–701. doi:10.1259/bjr/67705974

Pilar, P., and Wahlström, N. (2024). "Physics-informed neural networks with unknown measurement noise," in *6th annual learning for dynamics and control conference* (New York, United States: PMLR), 235–247.

Pinto, J., Blockley, N. P., Harkin, J. W., and Bulte, D. P. (2023). Modelling spatiotemporal dynamics of cerebral blood flow using multiple-timepoint arterial spin labelling mri. *Front. Physiology* 14, 1142359. doi:10.3389/fphys.2023.1142359

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378, 686–707. doi:10.1016/j.jcp.2018.10.045

Rossi, A., Mikail, N., Bengs, S., Haider, A., Treyer, V., Buechel, R. R., et al. (2022). Heart–brain interactions in cardiac and brain diseases: why sex matters. *Eur. heart J.* 43, 3971–3980. doi:10.1093/eurheartj/ehac061

Rotkopf, L. T., Ziener, C. H., von Knebel-Doeberitz, N., Wolf, S. D., Hohmann, A., Wick, W., et al. (2024). A physics-informed deep learning framework for dynamic susceptibility contrast perfusion mri. *Med. Phys.* 51, 9031–9040. doi:10.1002/mp.17415

Sahli Costabal, F., Yang, Y., Perdikaris, P., Hurtado, D. E., and Kuhl, E. (2020). Physics-informed neural networks for cardiac activation mapping. *Front. Phys.* 8, 42. doi:10.3389/fphy.2020.00042

Sarabian, M., Babaee, H., and Laksari, K. (2022). Physics-informed neural networks for brain hemodynamic predictions using medical imaging. *IEEE Trans. Med. imaging* 41, 2285–2303. doi:10.1109/TMI.2022.3161653

Silverman, A., and Petersen, N. H. (2020). *Physiology, cerebral autoregulation.* StatPearls Publishing

Tahsili-Fahadan, P., and Geocadin, R. G. (2017). Heart–brain axis: effects of neurologic injury on cardiovascular function. *Circulation Res.* 120, 559–572. doi:10.1161/CIRCRESAHA.116.308446

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. (2020). Deep learning on image denoising: an overview. *Neural Netw.* 131, 251–275. doi:10.1016/j.neunet.2020.07.025

van Herten, R. L., Chiribiri, A., Breeuwer, M., Veta, M., and Scannell, C. M. (2022). Physics-informed neural networks for myocardial perfusion mri quantification. *Med. Image Anal.* 78, 102399. doi:10.1016/j.media.2022.102399

Varela, M., Hajnal, J. V., Petersen, E. T., Golay, X., Merchant, N., and Larkman, D. J. (2011). A method for rapid *in vivo* measurement of blood t1. *NMR Biomed.* 24, 80–88. doi:10.1002/nbm.1559

Varela, M., Petersen, E. T., Golay, X., and Hajnal, J. V. (2015). Cerebral blood flow measurements in infants using look–locker arterial spin labeling. *J. Magnetic Reson. Imaging* 41, 1591–1600. doi:10.1002/jmri.24716

Wang, S., Yu, X., and Perdikaris, P. (2022). When and why pinns fail to train: a neural tangent kernel perspective. *J. Comput. Phys.* 449, 110768. doi:10.1016/j.jcp.2021.110768

Zimmermann, F. F., Kolbitsch, C., Schuenke, P., and Kofler, A. (2024). Pinqi: an end-to-end physics-informed approach to learned quantitative mri reconstruction. *IEEE Trans. Comput. Imaging* 10, 628–639. doi:10.1109/tci.2024.3388869

Zou, Z., Meng, X., and Karniadakis, G. E. (2024). Correcting model misspecification in physics-informed neural networks (pinns). *J. Comput. Phys.* 505, 112918. doi:10.1016/j.jcp.2024.112918

Zou, Z., Meng, X., and Karniadakis, G. E. (2025). Uncertainty quantification for noisy inputs–outputs in physics-informed neural networks and neural operators. *Comput. Methods Appl. Mech. Eng.* 433, 117479. doi:10.1016/j.cma.2024.117479

TYPE Original Research
PUBLISHED 16 April 2025
DOI 10.3389/fdgth.2025.1558589

OPEN ACCESS

EDITED BY
Xiaofeng Li,
Heilongjiang International University, China

REVIEWED BY
Heyam A. Marzog,
Al-Furat Al-Awsat Technical University, Iraq
Katya Arquilla,
University of Colorado Boulder, United States

*CORRESPONDENCE
Abhirup Banerjee
✉ abhirup.banerjee@eng.ox.ac.uk

RECEIVED 10 January 2025
ACCEPTED 25 March 2025
PUBLISHED 16 April 2025

CITATION
Sang Y, Banerjee A, Beetz M and Grau V (2025) Deep conditional generative model for personalization of 12-lead electrocardiograms and cardiovascular risk prediction.
Front. Digit. Health 7:1558589.
doi: 10.3389/fdgth.2025.1558589

Deep conditional generative model for personalization of 12-lead electrocardiograms and cardiovascular risk prediction

Yuling Sang[1,2], Abhirup Banerjee[2,3]*, Marcel Beetz[2] and Vicente Grau[2]

[1]Centre for Computational Biology, Duke-NUS Medical School, Singapore, Singapore, [2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom, [3]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom

Background: 12-lead electrocardiograms (ECGs) are a cornerstone for diagnosing and monitoring cardiovascular diseases (CVDs). They play a key role in detecting abnormalities such as arrhythmias and myocardial infarction, enabling early intervention and risk stratification. However, traditional analysis relies heavily on manual interpretation, which is time-consuming and expertise-dependent. Moreover, existing machine learning models often lack personalization, as they fail to integrate subject-specific anatomical and demographic information. Advances in deep generative models offer an opportunity to overcome these challenges by synthesizing personalized ECGs and extracting clinically relevant features for improved risk assessment.
Methods: We propose a conditional Variational Autoencoder (cVAE) framework to generate realistic, subject-specific 12-lead ECGs by incorporating demographic metadata, anatomical heart features, and ECG electrodes' positions as conditioning factors. This allows for physiologically consistent and personalized ECG synthesis. Furthermore, we introduce a revised Cox proportional-hazards regression model that utilizes the latent embeddings learned by the cVAE to predict future CVD risk. This approach not only enhances the interpretability of ECG-derived risk factors but also demonstrates the potential of deep generative models in personalized cardiac assessment.
Results: Our model is trained and validated on the UK Biobank dataset and *in silico* simulation data. By incorporating heart position and electrodes' positions, the generated ECGs demonstrate strong consistency with *in silico* simulations, providing insights into the relationship between cardiac anatomy and ECG morphology. Furthermore, our CVD risk prediction model achieves a C-index of 0.65, indicating that ECG signals, together with demographic and anatomical information, contain valuable prognostic information for stratifying subjects based on future cardiovascular risk.
Conclusion: This work marks a significant advancement in ECG analysis by providing a conditional VAE framework that not only improves ECG generation but also enriches our understanding of the relationship between ECG patterns and subject-specific information. Importantly, our approach enables clinically significant information to be extracted from 12-lead ECGs, providing valuable insights for predicting future CVD risks.

KEYWORDS

cardiac MRI, cardiovascular disease, cardiovascular risk prediction, ECG electrodes, ECG generation, variational autoencoder

1 Introduction

The electrocardiogram (ECG) is a well established, non-invasive diagnostic tool that records the electrical activity of the heart over time (1). However, the manual analysis of ECG data can be a time-consuming and labor-intensive process, requiring significant expertise in interpreting complex patterns and abnormalities in the heart's electrical activity. With the increasing use of wearable devices and other monitoring technologies, large volumes of ECG data can be generated on a daily basis (2), further exacerbating the challenge of manual analysis. As a result, there is a need for automatic techniques to facilitate the efficient diagnosis of heart diseases using the ECG.

Machine learning has emerged as a powerful tool for enabling automated analysis in a wide range of ECG-based tasks (3–9). While machine learning techniques have shown great promise, many of these methods require large amounts of labeled data to effectively train the model. This poses a significant challenge as obtaining and annotating large datasets can be time-consuming, expensive, and resource-intensive. Also, class imbalance is another common issue in ECG datasets, as certain cardiac abnormalities may be relatively rare compared to normal ECG patterns, which can lead to biased model performance (10). Furthermore, preserving patient privacy is another critical aspect of medical data sharing and usage, especially in the context of ECG data, which may contain personally identifiable and sensitive health information (11).

Researchers have tried to solve these problems through data augmentation. Classic data augmentation methods such as performing translation and adding noise can only obtain limited new additional information, which may lead to overfitting during the training process. In order to truly augment the dataset, deep generative models have attracted attention in recent years for the generation of high-quality synthetic medical data, and been applied successfully in ECG research. Previous deep generative models (12–14) have mainly focused on only single-lead ECG generation and lack the introduction of subject characteristics. 12-lead ECGs are the clinical gold standard, providing comprehensive spatial information about cardiac conduction, and incorporation of demographic and physiological features is crucial for understanding the relationship between ECG morphology and subject information. The inability to generate physiologically consistent multi-lead signals significantly restricts the applicability of these models in personalized cardiac assessments, as key inter-lead relationships and subject-specific variations are not considered.

Traditional simulation methods, such as the Extracellular-Membrane-Intracellular (EMI) model or the work of Mincholé et al. (15), which utilized computer simulation with torso-ventricular anatomical models to investigate the impacts of ventricular and torso anatomy on 12-lead ECGs, hypothesize that geometrical factors, including ventricular anatomy, heart orientation, location, and torso anatomy, differentially influence QRS complexes in 12-lead ECGs. Although these traditional biophysically-based models can be very precise, they are computationally intensive, with simulations requiring up to several hours (16), whereas generative models can synthesize ECG signals in milliseconds per sample.

Our study aims to bridge these gaps by introducing a conditional Variational Autoencoder (cVAE) framework that generates 12-lead ECGs conditioned on anatomical features. In our previous work (17), we included subject metadata and anatomical characteristics, such as heart positions and orientations, from cardiac Magnetic Resonance Imaging (MRI) to develop a cVAE model that can generate realistic 12-lead ECGs with ability to capture useful features from different conditions. However, the generated conditional ECGs only partially align with the *in silico* data, likely due to the absence of torso structural information in the model.

To address this limitation, in this study, we incorporate ECG electrode locations as additional input features. A widely used configuration for ECG measurement involves 10 electrodes: 4 electrodes placed on the limbs [left arm (LA), right arm (RA), left leg (LL), and right leg (RL)] and 6 electrodes positioned on the chest (V1 to V6). These chest electrodes provide detailed spatial information about the heart's electrical activity, enabling the formation of 12 leads and establishing a strong connection between the torso structure and ECG signals. With the recent development of automated 3D torso reconstruction (18, 19), we are able to obtain the precise electrodes' positions from each subject's clinical MRI. This additional information provides valuable constraints to the model, allowing it to generate ECGs that are not only realistic but also anatomically and physiologically consistent.

In order to demonstrate the efficacy of the latent representation achieved from the VAE architecture, we extend the model to perform future cardiovascular disease (CVD) risk prediction. The majority of contemporary algorithms focusing on CVD risk prediction are based on a limited set of subject attributes, e.g., age, smoking history, and blood pressure. Recently, efforts have been made to investigate a broader range of risk predictors, encompassing interaction terms and employing more sophisticated machine learning techniques to model CVD risk (20). However, these studies have only considered tabular data, neglecting other potential information sources such as ECG or MRI. Recent studies (21–23) have increasingly shown that ECG abnormalities are a promising predictor of CVD risk, making the direct use of ECG signals an attractive direction for risk stratification. However, most previous approaches have relied solely on ECG data without incorporating the underlying anatomical context. Specifically, variations in heart position and orientation can substantially alter ECG morphology by shifting the electrical axis and modifying the amplitude and duration of key waveforms (15, 19). If these anatomical effects are not accounted for, normal variations in heart position may be misinterpreted as pathological changes or, conversely, true abnormalities might be obscured. By incorporating heart position and orientation, our model can disentangle these anatomical influences from disease-related signals. Therefore, our work explores the novel integration of heart data with ECG signals, aiming not only to generate more realistic ECGs but also to enhance the accuracy of CVD risk prediction by incorporating critical anatomical context.

Our study makes the following key contributions:

1. We develop a novel cVAE framework capable of generating 12-lead ECGs and incorporate patient-specific conditions.
2. We demonstrate that incorporating heart position and electrodes placement significantly improves the fidelity of synthetic ECG signals, capturing inter-lead dependencies and individual variability.
3. We introduce a revised Cox proportional-hazards model, leveraging ECG-derived latent embeddings to enhance CVD risk prediction.
4. ECG signals, combined with anatomical context, can stratify subjects based on their future cardiovascular risk (C-index = 0.65), providing valuable insights for personalized cardiac assessments.

2 Materials and methods

2.1 ECG dataset

Our research has been conducted using the UK Biobank Resource under Application Number "40161" (24). In total, we have ECG files from 37,508 volunteers, together with their personal information including age, sex, BMI, and their clinical imaging information.

Each ECG file in the UK Biobank dataset contains a 10-s sample recorded at 500 Hz with 5,000 data points per lead. Additionally, UK Biobank provides a median beat waveform, which is computed by extracting individual heartbeats from the 10-s segment, aligning them, and calculating the median waveform across all beats. This median beat contains approximately 600 data points and serves as a representative single heartbeat, The majority of our experiments are performed on the shorter median data, since the averaging process can help to reduce noise and artifacts in the signal, providing a cleaner and accurate representation of the cardiac activity. It not only allows us to focus on specific features of the ECG, such as the QRS complex, without the confounding effects of beats variability in the longer recording, but requires less computational power and time as well. The ECG data require some additional pre-processing to remove artifacts like baseline drift, which was removed using a finite impulse response band-pass filter between 3–45 Hz, inspired by an entry to the Computing in Cardiology (CinC) 2017 challenge (25).

The age and sex information of the subjects are included in the UK Biobank ECG files. The BMI can be located within the "Body Size Measures" category in the UK Biobank, accessible through each subject's unique identification number.

The UK Biobank dataset we use includes 21,083 cardiac MRI cases in total, and they were acquired at the same date as the ECG acquisitions (26). These cardiac MRI are used to calculate subject-specific information, including heart positions, orientations and electrode positions.

2.2 CVD risk prediction dataset

In this project, we define CVD as a composite of any of the following ICD-10 diagnosis codes: I20 (angina pectoris), I21 (acute myocardial infarction), I22 (subsequent myocardial infarction), I23 (certain current complications following acute myocardial infarction), I24 (other acute ischaemic heart diseases), I25 (chronic ischaemic heart disease), and I50 (heart failure). This is similar to the research of Alaa et al. (20), but we exclude I60–I69 (cerebrovascular diseases), as we assume that the link between ECG and cerebrovascular disease is relatively weak. We also exclude vascular dementia, since at the time of our study we do not have access to its ICD-10 code. We apply our model only on the cases whose CVD event date is posterior to the ECG acquisition date, which we refer as incident cases. We identify all subjects for which a CVD event was recorded before ECG acquisition as prevalent CVD cases (27). The diagram of our dataset preparation is shown in Figure 1.

In total, we have 37,508 subjects with successful ECG recordings. As detailed in Section 2.1, a finite impulse response band-pass filter is applied to correct baseline drift in the signals. However, this method does not address short peak artifacts, which can significantly affect our model training. To mitigate this issue, we remove all signals with absolute amplitudes exceeding 800 mV/100 in any lead, resulting in the exclusion of 853 subjects. A further 25 subjects are excluded due to missing CVD diagnoses. Next, we exclude 2,917 subjects with prevalent CVD diagnosis from our dataset leaving 33,713 subjects. Among them, we separately have 925 cases with incident CVD diagnosis and 32,788 healthy subjects with no CVD records at the time of this study.

We allocate 80% of each healthy and CVD group into the training set and the remaining 20% into the test set for CVD risk prediction. This stratification was applied separately to each of the 7 CVD subtypes, ensuring that their proportions remained consistent across both sets. By maintaining balanced representation, we reduce the potential for certain diseases to be over- or under-represented, thereby improving model accuracy and generalizability.

2.3 Heart position and orientation

The heart position and orientation data are calculated using information from the cardiac MRI. In general, a standard cardiac MRI acquisition includes a stack of 2D short-axis (SAX) slices, which cover the left and right ventricles from apex to base, as well as a 2-chamber long axis (LAX) slice and a 4-chamber LAX slice (28). As shown in Figure 2A, we define the heart position as the intersection between three planes: 2-chamber LAX plane, 4-chamber LAX plane, and the middle plane of the SAX view stack. The definition of a plane is 3D space is given by Equation 1:

$$\mathbf{n} \cdot (\mathbf{X} - P) = 0 \qquad (1)$$

where:

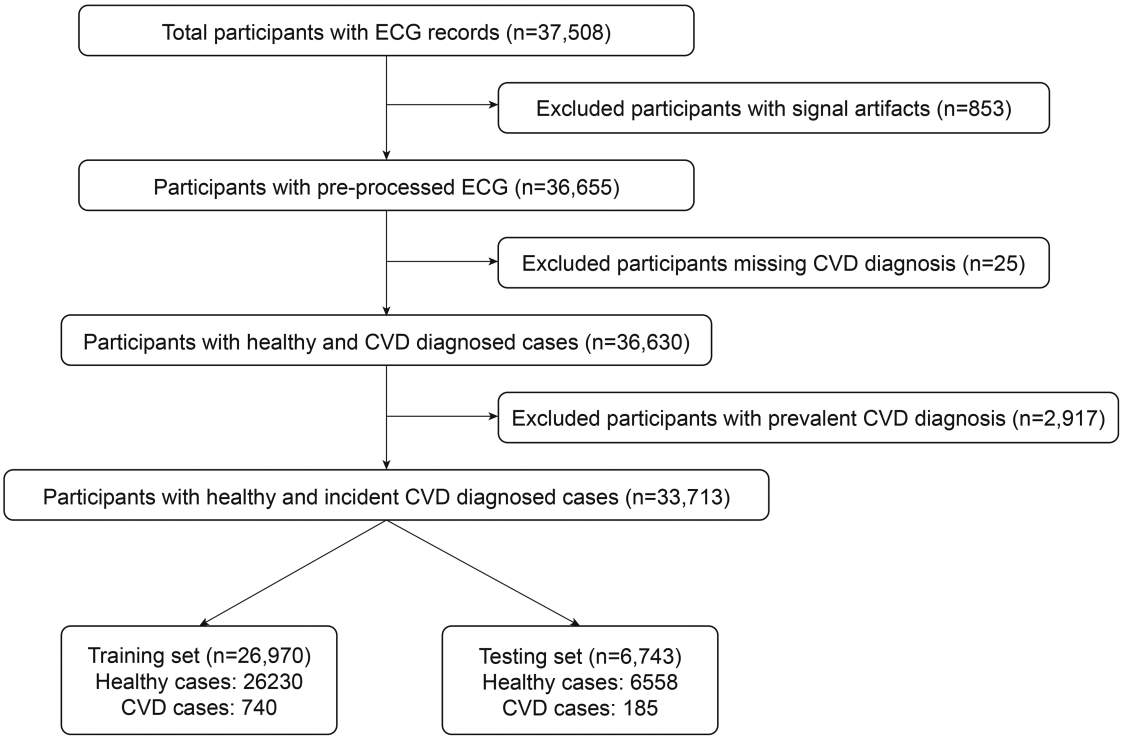

Total participants with ECG records (n=37,508)
Excluded participants with signal artifacts (n=853)
Participants with pre-processed ECG (n=36,655)
Excluded participants missing CVD diagnosis (n=25)
Participants with healthy and CVD diagnosed cases (n=36,630)
Excluded participants with prevalent CVD diagnosis (n=2,917)
Participants with healthy and incident CVD diagnosed cases (n=33,713)
Training set (n=26,970)
Healthy cases: 26230
CVD cases: 740
Testing set (n=6,743)
Healthy cases: 6558
CVD cases: 185

FIGURE 1
CVD risk prediction dataset preparation diagram.

- $\mathbf{n} \in \mathbb{R}^3$ is the normal vector of the plane;
- $\mathbf{X} = (x, y, z) \in \mathbb{R}^3$ is an arbitrary point on the plane; and
- $P = (P_x, P_y, P_z)$ is a known point on the plane, extracted from the DICOM metadata.

The specific plane equations for the three anatomical planes are shown in Equations 2–4:

$$\mathbf{n}_{\text{SAX}} \cdot (\mathbf{X} - P^{\text{SAX}}) = 0 \tag{2}$$

$$\mathbf{n}_{\text{2CH}} \cdot (\mathbf{X} - P^{\text{2CH}}) = 0 \tag{3}$$

$$\mathbf{n}_{\text{4CH}} \cdot (\mathbf{X} - P^{\text{4CH}}) = 0 \tag{4}$$

where $\mathbf{n}_{\text{SAX}}, \mathbf{n}_{\text{2CH}}, \mathbf{n}_{\text{4CH}}$ are the normal vectors of the SAX, 2-chamber LAX, and 4-chamber LAX planes, respectively. $P^{\text{SAX}}, P^{\text{2CH}}, P^{\text{4CH}}$ are the image position points for each plane. By solving this system of three linear equations, we obtain the heart's center position, as shown in Equation 5:

$$P_{\text{heart}} = (x_h, y_h, z_h) = \text{Intersection}(\text{SAX},\ \text{2CH},\ \text{4CH}) \tag{5}$$

The heart orientation is defined relative to the standard anatomical coordinate system using a new heart-specific coordinate system based on the SAX and 4-chamber LAX planes. This coordinate system is denoted as $(\mathbf{e}_X, \mathbf{e}_Z, \mathbf{e}_Y)$. The new X-axis is computed as the normalized intersection vector between the SAX and 4-chamber LAX planes:

$$\mathbf{e}_X = \frac{\mathbf{L}_{\text{SAX-4CH}}}{\|\mathbf{L}_{\text{SAX-4CH}}\|} \tag{6}$$

where $\mathbf{L}_{\text{SAX-4CH}} = \mathbf{n}_{\text{SAX}} \times \mathbf{n}_{\text{4CH}}$ is the direction vector of the line formed by the intersection of the SAX and 4-chamber LAX planes. $\|\mathbf{L}_{\text{SAX-4CH}}\|$ is the vector norm, ensuring $\mathbf{e}_X$ is a unit vector.

The new Z-axis is chosen to be perpendicular to the 4-chamber LAX plane, while ensuring it remains orthogonal to $\mathbf{e}_X$:

$$\mathbf{e}_Z = \mathbf{n}_{\text{4CH}} - (\mathbf{n}_{\text{4CH}} \cdot \mathbf{e}_X)\mathbf{e}_X \tag{7}$$

where $\mathbf{n}_{\text{4CH}}$ is the normal vector of the 4-chamber LAX plane. The term $(\mathbf{n}_{\text{4CH}} \cdot \mathbf{e}_X)\mathbf{e}_X$ removes the component of $\mathbf{n}_{\text{4CH}}$ that is parallel to $\mathbf{e}_X$, ensuring orthogonality.

The new Y-axis is computed as the cross-product of $\mathbf{e}_X$ and $\mathbf{e}_Z$:

$$\mathbf{e}_Y = \mathbf{e}_X \times \mathbf{e}_Z \tag{8}$$

Equations 6–8 ensure that $(\mathbf{e}_X, \mathbf{e}_Y, \mathbf{e}_Z)$ forms a right-handed orthonormal coordinate system. The Euler angles describe the rotation between the heart coordinate system $(\mathbf{e}_X, \mathbf{e}_Y, \mathbf{e}_Z)$ and the standard anatomical coordinate system $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, as described in

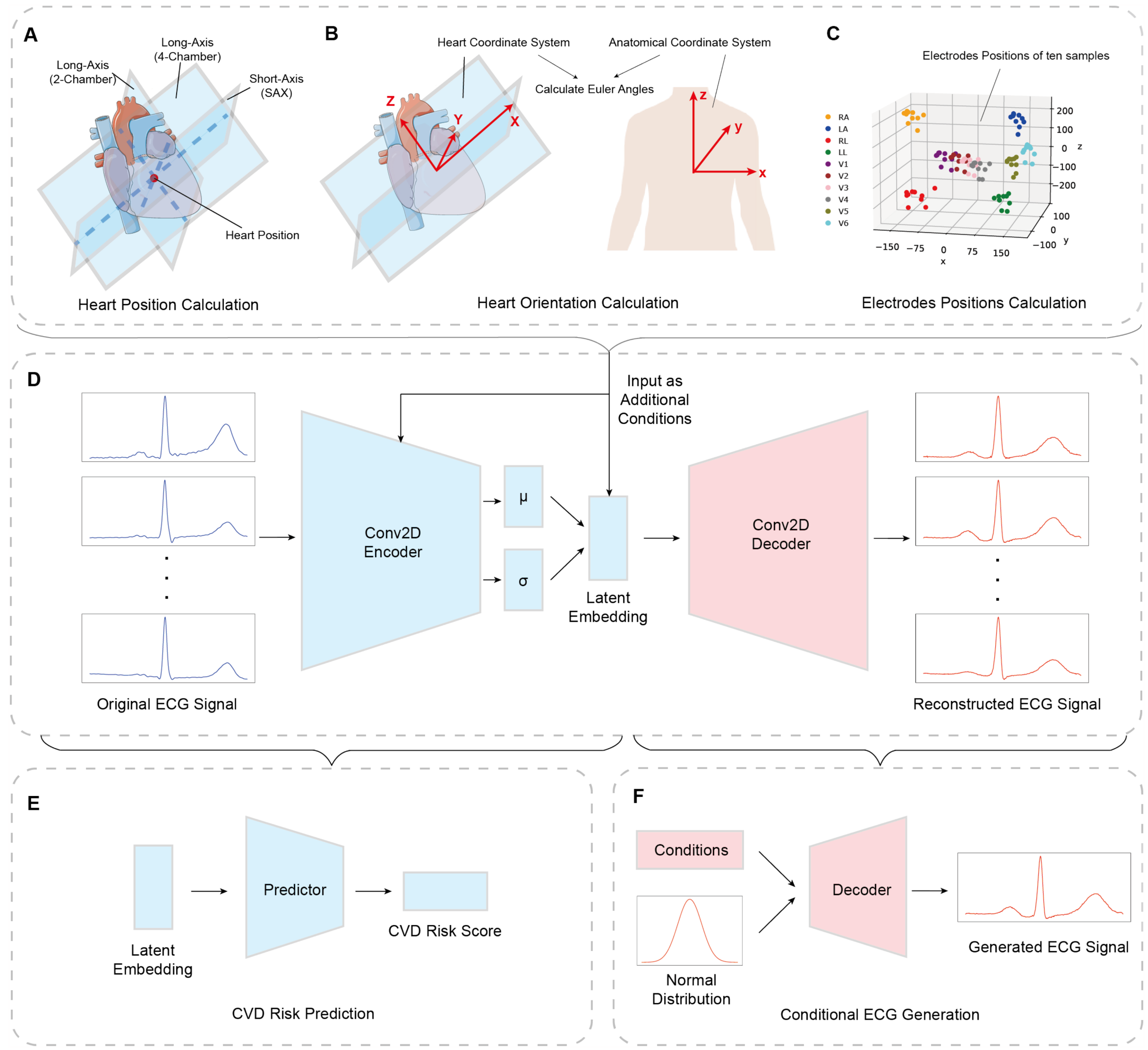


FIGURE 2
Overall pipeline for utilizing heart and torso information in conditional ECG generation and CVD risk prediction. **(A)** Heart position is calculated by the intersection of 2-chamber view, 4-chamber view, and middle short-axis view. **(B)** Heart orientation is represented by Euler angles between heart coordinate system and anatomical coordinate system. **(C)** Electrodes' positions are achieved from cardiac MRI (18) and transformed to heart coordinate system. **(D)** The conditional VAE architecture with the heart and torso information as additional condition inputs added to the first fully-connected layer of encoder and latent space. **(E)** The CVD risk prediction model. An additional predictor is concatenated to the latent embedding, which provides the risk score to realize the revised Cox proportional hazard regression model. **(F)** The conditional ECG generation is performed by trained decoder, which takes random sampling from normal distribution and condition inputs.

Equations 9–11:

$$\alpha = \cos^{-1}\left(\frac{\mathbf{e}_X \cdot \mathbf{x}}{\|\mathbf{e}_X\| \cdot \|\mathbf{x}\|}\right) \tag{9}$$

$$\beta = -\cos^{-1}\left(\frac{\mathbf{e}_Z \cdot \mathbf{z}}{\|\mathbf{e}_Z\| \cdot \|\mathbf{z}\|}\right) \tag{10}$$

$$\gamma = -\cos^{-1}\left(\frac{\mathbf{e}_X \cdot \mathbf{N}}{\|\mathbf{e}_X\| \cdot \|\mathbf{N}\|}\right) \tag{11}$$

where **N** is the normal vector of the anatomical XOY plane.

2.4 Electrode positions

We use the work of Smith et al. (19) for estimating the electrodes' positions for each subject. The method applies a U-net deep learning network for automated torso segmentation and contour extraction from the localizer and scout cardiac MRI from the UK Biobank dataset (18). The undesired section including head, neck, and arms and potential artifacts such as shadow regions are removed using a preprocessing algorithm. Finally, a statistical shape model is used over sparse 3D contours to generate 3D torso meshes, with the electrodes' positions estimated on the 3D torso meshes.

Due to the relative slow speed of this algorithm, which usually takes 30–60 min for one case, we reconstruct a total of 1,834 3D torso meshes, and measure the ten electrodes' positions for standard 12-lead ECGs, which include four limb electrodes including left arm, right arm, left leg, and right leg, and six precordial electrodes corresponding to six precordial leads. Figure 2C presents electrodes' locations of ten sample cases from our training set.

The electrodes generated from torso meshes are 3D variables located in the anatomical coordinate system presented in Figure 2B. In order for each subject's location information to be more accurately comparable and representative of the anatomical characteristics of the heart, we utilize the heart position and orientation calculated before, to transfer the locations from anatomical coordinate system to heart coordinate system. In this way we capture the corresponding relationship between the electrode coordinates and the heart coordinates while treating the heart coordinates as the origin. Therefore, the electrodes' positions information is able to contain both torso and heart features.

2.5 Conditional VAE architecture

Assuming that the original data set is $\mathbf{x}$, the encoder produces a hidden variable $\mathbf{z}$ and the decoder produces the reconstructed dataset $\hat{\mathbf{x}}$. The VAE aims to learn the marginal likelihood of the input through this generative process, as defined in Equation 12:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \tag{12}$$

where ϕ, θ parameterize the distributions of the VAE encoder and decoder respectively. Here, $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior distribution of the latent variable $\mathbf{z}$ given the input $\mathbf{x}$, and $p_\theta(\mathbf{x}|\mathbf{z})$ represents the likelihood of the input given the latent variable, modeled by the decoder. Based on the evidence lower bound (ELBO), the training process of VAE uses the loss function as Equation 13:

$$\mathcal{L} = -\mathbb{E}[\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \tag{13}$$

where $-\mathbb{E}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ in our experiment is chosen as the mean-squared error between the original $\mathbf{x}$ and the reconstructed $\hat{\mathbf{x}}$, denoted as $\mathcal{L}_{\text{recons}}$. $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ represents the Kullback-Leibler (KL) divergence between predefined posterior $p(\mathbf{z}) \sim \mathcal{N}(\mu, \sigma)$ and the latent space distribution $q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mu_z, \sigma_z)$ produced by our network, denoted as $\mathcal{L}_{KL}$. The posterior $p(\mathbf{z})$ is set as a standard normal distribution for easy computation.

The structure of the cVAE is similar to VAE, except that category information $\mathbf{y}$ is added as part of the input data, which is used to control sample generation for specified categories. The modified objective function of cVAE is presented in Equation 14:

$$\mathcal{L} = -\mathbb{E}[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})\|p(\mathbf{z})) \tag{14}$$

Figure 2D shows the revised cVAE network architecture. In the encoder part, the ECG data, with the dimension of 1 × 12 × 400, is treated as the input, followed by two convolutional blocks, each of which includes a 2-dimensional convolution layer, a batch normalization layer, and an Exponential Linear Unit (ELU) activation function. Next, we have an Average Pooling layer and the output is flattened. We use two fully-connected layers to produce two 64-dimensional vectors: one is interpreted as the mean, while the other one is considered as the logarithms of the variance of 64 normal distributions. In the final stage, a sampling layer is used to get a 64-dimensional latent space sampled from the distributions mentioned above. The decoder part is symmetrical to the encoder part, which uses upsampling layers and 2-dimensional deconvolution layers, to reconstruct the 12 lead ECGs.

Physiologically, heart position, orientation, and electrode locations define the spatial relationship between the heart's electrical activity and the recording leads, thereby affecting ECG waveform morphology. To ensure that the generative model learns these dependencies, in the encoder, the conditional information is concatenated to the first fully connected layer, ensuring that the learned latent representation $\mathbf{z}$ captures the variability introduced by anatomical differences. In the decoder, the same conditional inputs are incorporated alongside $\mathbf{z}$ to modulate ECG generation, enforcing physiological consistency by reconstructing ECG waveforms that align with the given heart position, orientation, and electrode locations. Its dimension c depends on the information category: for heart position and orientation, these are three-dimensional coordinates and angles respectively, and for electrodes' positions are 10×3 dimensional coordinates.

The model is implemeted in Python3 using PyTorch. Adam optimizer (29) is used with a learning rate of 0.001. For each VAE, we assigned 80% of the dataset as the training set and the rest as test set. The batch size is set as 64, and the training process is performed for 80 epochs. We run all experiments on NVIDIA A100 Tensor Core GPU.

2.6 Risk prediction model architecture

The revised network presented in Figure 2E is the addition of an extra predictor connected to the latent space so that we are able to analyze the representations and features contained in latent space and achieve a risk score output. For the predictor, we perform all experiments on a single fully connected layer, with 64 latent space dimension as input, and one dimensional risk score as output, obtained using a sigmoid function.

The loss function of this predictor network consists of three parts. The first two parts are the same as previous sections, i.e., $\mathcal{L}_{KL}$ and $\mathcal{L}_{\text{recons}}$, as shown in Equations 13 and 14. We use the

Cox proportional hazard regression model to realize the last survival loss part. Typically in a linear Cox model, the hazard function has the form defined in Equation 15:

$$h(t, x_1, \ldots, x_m, \beta_1, \ldots, \beta_m) = h_0(t) \exp\{\beta_1 x_1 + \cdots + \beta_m x_m\} \tag{15}$$

where $h_0(t)$ is the baseline hazard function, which would correspond to a hypothetical subject whose covariate values are all zeros. The $\exp\{\beta_1 x_1 + \cdots + \beta_m x_m\}$ is called the relative risk of a subject. Predictor covariate variables, x_i , are weighted by β_i, to adjust this baseline hazard function for each subject. These weights, β', are estimated by maximising the Cox proportional hazards partial likelihood function:

$$\log \mathcal{L}(\beta) = \sum_{i=0}^{n} \delta_i \left(\beta' x_i - \log \sum_{j \in R(t_i)} e^{\beta' x_j} \right) \tag{16}$$

where x_i is the vector of predictor covariate variables, δ_i is a boolean variable indicating event status, and $R(t_i)$ is the set of subjects yet to have an event or be censored at time t for subject i. Equation 16 can be adapted for a neural network by replacing $\beta' x_i$ with the output of a network.

Therefore, in order to optimize our VAE network training for survival analysis, we replace $\beta' x_i$ with the output of our predictor, as shown in Equation 17, to form our survival loss function:

$$\mathcal{L}_{\text{surv}} = \frac{1}{N} \sum_{i=0}^{N} \delta_i \left(r_i - \log \sum_{j \in R(t_i)} e^{r_j} \right) \tag{17}$$

where N is our batch size and r_i is the sigmoid of the output of the model, i.e., the log-hazard ratio of subject i. Preliminary work with the survival model showed that the exponent term in $\mathcal{L}_{\text{surv}}$ can cause the untrained predictor head to exponentiate large numbers leading to numerical instability. To prevent this, we apply a sigmoid function to the output of the model, both ensuring that large exponents are not possible and keeping the relative order of risk for subjects unchanged, since the sigmoid is monotonically increasing.

Therefore, the loss for our overall model is written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recons}} + \mathcal{L}_{KL} + \mathcal{L}_{\text{surv}}. \tag{18}$$

As this study represents an initial investigation, the three loss terms in Equation 18 are assigned equal weights. In the future, techniques such as grid search or other hyperparameter optimization methods can be utilized to systematically determine the optimal weight configuration.

3 Results

3.1 ECG conditioned on heart position and orientation

We used all 21,083 UK Biobank cases that include both ECG and MRI data to train our model. After training, by modifying the conditioning inputs (i.e., heart position and orientation), we generated synthetic ECGs reflecting various cardiac poses, which we then compared with the simulation trends reported by Mincholé et al. The conditional ECG generation is performed by the pre-trained decoder which takes random samples from a normal distribution and conditional inputs, shown in Figure 2F.

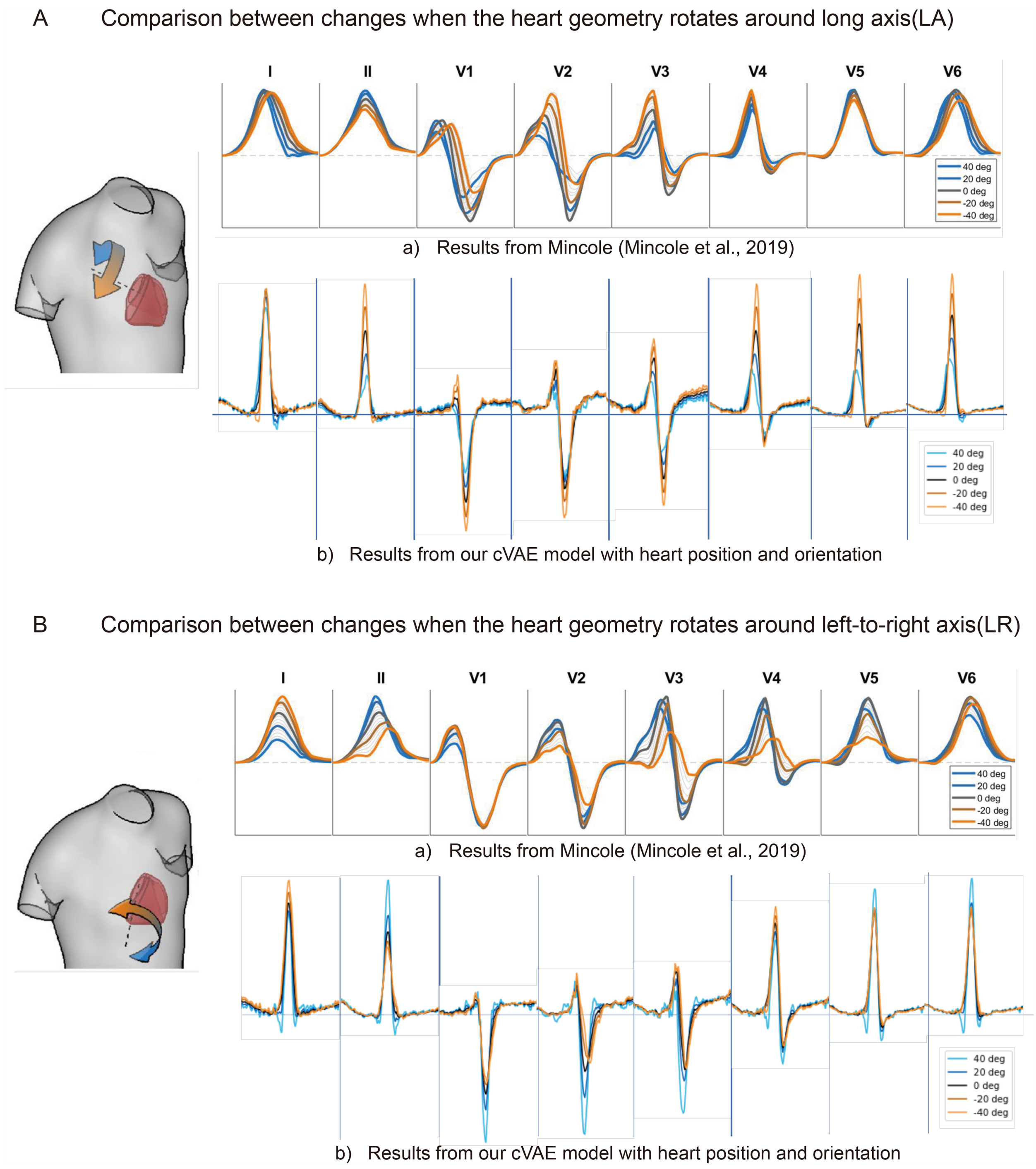
Figures 3 and 4 show the results that reflect the learned effect of heart rotation and translation, respectively. For rotation, we first rotate the heart along the long axis, which is the Z axis of the heart coordinate system, and then left-to-right ventricle axis, which is the Y axis of the heart coordinate system. For translation, we move the heart along the lateral and cranio-caudal directions, which would be represented as the heart position coordinate (x, y, z) changes, so that moving along the lateral direction and cranio-caudal direction means changing the value of x and z respectively.

We compare our generated ECGs with the work of Mincholé et al. (15). As shown in Figure 3A(b), rotation along the long axis influences R, S, and T waves in almost all the ECG leads. The heart rotated more counterclockwise results in an increase in the amplitude of these waves. After comparison, we find only lead V4 completely agrees with the result of Mincholé et al. (15) [Figure 3A(a)], while leads V1 to V3 have the same change on R wave but the opposite on S wave. The rest of the leads show different features, for in results of Mincholé et al. (15) long axis rotation exerts a limited influence on leads I, V5, and V6.

Figure 3B(b) shows the amplitude of R and S waves increase in lead II and V1–V3 when we rotate more counterclockwise along the left-to-right ventricular axis. More clockwise rotation affects the morphology of S wave in leads V2 and V3. Five leads I, II, V2, V3, and V5 in our work share the same amplitude features with results of Mincholé et al. (15). Our results also reflect the influence on the morphology, but the degree of change is not as prominent as Figure 3B(a).

In Figure 4A(b), when the heart moves more to the left-hand direction, the R wave and T wave amplitudes increase in leads I, II and V4-V6, while the S wave amplitude increases in all precordial leads. After comparison, we find only leads V5 and V6 agree with the findings of Mincholé et al. (15) [Figure 4A(a)], while other leads reflect the opposite influence.

Finally, we analyze the translation along the cranio-caudal direction. Figure 4B(b) shows that translation along this direction mainly affects the amplitude of T wave of leads V2-V4. We also notice an increase after translating the heart more to the inferior direction in leads V4-V6. Compared to the work of Mincholé et al. (15) [Figure 4B(a)], only lead II completely agrees. While our V4 and V5 have similar response to this translation, the degree of change in work of Mincholé et al. (15) are much greater.



A Comparison between changes when the heart geometry rotates around long axis(LA)

a) Results from Mincole (Mincole et al., 2019)

b) Results from our cVAE model with heart position and orientation

B Comparison between changes when the heart geometry rotates around left-to-right axis(LR)

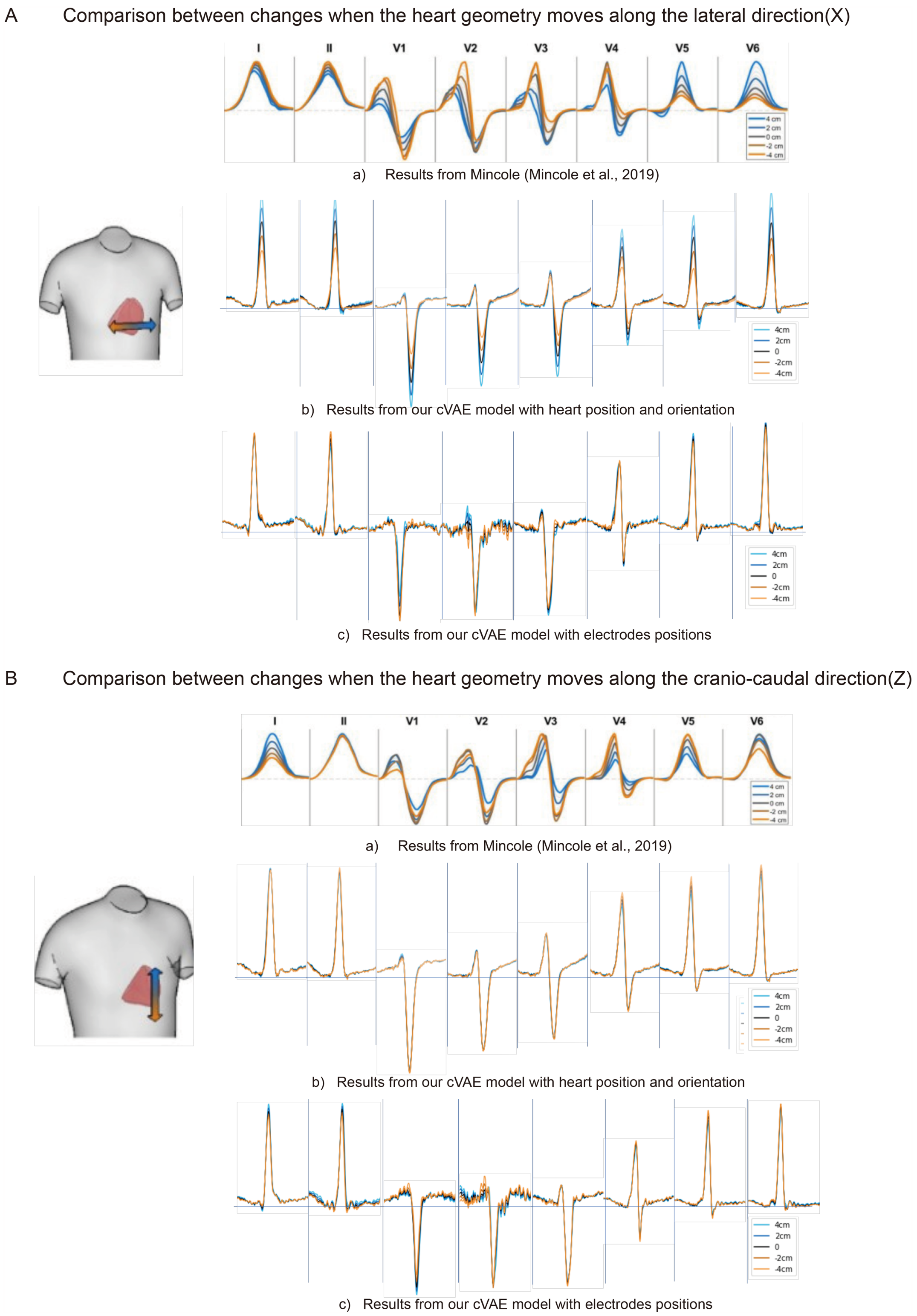
a) Results from Mincole (Mincole et al., 2019)

b) Results from our cVAE model with heart position and orientation

FIGURE 3
The comparison of ECG changes between previous work (15) and proposed cVAE model in leads I, II, and V1 to V6 when heart rotates around long axis **(A)** and around left-to-right axis **(B)** in 40, 20, 0, −20, 40° separately.

3.2 ECG conditioned by electrode positions

While the previous results demonstrate that our network successfully extracts valuable and relevant features from the ECGs, incorporating only heart position and orientation may present certain limitations. For instance, only considering absolute heart coordinates without accounting for their relative positions in the torso structure may reduce comparability across subjects, as the anatomical coordinate origin is determined by the scanner. This highlights the potential benefits of incorporating additional factors, such as torso structure, to enhance the accuracy and generalizability of our approach. We include the ten electrode positions to fix our torso structure when we perform the heart position translation. Each electrode


A Comparison between changes when the heart geometry moves along the lateral direction(X)

a) Results from Mincole (Mincole et al., 2019)

b) Results from our cVAE model with heart position and orientation

c) Results from our cVAE model with electrodes positions

B Comparison between changes when the heart geometry moves along the cranio-caudal direction(Z)

a) Results from Mincole (Mincole et al., 2019)

b) Results from our cVAE model with heart position and orientation

c) Results from our cVAE model with electrodes positions

FIGURE 4
The comparison of ECG changes between previous work (15) and proposed cVAE model in leads I, II, and V1 to V6 when heart moves along the lateral direction (A) and along the cranio-caudal direction (B) in 4, 2, 0, −2, −4 cm separately.

position is transformed from anatomical coordinate system to heart coordinate system using heart position and orientation Euler angles. Therefore, when evaluating the influence of heart information, the electrodes' positions are the only condition inputs to the model, which contain both heart position, orientation, and torso information.

For the analysis Figures 4A(c),B(c), we use a subset of 1,834 real cases that include electrode position information. By fixing the electrode positions to control for torso influence and modifying heart position inputs, we generate ECGs that are compared with the morphological trends observed in Mincholé's work. From Figure 4A, in general, the generated ECGs using ten electrodes have the same quality as the results using only heart positions, except with more noise in the generated leads V1 and V2 signals. This increased noise may result from the mismatch between the 3D nature of electrode positions and the 1D latent space used in our model, which introduces additional complexity in the decoding process.

The overall impact of heart information on the generated signals are more obvious than the one using electrodes, with more clear difference when we move the heart. However, if we treat the simulated signals in Figure 4A(a) as the standard, we can discover more accurate features or trends presented in the electrodes based model. When we look at leads I and II, Figure 4A(b,c) reacts to the position change in a completely opposite way. While R peak amplitude increases with the right to left movement of the heart in Figure 4A(b), it decreases in Figure 4A(c). When it comes to precordial leads, in leads V1 and V4 our electrodes' results of Figure 4A(c) also have more consistency with the simulated results than ones with heart-position only [Figure 4A(b)]. When the heart moves more to the left, the S wave peak of lead V1 increases, while in lead V4, the R wave peak increases and the S wave peak decreases. Those characteristics are exhibited in the opposite direction in Figure 4A(b).

In Figure 4B, more noise can be found in leads V1 and V2 in the model with electrodes' positions. Compared to Figure 4B(b), the influence of Z direction change is revealed more clearly using electrodes. Especially in leads I and II of the model with electrodes' positions [Figure 4B(c)], when the heart moves towards the head direction, the R wave amplitude will get increased, which is also reflected by the simulated results in Figure 4B(a). As comparison, the heart movement in Z direction has little influence on the final generated signals in our model with heart information only [Figure 4B(b)]. Regarding the precordial leads V1–V6, our two networks in Figures 4B(b),B(c) do not reveal large differences about the reaction to the heart position change. In leads V2–V6, the R and S wave amplitudes get larger if we move the heart more towards the feet. In general, the features in both Figures 4B(b),B(c) demonstrate more consistent results with simulated results in Figure 4B(a), except lead V6 which shows the opposite.

3.3 CVD risk prediction

We plot the Kaplan–Meier estimate curve of the full dataset before any stratification, as shown in Figure 5A. During seven years of follow-up observation, 5.5% of our total subjects have been diagnosed with CVD.

We train our network and achieve the score for each subject's future CVD risk in our test set, and accordingly divide them into two groups: low CVD risk and high CVD risk, using the median risk as threshold. Figure 5B reveals the Kaplan–Meier estimate of both groups of our test set. In Figure 5B, we can notice a clear difference between low CVD risk and high CVD risk groups. The CVD event occurs in 2% of subjects in the low risk group, and 6% of subjects in the high risk group over a nearly 7-years observation period.

Instead of considering the absolute survival times for each occurrence, survival analysis frequently uses the relative risk of an event (30, 31). To evaluate this, we use the concordance index (C-index), a widely used metric in survival analysis. Unlike classification metrics such as AUC-ROC, sensitivity, and

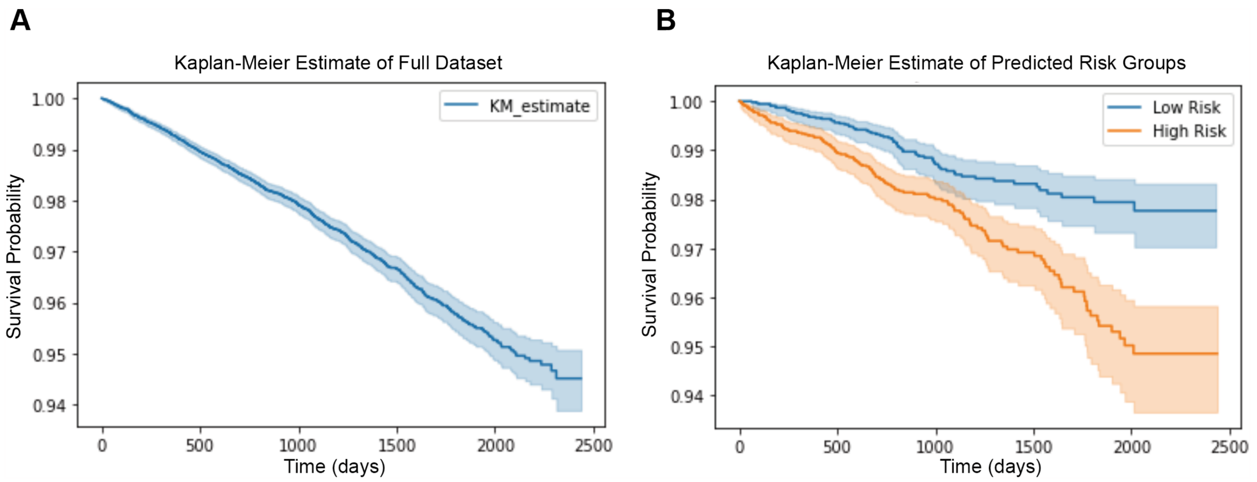


FIGURE 5
(A) Kaplan–Meier plot of the full dataset before stratification, showing survival probabilities for all subjects. (B) Kaplan–Meier plot of the test set, divided into low-risk (blue line) and high-risk (orange line) groups based on risk scores predicted by our model. The shaded areas represent the 95% confidence intervals (CIs).

specificity, which require binary labels, the C-index assesses how well the predicted risk scores preserve the correct ranking of event times. This makes it particularly suitable for our task, where the goal is quantifying relative CVD risk rather than classifying individuals into discrete risk categories.

Our model achieves a C-index of 0.63, indicating that ECG-derived risk scores successfully rank individuals based on their future CVD risk with performance significantly above random chance (C-index = 0.5). While existing CVD risk models often achieve higher C-index values by incorporating comprehensive clinical and lifestyle factors (e.g., blood pressure, cholesterol, and smoking history), our study focuses specifically on evaluating the prognostic value of ECG morphology alone.

We also explore whether the additional information can improve the performance of the network. Therefore, we first introduce the sex and age to the encoder and next the electrodes' positions. The idea was that sex and age are directly predictive of incident CVD, while the electrodes' positions could be used by the network to contextualize the shape of the ECG and refine the prediction. For the prediction model including sex and age, we use the same training set and test set as in the previous sections. The first row of Table 1 shows the result of our baseline model with C-index of 0.63. The inclusion of heart position information resulted in an increase of 3% in the concordance index, indicating an improvement in the model's ability to correctly rank individuals by CVD risk.

For the model including electrodes' positions, due to the limited size of our processed dataset as discussed in Section 2.4, we include 1,600 healthy cases and 100 cases with CVD diagnosed and maintain the same group proportion as in the previous experiment. From the results presented in Table 2, we find that the baseline model only achieves the C-index of 0.58. The addition of sex and age information increases the baseline model result to 0.61, with a 5% improvement. By incorporating electrodes' positions relative to heart coordinate system, the revised model provides a 1.7% increase in C-index, from 0.58 to 0.59. However, including the electrodes' position along with sex and age information do not further improve the predictive performance.

TABLE 1 C-index result for ECG baseline prediction model and model with additional demographic information.

Model	C-index
Baseline	0.63
Baseline + sex + age	0.65

TABLE 2 C-index result for ECG baseline model, model with demographic information, and electrodes' positions.

Model	C-index
Baseline	0.58
Baseline + electrode positions	0.59
Baseline + sex + age	0.61
Baseline + sex + age + electrode positions	0.61

4 Discussions

In this work, we have developed a conditional VAE model to generate 12-lead ECGs, which takes heart position, orientation, and electrodes' positions as conditions. The results of our cVAE model show that the heart position and orientation have a significant impact on the generated ECGs, which is consistent with previous research (15). However, the influence of heart position and orientation on the generated ECGs is not as prominent as the simulated results. One possible explanation is that our position definition is not accurate enough because we only calculate the intersection of three cardiac MRI planes. An alternative explanation is in the work of Mincholé et al. (15) the torso structure was fixed for simulation, while in our research the torso of each subject can vary. Additionally, after comparison we find that some rotation degree and translation distance in the work of Mincholé et al. (15) are too large to the extent that they do not occur in real subjects.

When we include electrodes' positions as input, they should also contain heart position and orientation information. During training, electrode positions help capture the influence of torso anatomy on ECG morphology. During generation, fixing electrodes' positions allows us to control for torso-related variability, ensuring that observed ECG changes are primarily driven by modifications in heart position and orientation. Therefore, in this experiment we are able to reduce the influence of torso on our final generated results. From Figure 4, we can notice with the addition of electrodes' positions, the consistency between our generated signals and simulated *in silico* signals of Mincholé et al. (15) gets improved. This illustrates that our model including the electrodes' positions is capable of capturing useful features that represent the individual characteristics well, though there is more noise in the final generated signals. A potential explanation for this issue lies in the difference between the 3D nature of the electrodes' positions (3 × 10 coordinates) and the 1D latent space (64 dimensions) used in our experiments. This mismatch introduces additional complexity, which may challenge the decoder's ability to effectively interpret and reconstruct the information. To address this, further parameter tuning or introduction of a separate encoder for electrodes' positions could help achieve better results.

While comparing our generated outcomes with the work of Mincholé et al. (15), it is important to acknowledge that the comparison is largely qualitative in nature, given that their work does not provide actual values to enable a more comprehensive, quantitative comparison. Thus, although this comparison provides some initial insights into the relative performance of our model, further quantitative analysis would be required to provide a more definitive evaluation of the model's performance. Additionally, the work of Mincholé et al. (15) mainly focused on the QRS complex of the ECG, while the other crucial components of the ECG waveform, such as the P and T waves, have not been examined. To address this limitation, our future work will focus on integrating detailed biophysical parameters into our generative model, enabling a more precise quantitative comparison between our synthesized ECGs and simulation-based results.

About the CVD risk prediction model, the results from Figure 5 suggest that our cVAE with predictor successfully learned to stratify subjects by CVD risk using features extracted from 12-lead ECG signals. This indicates that there is useful information related to their future CVD risk contained in ECG recordings, and our model has the ability to capture it. The baseline model in the current study attained a C-index of 0.63, suggesting a moderate predictive performance that necessitates further refinement. Although the C-index provides a useful quantification of model performance, its standalone value might not fully encapsulate the model's clinical applicability. The future works could further explore the ECG of the subjects defined as high risk group by our network, and analyze their ECG measurements in detail in order to find common characteristics for certain diseases.

When we include additional demographic information to our prediction network, as shown in Tables 1 and 2, it improves the C-index by 3% and 5%. This is consistent with previous findings of Alaa et al. (20), which highlighted the importance of age and sex in CVD risk evaluation. While this suggests that sex and age contribute to risk prediction, the relatively modest increase reflects the fact that ECG waveforms already encode physiological characteristics associated with these demographic factors. Our future work will explore the inclusion of additional subject information commonly used in traditional risk evaluation methods, such as the Framingham Risk Score factors (e.g., smoking history, blood pressure), to assess whether incorporating a broader range of clinical variables could further enhance model performance.

In Table 2, we notice that the addition of electrodes' positions does not improve the C-index. One possible explanation is that the relationship between electrodes' positions and CVD risk is already partially captured within the ECG waveforms themselves. Since ECG morphology inherently encodes subject-specific anatomical and physiological characteristics, some of the variability introduced by differences in electrode positioning may have already been learned by the model. Due to the high dimensionality of the electrode position data (3×10 coordinates), the single fully connected layer in our current model may not be expressive enough to fully map these features into the latent space for risk prediction. A more complex network architecture could be explored in future to better leverage electrode position information for improved prediction performance.

While our study investigates general CVD risk prediction, further work is needed to explore how changes in ECG amplitude and duration, resulting from variations in heart position and electrode placement, impact the prediction of specific cardiovascular diseases. Certain ECG-derived biomarkers, such as ST-segment deviations or QRS complex amplitudes, are directly influenced by these factors and play a crucial role in diagnosing conditions such as myocardial infarction or hypertrophy. A future extension of our work could involve evaluating how disease-specific classification models respond to these anatomical influences, improving the interpretability and robustness of ECG-based prediction methods.

5 Conclusion

In this work, we have developed a cVAE-based ECG generation model, incorporating the electrodes' positions to include torso information. This approach has markedly improved the consistency between our generated signals and previous *in silico* studies, surpassing the performance of models that relied solely on heart position and orientation. Through the meaningful latent space representation learned by our cVAE model, we highlight the ability of ECG signals alone to predict future CVD risk. Furthermore, by incorporating additional conditioning factors such as age, sex, and electrodes' positions, we demonstrate that these structured inputs provide additional guidance, further refining risk estimation. Our findings underscore the potential of generative approaches to extract clinically relevant features from 12-lead ECG signals, supporting the development of more personalized and data-driven CVD risk assessment models.

Data availability statement

The data supporting the conclusions of this article will be made available by the authors upon reasonable request.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

YS: Data curation, Investigation, Software, Writing – review & editing, Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft; AB: Data curation, Investigation, Software, Writing – review & editing, Supervision, Funding acquisition; MB: Conceptualization, Data curation, Investigation, Methodology, Supervision, Validation, Visualization, Writing – review & editing, Software; VG: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing, Data curation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. AB is supported by the Royal Society University Research Fellowship (Grant No. URF\R1\221314).

Acknowledgments

This work used the facilities of High Performance Computing (HPC) of the Institute of Biomedical Engineering (IBME), University of Oxford and the HPC of the Advanced Research Computing (ARC) service of the University of Oxford. This research has been conducted using the UK Biobank Resource under Application Number "40161".

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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. Waller AD. A demonstration on man of electromotive changes accompanying the heart's beat. *J Physiol (Lond).* (1887) 8:229–34. doi: 10.1113/jphysiol.1887.sp000257

2. George S, Rodriguez I, Ipe D, Sager PT, Gussak I, Vajdic B. Computerized extraction of electrocardiograms from continuous 12-lead holter recordings reduces measurement variability in a thorough QT study. *J Clin Pharmacol.* (2012) 52:1891–900. doi: 10.1177/0091270011430505

3. Beetz M, Banerjee A, Grau V. Multi-domain variational autoencoders for combined modeling of MRI-based biventricular anatomy and ECG-based cardiac electrophysiology. *Front Physiol.* (2022) 13:886723. doi: 10.3389/fphys.2022.886723

4. Beetz M, Banerjee A, Sang Y, Grau V. Combined generation of electrocardiogram and cardiac anatomy models using multi-modal variational autoencoders. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022). p. 1–4.

5. Benali R, Bereksi Reguig F, Hadj Slimane Z. Automatic classification of heartbeats using wavelet neural network. *J Med Syst.* (2012) 36:883–92. doi: 10.1007/s10916-010-9551-7

6. Kampouraki A, Manis G, Nikou C. Heartbeat time series classification with support vector machines. *IEEE Trans Inf Technol Biomed.* (2008) 13:512–8. doi: 10.1109/TITB.2008.2003323

7. Li L, Camps J, Banerjee A, Beetz M, Rodriguez B, Grau V. Deep computational model for the inference of ventricular activation properties. In: *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers.* Springer (2023). p. 369–80.

8. Li L, Camps J, Jenny Wang Z, Beetz M, Banerjee A, Rodriguez B, et al. Toward enabling cardiac digital twins of myocardial infarction using deep computational models for inverse inference. *IEEE Trans Med Imaging.* (2024) 43:2466–78. doi: 10.1109/TMI.2024.3367409

9. Zhang L, Peng H, Yu C. An approach for ECG classification based on wavelet feature extraction and decision tree. In: *2010 International Conference on Wireless Communications & Signal Processing (WCSP).* IEEE (2010). p. 1–4.

10. Chen C, Li L, Beetz M, Banerjee A, Gupta R, Grau V. Large language model-informed ECG dual attention network for heart failure risk prediction. *arXiv* [Preprint]. *arXiv:2403.10581* (2024).

11. McLachlan S, Dube K, Gallagher T. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI).* IEEE (2016). p. 439–48.

12. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv* [Preprint]. *arXiv:1706.02633* (2017).

13. Kuznetsov V, Moskalenko V, Zolotykh NY. Electrocardiogram generation and feature extraction using a variational autoencoder. *arXiv* [Preprint]. *arXiv:2002.00254* (2020).

14. Zhu F, Ye F, Fu Y, Liu Q, Shen B. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. *Sci Rep.* (2019) 9:6734. doi: 10.1038/s41598-019-42516-z

15. Mincholé A, Zacur E, Ariga R, Grau V, Rodriguez B. MRI-based computational torso/biventricular multiscale models to investigate the impact of anatomical variability on the ECG QRS complex. *Front Physiol.* (2019) 10:458916. doi: 10.3389/fphys.2019.01103

16. Jæger KH, Tveito A. Deriving the bidomain model of cardiac electrophysiology from a cell-based model; properties and comparisons. *Front Physiol.* (2022) 12:811029. doi: 10.3389/fphys.2021.811029

17. Sang Y, Beetz M, Grau V. Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI).* IEEE (2022). p. 1–5.

18. Smith HJ, Banerjee A, Choudhury RP, Grau V. Automated torso contour extraction from clinical cardiac MR slices for 3D torso reconstruction. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* IEEE (2022). p. 3809–13.

19. Smith HJ, Rodriguez B, Sang Y, Beetz M, Choudhury R, Grau V, et al. Anatomical basis of sex differences in human post-myocardial infarction ECG phenotypes identified by novel automated torso-cardiac 3D reconstruction. *arXiv* [Preprint]. *arXiv:2312.13976* (2023).

20. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS ONE.* (2019) 14: e0213653. doi: 10.1371/journal.pone.0213653

21. Emdin CA, Wong CX, Hsiao AJ, Altman DG, Peters SA, Woodward M, et al. Atrial fibrillation as risk factor for cardiovascular disease and death in women compared with men: systematic review and meta-analysis of cohort studies. *BMJ.* (2016) 352:h7013. doi: 10.1136/bmj.h7013

22. Turnbull I, Camm C, Halsey J, Du H, Chen Z, Clarke R. Population prevalence of ECG abnormalities and risk of incident CVD outcomes: 5-year follow-up of 25,000 Chinese adults. *Eur Heart J.* (2023) 44:ehad655–2380. doi: 10.1093/eurheartj/ehad655.2380

23. Wu G, Wu J, Lu Q, Cheng Y, Mei W. Association between cardiovascular risk factors and atrial fibrillation. *Front Cardiovasc Med.* (2023) 10:1110424. doi: 10.3389/fcvm.2023.1110424

24. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* (2015) 12:e1001779. doi: 10.1371/journal.pmed.1001779

25. Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A wide and deep transformer neural network for 12-lead ECG classification. In: *2020 Computing in Cardiology.* IEEE (2020). p. 1–4.

26. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson.* (2015) 18:1–7. doi: 10.1186/s12968-016-0227-4

27. Strain T, Wijndaele K, Sharp SJ, Dempsey PC, Wareham N, Brage S. Impact of follow-up time and analytical approaches to account for reverse causality on the association between physical activity and health outcomes in UK Biobank. *Int J Epidemiol.* (2020) 49:162–72. doi: 10.1093/ije/dyz212

28. Banerjee A, Camps J, Zacur E, Andrews CM, Rudy Y, Choudhury RP, et al. A completely automated pipeline for 3D reconstruction of human heart from 2D cine magnetic resonance slices. *Philos Trans R Soc A.* (2021) 379:20200257. doi: 10.1098/rsta.2020.0257

29. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* [Preprint]. *arXiv:1412.6980* (2014).

30. Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med.* (1984) 3:143–52. doi: 10.1002/sim.4780030207

31. Pencina MJ, D'agostino RB. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med.* (2004) 23:2109–23. doi: 10.1002/sim.1802

TYPE Original Research
PUBLISHED 12 May 2025
DOI 10.3389/fdgth.2025.1583399

Interpretable AI-driven multi-objective risk prediction in heart failure patients with thyroid dysfunction

Massimo Iacoviello[1], Vito Santamato[2], Alessandro Pagano[3] and Agostino Marengo[2]*

[1]Department of Clinical and Experimental Medicine, University of Foggia, Foggia, Italy, [2]Department of Agriculture, Food, Natural Resources and Engineering Sciences, University of Foggia, Foggia, Italy, [3]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

OPEN ACCESS

EDITED BY
Ulrich Parlitz,
Max Planck Institute for Dynamics and Self-Organization, Germany

REVIEWED BY
Paraskevi Papadopoulou,
American College of Greece, Greece
Ge Zhang,
The First Affiliated Hospital of Zhengzhou University, China

*CORRESPONDENCE
Agostino Marengo
✉ agostino.marengo@unifg.it

RECEIVED 26 February 2025
ACCEPTED 29 April 2025
PUBLISHED 12 May 2025

CITATION
Iacoviello M, Santamato V, Pagano A and Marengo A (2025) Interpretable AI-driven multi-objective risk prediction in heart failure patients with thyroid dysfunction.
Front. Digit. Health 7:1583399.
doi: 10.3389/fdgth.2025.1583399

Introduction: Heart Failure (HF) complicated by thyroid dysfunction presents a complex clinical challenge, demanding more advanced risk stratification tools. In this study, we propose an AI-driven machine learning (ML) approach to predict mortality and hospitalization risk in HF patients with coexisting thyroid disorders.

Methods: Using a retrospective cohort of 762 HF patients (including euthyroid, hypothyroid, hyperthyroid, and low T3 syndrome cases), we developed and optimized several ML models—including Random Forest, Gradient Boosting, Support Vector Machines, and others—to identify high-risk individuals.

Results: The best-performing model, a Random Forest classifier, achieved robust predictive accuracy for both 1-year mortality and HF-related hospitalization (area under the ROC curve ~0.80 for each). We further employed model interpretability techniques (Local Interpretable Model-agnostic Explanations, LIME) to elucidate key predictors of risk at the individual level. This interpretability revealed that factors such as atrial fibrillation, absence of cardiac resynchronization therapy, amiodarone use, and abnormal thyroid-stimulating hormone (TSH) levels strongly influenced model predictions, providing clinicians with transparent insights into each prediction. Additionally, a multi-objective risk stratification analysis across thyroid status subgroups highlighted that patients with hypothyroidism and low T3 syndrome are particularly vulnerable under high-risk conditions, indicating a need for closer monitoring and tailored interventions in these groups.

Discussion: In summary, our study demonstrates an innovative AI methodology for medical risk prediction: interpretable ML models can accurately stratify mortality and hospitalization risk in HF patients with thyroid dysfunction, offering a novel tool for personalized medicine. These findings suggest that integrating explainable AI into clinical workflows can improve prognostic precision and inform targeted management, though prospective validation is warranted to confirm realworld applicability.

KEYWORDS

artificial intelligence, machine learning, heart failure, thyroid dysfunction, risk stratification, explainable AI

1 Introduction

Heart Failure (HF) is one of the leading causes of morbidity and mortality globally, imposing a significant burden on healthcare systems and the quality of life of patients. Concurrently, thyroid dysfunctions, particularly hypothyroidism, have been associated with worsening clinical outcomes in patients with HF, adversely affecting prognosis. Recent studies underscore that subclinical hypothyroidism (SH) significantly raises the risk of cardiovascular mortality in HF patients, emphasizing the need for precise monitoring and intervention strategies (1). Optimal ranges of thyroid-stimulating hormone (TSH) and free thyroxine (FT4) levels are linked to reduced mortality risks, suggesting that both high and low extremes can worsen HF outcomes (2). Previous studies have demonstrated that hypothyroidism can negatively impact cardiac function and increase the risk of developing HF. Recent meta-analyses have confirmed that subclinical hypothyroidism is associated with an increased risk of all-cause mortality and hospitalization in patients with HF, highlighting the importance of thyroid evaluation in this population (3). However, the relationship between hypothyroidism, HF, and mortality remains complex and multifactorial, requiring further exploration for optimal patient management.

The complexity of clinical management of this patient cohort underscores the need for advanced tools for accurate and personalized risk assessment. Machine learning (ML) has shown revolutionary capabilities in the medical field, particularly in predictive medicine, where complex models such as XGBoost, Random Forest, and LightGBM have managed large volumes of clinical data and identified complex patterns not immediately apparent to human analysis (4). Recent advancements, such as the use of SF-IIAdaboost algorithms integrating IoT and AI, have achieved high predictive accuracy in cardiovascular contexts, underscoring the potential for enhanced prognostic precision (5). The use of advanced ML algorithms has enabled the identification of clinical and biochemical features that predict mortality risk, examining how these interact with each other and with the patient's baseline condition. Such models have been shown to improve risk stratification and treatment personalization in patients with HF, including those in a hypothyroid state (6). In patients with HF, ML analysis has identified prognostic phenotypes, facilitating the application of precision medicine. This approach is particularly relevant for hypothyroid patients, who present a unique disease dynamic compared to patients with overt thyroid dysfunction (7).

This work aims to explore the application of ML in estimating the mortality risk in hypothyroid patients suffering from HF, with a particular emphasis on the analysis of age and TSH levels as prognostic factors. Through the analysis of a large cohort of cardiac patients stratified by thyroid conditions, this study aims to develop ML models that provide accurate estimates for two main targets: mortality and hospitalization in this specific population. Our goal is twofold: on one hand, to contribute to the scientific literature by offering insights into the underlying mechanisms of the association between thyroid conditions and HF; on the other hand, to provide healthcare providers with an innovative tool for improving risk stratification and personalizing therapeutic strategies.

The core of this work involves the presentation of the research methods used to develop the ML models, including feature selection, model training, and validation. Finally, the results are analyzed in detail, highlighting how various factors contribute to predicting the risk of mortality and hospitalization in patients with HF and how these models can be employed in clinical practice to support more informed therapeutic decisions.

The use of ML in predicting mortality risk in patients with HF could mark a significant advancement in managing this complex intersection of conditions. This study aims to explore such potential, opening new frontiers in cardiovascular and endocrinological research. By highlighting these computational underpinnings, the manuscript extends the theoretical understanding of explainable AI in clinical contexts and bridges the gap between algorithmic transparency and medical applicability. The article begins in Section 2 with a comprehensive background, offering an overview of related studies and showcasing the unique benefits and objectives of this research. In Section 3, the methodology is detailed, guiding readers through the study's innovative approach. Section 4 dives into a discussion of the primary findings, spotlighting key results and their implications. Finally, the conclusion ties everything together, underscoring the study's contributions and future directions.

2 Background

The growing awareness of the negative impact of hypothyroidism on patients with HF underscores the need for comprehensive risk assessment and personalized management strategies. Studies have shown that hypothyroidism, including its subclinical form, is prevalent among HF patients and significantly contributes to an increased risk of mortality, hospitalization, and deterioration of cardiac function. Amiodarone, a commonly used antiarrhythmic drug, has been identified as a determining factor in the onset of hypothyroidism in this population (8). Research highlights the importance of monitoring TSH levels as a key indicator of thyroid function in these patients. It has been demonstrated that correcting thyroid hormone deficiency, indicated by elevated TSH levels, leads to improvements in cardiac function while simultaneously reducing the risk of hospitalization and mortality. Conversely, worsening thyroid function, characterized by rising TSH levels, is associated with a decline in cardiac function and adverse outcomes (9, 10). Beyond traditional risk markers, the role of N-terminal pro-B-type natriuretic peptide (NT-proBNP) has emerged as a significant prognostic factor in patients with suspected HF. Even in the absence of echocardiographic evidence of HF, elevated NT-proBNP levels, combined with factors such as advanced age, male sex, chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), and dementia, have been associated with higher mortality (11). These findings highlight the complex interaction between HF and thyroid dysfunction, suggesting a need for more sophisticated approaches for accurate risk stratification and timely interventions.

The emergence of Machine Learning (ML) algorithms such as XGBoost, Random Forest, and LightGBM offers a promising avenue forward. These ML algorithms have demonstrated their ability to discern complex prognostic patterns and improve treatment personalization in various healthcare contexts, including predicting acute kidney injury (AKI) following percutaneous coronary intervention (PCI) in patients with acute coronary syndrome (ACS) (12). In HF, recent studies indicate that ML models enhance predictive accuracy for mortality and readmission by integrating comprehensive clinical data and managing issues like data imbalance and incompleteness (13). Advanced deep learning techniques, such as multi-head self-attention, further improve model performance, particularly in handling complex and diverse datasets common in HF populations (14). Applying ML algorithms in this context may improve the precision of risk assessment and support more personalized management of patients with HF and hypothyroidism, although prospective validation is still required. By harnessing the power of these algorithms, we could develop predictive models capable of accurately identifying high-risk individuals for adverse outcomes, allowing for targeted interventions and improved patient outcomes. Additionally, the integration of variables such as age and TSH levels into ML models could provide further insights into the delicate balance between cardiac and thyroid function. By incorporating these factors, the resulting models may achieve higher predictive accuracy, guiding clinical decisions and leading to personalized treatment strategies.

2.1 Related studies and benefits

Recent scientific literature highlights the effectiveness of ML in predicting complex clinical outcomes, such as mortality and hospitalization, especially in patients with endocrine and cardiovascular comorbidities. Some studies have explored the use of ML to analyze autoimmune and endocrine diseases, revealing the significant role that conditions like diabetes and thyroid disorders play in elevating mortality rates (15). Similarly, other studies have applied ML to diagnose forms of secondary hypertension, showing how abnormal TSH levels can influence cardiovascular risk (16). Additionally, models have emerged linking diabetes and hypothyroidism with increased mortality in COVID-19 patients requiring hospitalization (17), while other research has developed algorithms to predict atrial fibrillation associated with thyrotoxicosis, emphasizing the importance of thyroid profiles in heart disease (18). Further investigations into the connection between subclinical hypothyroidism and cardiovascular diseases have also examined the potential for accurately predicting mortality and hospitalization in patients with HF (19, 20). ML models that incorporate social determinants of health have also shown promise in predicting in-hospital mortality for HF patients, illustrating the benefits of integrating clinical and social factors to improve outcomes in complex cardiovascular cases (21). Efforts to enhance cardiovascular risk predictions by integrating factors such as diabetes and thyroid health have further refined risk stratification models (22). Additionally, there is promising research on ML frameworks that predict postprocedural outcomes in interventional radiology using random forest models, offering insight into complications, mortality, and length of stay (23). However, these studies often treat thyroid dysfunctions as one of many risk variables, without fully exploring their specific impact on patients with cardiovascular conditions.

This study stands out by providing a detailed, targeted analysis of the influence of thyroid conditions on clinical outcomes through an innovative ML approach. Unlike previous studies, this work focuses specifically on the impact of thyroid dysfunctions, making each prediction more precise and tailored to clinical management. Additionally, by using Local Interpretable Model-agnostic Explanations (LIME), predictions are both transparent and individualized, allowing clinicians to clearly see how each clinical variable contributes to the risk of mortality or hospitalization for each patient, thereby supporting more informed and personalized decision-making.

The ML analysis also extends to specific patient subgroups, such as euthyroid and hypothyroid patients, making this study uniquely comprehensive compared to existing literature. Through advanced predictive modeling, the study has identified the absence of Cardiac Resynchronization Therapy (CRT) as a critical risk factor for mortality in patients with thyroid dysfunctions, suggesting that targeted interventions could improve patient prognosis. Another key finding is the association between low TSH levels and reduced hospitalization risk in euthyroid patients, introducing new parameters to monitor even in the absence of overt hypothyroidism or hyperthyroidism. Finally, ML has enabled the identification of an increased mortality risk associated with Amiodarone use in patients with LT3, offering practical insights for optimizing therapeutic decisions in cardiology.

In summary, this study not only enriches scientific knowledge but also serves as an innovative pillar for precision medicine in managing patients with thyroid and cardiovascular comorbidities. The advanced use of ML enables more accurate and personalized predictions, thus transforming the quality of clinical care.

2.2 Patient selection

In this study, we examined a cohort of 762 patients to assess significant clinical outcomes such as HF hospitalization and mortality over the follow-up period. The patients were monitored for durations ranging from less than a month to almost 12.7 years, with an average follow-up period of approximately 4.5 years (9).

The selection of participants was meticulously conducted to include only those subjects for whom complete data were available regarding arrival date, follow-up date, age, sex, and key clinical events such as mortality and HF hospitalization. No patient was excluded due to a lack of essential data, thus maintaining the integrity of the cohort.

From a demographic perspective, the average age of participants at the time of arrival was 63.5 years, ranging from 14 to 89 years. Males constituted 78% of the cohort, demonstrating a prevalence of this gender. This sex imbalance reflects the characteristics of the referred population but may also introduce gender-related bias, particularly relevant given the higher prevalence of thyroid dysfunction in females. Regarding clinical outcomes, about 30% of the patients died, and 22% experienced at least one episode of HF hospitalization during the follow-up period. All consecutive outpatients with CHF referred to the HF Unit of the University Policlinic Hospital of Bari from January 2006 to December 2016 were retrospectively evaluated. All the evaluations with patients in stable clinical conditions from at least 30 days and in conventional medical and electrical therapy from at least 3 months were considered. The adoption of well-defined inclusion criteria minimized potential biases arising from incomplete data and enhanced the representativeness and generalizability of the results. For patients who developed thyroid dysfunction after their initial evaluation, the clinical timepoint corresponding to the diagnosis of hypothyroidism, hyperthyroidism, or low-T3 syndrome was considered as the analytical baseline (9). This allowed for consistent classification of thyroid status and ensured that risk predictions were anchored to the relevant endocrine condition.

3 Materials and methods

The study is based on a dataset of 762 patients and employs ML techniques implemented in Python to build predictive models that estimate the risks of mortality and hospitalization. The main objective is to analyze the influence of various clinical characteristics, including thyroid variables, on these outcomes.

The analyses were conducted using Orange Data Mining software version 3.36.2 on an Apple M1 Pro system equipped with 16 GB of RAM and 1 TB of storage, operating on macOS Sonoma 14.2.1. This setup, combined with the use of advanced ML techniques, ensured the efficiency and reproducibility of our analyses. The importance of such ML methodologies in extracting meaningful insights and predictive models from complex datasets has been previously highlighted and validated in similar studies in the field of health performance assessment, such as efficiency and mobility (24–27) and for predicting neurodevelopmental disorders in children (28). The methodological phases of the study, illustrated in Figure 1, were developed in a Python environment, highlighting the key steps of the analysis.

The methodological workflow, illustrated in Figure 1, follows a multi-step approach organized into key phases:

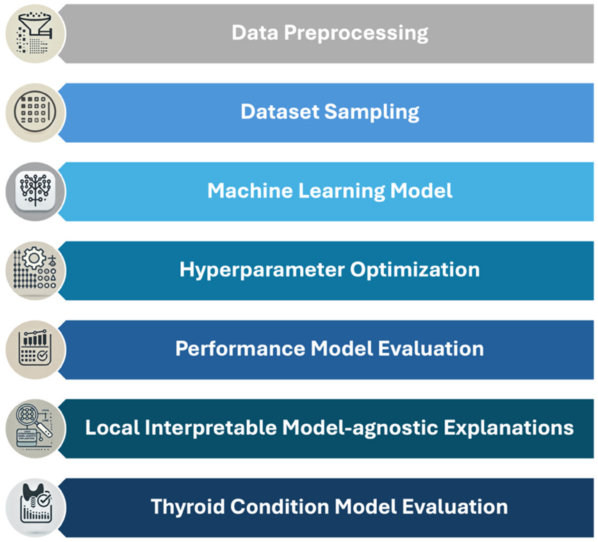

Data Preprocessing
Dataset Sampling
Machine Learning Model
Hyperparameter Optimization
Performance Model Evaluation
Local Interpretable Model-agnostic Explanations
Thyroid Condition Model Evaluation

FIGURE 1
Workflow diagram for data collection and ML model training.

1. Data preprocessing and handling of missing data: Missing data is managed through model-based methods that leverage relationships among variables to estimate missing values, preserving the original distribution and minimizing potential bias.
2. Dataset sampling: To assess model robustness, the dataset is split into a training set and a test set, allowing for rigorous validation of predictive performance.
3. Selection of ML models: Various ML algorithms are tested, including Random Forest, Gradient Boosting, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Neural Networks, Decision Trees, AdaBoost, Stochastic Gradient Descent, and Logistic Regression.
4. Internal Validation and Hyperparameter Optimization: Techniques such as grid search and cross-validation are employed to optimize hyperparameters, ensuring that model performance generalizes and is not limited to the training set alone.
5. Performance Model Evaluation: An evaluation function is created to automate model assessment on the test data, calculating metrics such as area under the ROC curve (AUC), accuracy, F1-score, precision, recall, and MCC to facilitate model comparison.
6. Model interpretation with LIME: To interpret predictions, LIME is used, highlighting the contribution of each variable to the final prediction and providing visual representations accessible to a non-technical audience.
7. Evaluation of models on different thyroid conditions: Models are evaluated on both the entire dataset and subgroups based on thyroid conditions (Euthyroidism, Hypothyroidism, Hyperthyroidism, and Low T3 Syndrome). This approach allows exploration of how model performance varies according to different thyroid conditions.

In summary, the study adopts a ML approach to develop and validate predictive models for mortality and hospitalization risks in cardiology and endocrinology patients. The workflow incorporates multiple phases, from data preprocessing to model interpretation, with particular attention to the influence of thyroid variables.

3.1 Dataset

Initially, we collected a broad set of clinical data, including both numerical and categorical variables ranging from demographic to biochemical parameters (29). In accordance with the study conducted by Iacoviello et al. in 2020, for each patient, the baseline evaluation was conducted during the first recorded medical visit. At this stage, a comprehensive medical history, physical examination, 12-lead ECG, mono- and two-dimensional echocardiographic evaluation, and blood samples were collected. For patients who subsequently developed thyroid disorders, the evaluation corresponding to the diagnosis of hypothyroidism, hyperthyroidism, or low T3 syndrome (LT3) was considered as the baseline. During the medical visit, the presence of ischemic cardiomyopathy, arterial hypertension, atrial fibrillation, and diabetes mellitus was carefully documented, along with any previous thyroid disease diagnosis. Data on HF therapy and any prior or ongoing treatment with amiodarone were also gathered. Additionally, information regarding the thyroid disease diagnosis was recorded. The 12-lead ECG was used to assess heart rhythm and rate. Echocardiographic recordings were obtained using a phased-array echo-Doppler system (Sonos 5500, Philips, Netherlands; from September 2008 onward, Vivid 7, GE, Wisconsin, USA) to estimate the left ventricular ejection fraction (LVEF) using the Simpson method. At baseline, levels of sodium (mEq/L), serum creatinine concentrations (mg/dl), and hemoglobin (g/dl) were measured. The glomerular filtration rate (GFR, ml/min) was calculated using the EPI formula (30). Additionally, amino-terminal brain natriuretic peptide (NT-proBNP, Dade Behring, Eschborn, Germany), free T3 (fT3), free thyroxine (fT4), and TSH levels were measured through immunoassays, using the reference ranges provided by the kit manufacturers (Advia Centaur, Bayer HealthCare, Diagnostics Division, Tarrytown, NY, US until 2011, and subsequently Dimension Vista, Siemens Healthcare Diagnostics, Erlangen, Germany). The resulting dataset with the selected variables is shown in Table 1.

The table provides a comprehensive description of the variables used to feed our ML model for predicting two key clinical outcomes: mortality and hospitalization. The variables are organized into two main categories, namely *Target*, which includes the clinical outcomes of interest, and *Feature*, which comprises the relevant clinical and demographic factors selected to optimize the predictive accuracy of the model.

In the *Target* category, there are two variables, "Mortality" and "Hospitalization," which respectively indicate the occurrence of patient mortality and hospitalization. Each is coded as a categorical variable, with the value 1 representing the occurrence of the event and the value 0 indicating its absence. These targets serve as the dependent variables of the model, which is trained to identify and classify the risks associated with each outcome.

The *Features* include a range of demographic and clinical variables, carefully selected to identify significant correlations and enhance the model's predictive capabilities. Among the demographic characteristics, *MALE GENDER* indicates the

TABLE 1 Overview of variables in the dataset.

Model variable	Variable name	Description	Type variable
Target	Mortality	Patient mortality event (1: Yes, 0: No)	Categorical
	HF hospitalization	Patient hospitalization (1: Yes, 0: No)	
FEATURE	Male gender	Patient's gender (1: male, 0: female)	
	Ischemic cardiomiopaty	Presence of ischemic cardiomyopathy (1: present, 0: absent)	
	Diabetes	Diabetes diagnosis (1: Diabetic, 0: non-diabetic)	
	ACEi/ARBs	Use of ACE inhibitors or ARBs (1: Use, 0: no use)	
	Beta-blockers	Use of beta-blockers (1: Use, 0: no use)	
	Diuretics	Use of diuretics (1: Use, 0: no use)	
	Aldosterone antagonists	Use of aldosterone antagonists (1: Use, 0: no use)	
	Amiodarone	Use of amiodarone (1: Use, 0: no use)	
	ICD	Implantable defibrillator (1: Present, 0: absent)	
	CRT	Cardiac resynchronization therapy (1: present, 0: absent)	
	NYHA class	NYHA functional class (1, 2, 3)	
	Atrial fibrillation	Presence of atrial fibrillation (1: present, 0: absent)	
	Age	Patient's age (years)	Numerical
	BMI	Body mass index (kg/m^2)	
	Systolic arterial pressure	Systolic blood pressure (mmHg)	
	LVEF	Calculated ejection fraction (percentage)	
	GFR-EPI	Estimated glomerular filtration rate (ml/min/1.73 m^2)	
	Natremia	Blood sodium concentration (mmol/L)	
	NT-proBNP	NT-proBNP levels in blood (pg/ml)	
	FT3	Free triiodothyronine levels (pmol/L)	
	FT4	Free thyroxine levels (pmol/L)	
	TSH	TSH levels (mU/L)	

patient's gender, with 1 for male and 0 for female, an important attribute as gender can influence HF prognosis. The patient's age is represented by the continuous numeric variable *AGE*, allowing the model to capture risk variations associated with advanced age. The body mass index *BMI*, expressed in kg/m², is also included as a general health indicator, potentially associated with overall cardiovascular risk.

The clinical variable set consists of critical diagnostic information, such as the presence of ischemic cardiomyopathy, described by the variable *ISCHEMIC CARDIOMYOPATHY*, and diabetes diagnosis, represented by the *DIABETES* variable. Both are binary variables distinguishing between patients with and without these conditions, each known to negatively impact the progression of HF. Other clinical variables include pharmacological treatments followed by the patients, such as the use of ACE inhibitors or angiotensin receptor blockers ACEinhibitor/ANGIOTENSIN II RECEPTOR BLOCKERS (*ACEi/ARBs*), *BETA-BLOCKERS*, *DIURETICS*, and *MINERALCORTICOID RECEPTOR ANTAGONISTS*. These medications, coded as 1 for use and 0 for non-use, play a crucial role in managing symptoms and preventing cardiovascular complications. The use of *AMIODARONE*, an antiarrhythmic drug, is similarly included as a binary variable, as it is relevant for patients with severe arrhythmias. *ATRIAL FIBRILLATION* is a key clinical feature indicating the presence of atrial fibrillation, coded as 1 for present and 0 for absent. This variable is essential for HF patients, as atrial fibrillation can exacerbate symptoms and increase the risk of adverse events.

The model also incorporates instrumental characteristics, such as the presence of an implantable cardioverter-defibrillator *ICD* and cardiac resynchronization therapy *CRT*, both coded to indicate the presence or absence of the device, respectively with 1 and 0. The patient's *NYHA* functional class, categorized with values from 1 to 3, is another critical clinical parameter, as it reflects the severity of HF symptoms and helps predict the risk of adverse events.

The dataset further includes a series of relevant physiological and biochemical parameters, such as systolic blood pressure, measured in mmHg, and the calculated ejection fraction (*LVEF*), expressed as a percentage, which represent the level of blood pressure and the heart's contractile capacity, respectively. Renal function is evaluated through the estimated glomerular filtration rate by EPI formula (*GFR-EPI*), measured in ml/min/1.73 m², while blood sodium concentration (*NATREMIA*) provides insights into electrolyte balance and fluid regulation, both relevant to cardiovascular function. Amino-terminal Brain Natriuretic Peptide (*NT-proBNP)*, a biomarker of HF severity, is also included and measured in pg/ml to quantify the condition's severity.

The dataset is completed by the levels of the thyroid hormones *FT3* and *FT4*, along with *TSH*, which offer valuable information about the patient's thyroid function. These variables are particularly significant for patients with thyroid dysfunction, given their potential impact on outcomes in HF.

This set of variables forms a robust and multidimensional data foundation essential for training ML models. Through this wide array of clinical and demographic features, the ML model can process complex details and identify significant patterns, thereby providing valuable support in predicting clinical risks and personalizing therapies for patients with HF and associated comorbidities.

3.2 Preprocessing and data sampling

These data were meticulously cleaned to eliminate anomalies and missing values, thereby ensuring the integrity of the dataset used for model training. The handling of 0.2% missing data was performed using the *model-based imputer* with a simple tree model, through Orange (version 3.36.2), a data mining software built on open-source Python libraries for scientific computing, such as NumPy and SciPy. The *Impute* widget was used for this purpose, allowing the construction of models to predict missing values based on the available data in other variables. With the integration of advanced Python libraries, Orange provides a powerful interface for imputation and scientific calculations, enabling accurate estimation of missing values with a simple decision tree while preserving dataset integrity, even with a low percentage of missing data.

Mathematically, the imputation process can be represented as follows: each missing value X_i is estimated using other observed variables X_{-i} through a function f derived from a simple decision tree, as shown in (Equation 1):

$$\widehat{X}_i = f(X_{-i}) \tag{1}$$

Where $\widehat{X}_i$ denotes the imputed value for the variable X_i, X_{-i} represents the set of all other observed variables used as predictors, and f is the function constructed by the decision tree to predict the missing values.

For continuous variables, this function imputes missing values as the mean of known values within the relevant leaf node, as described in (Equation 2):

$$\widehat{X}_i = \frac{1}{n} \sum_{j \in leaf} X_j \tag{2}$$

where n is the number of samples in the same leaf node and X_j represents each known value of X_i within that node. The summation $\sum_{j \in leaf} X_j$ calculates the total of known values for X_i in the node, with the division by n yielding the mean.

Equations 1, 2 together provide the general method for accurately filling in missing values, preserving dataset integrity for effective model training.

The dataset was divided into a training set (70%) and a validation set (30%), using this split to minimize the risk of overfitting and to verify the model's ability to generalize to unseen data. This split was done in Python using the *train_test_split* command of the *sklearn library.*

Formally, if we consider $\boldsymbol{X}$ as the set of independent variables (features) and $\boldsymbol{y}$ as the target, we can represent the data separation as shown in (Equations 3, 4):

$$(X_{train},\ y_{train}) = \{(X_i,\ y_i)|\ i \in Training\ set\} \tag{3}$$

$$(X_{test},\ y_{test}) = \{(X_i,\ y_i)|\ i \in Validation\ set\} \tag{4}$$

where X_{train} and y_{train} represent the features and targets of the training set, respectively, while X_{test} and y_{test} represent the features and targets of the validation set.

For each model, after training on the training set, we calculate evaluation metrics on the validation set to assess model performance. The evaluation function, denoted as *Metric*, measures the performance of the optimized model using the validation set observations, as shown in (Equation 5):

$$Metric = \frac{1}{N}\sum_{i=1}^{N} L(f(X_{test,i}\ ,\ \theta_{opt}),\ y_{test,i}) \tag{5}$$

Where $f(X_{test,i}\ ,\ \theta_{opt})$ is the model's prediction for test data point $X_{test,i}$, using the optimized parameters θ_{opt}. $y_{test,i}$ represents the actual target value for $X_{test,i}$. L is a loss function that quantifies the difference between the prediction and the actual value (e.g., mean squared error for regression or cross-entropy for classification). N is the number of observations in the validation set.

Equations 3, 4 describe the division of data into training and validation sets, while (Equation 5) defines the evaluation metric to assess model performance after optimization. This approach ensures that the model is tested on unseen data, providing a reliable measure of its generalization capabilities.

3.3 Validation and optimization process for ML models

We explored a broad range of ML algorithms, including Gradient Boosting, Naive Bayes, Random Forest, AdaBoost, Logistic Regression, SVM, SGD, Decision Trees, and KNN, optimizing each to enhance the accuracy of predictions for mortality and hospitalization risks (31). Previous studies have demonstrated the effectiveness of ML in cardiovascular risk stratification, showing that these models outperform traditional methods in handling complex datasets and modeling non-linear relationships, thus providing higher sensitivity and specificity (32, 33). The implementation was carried out in a Python environment, using advanced libraries such as *pandas, numpy*, and *scikit-learn*, with a script that managed data loading, cleaning, and splitting for model training and validation.

The selected features include 10 numerical and 11 categorical variables, as outlined in Table 1. After dividing the dataset into a training set (70%) and a validation set (30%) using the *train_test_split* function from *scikit-learn*, we created pipelines for each model, applying feature standardization via *StandardScaler*. Feature standardization was performed using the following formula (Equation 6):

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{6}$$

Where X represents the original value of the feature, μ is the mean of the feature values in the training set, σ is the standard deviation of the feature in the training set. This transformation scales the features to have a mean of zero and a standard deviation of one, improving the stability and performance of ML algorithms, especially those sensitive to data scaling.

We developed two distinct predictive models, focusing on mortality and hospitalization events as target variables for our patient cohort. Each model was trained separately on target-specific data and validated to ensure the reliability of the results. To minimize variance and improve the robustness of performance estimates, we applied 10-fold cross-validation, in line with established methods (34). The training process included a class balancing phase to address the data imbalance for mortality and hospitalization targets, a common issue in clinical datasets. Using SMOTE (Synthetic Minority Over-sampling Technique), we balanced the training set for each target by creating synthetic samples of the minority class, enhancing the models' ability to handle imbalanced data. This approach improved the sensitivity and specificity of the models, reducing the risk of misclassifying high-risk patients. The developed models were rigorously validated using standard metrics such as the AUC, accuracy, sensitivity, and specificity (35). For each model, we implemented a hyperparameter tuning phase using Python's *GridSearckCh,* a tool provided by the scikit-learn library that enables an exhaustive search for the optimal combination of hyperparameters to maximize model performance. *GridSearchCk* evaluates each combination specified in a predefined parameter grid, applying cross-validation to ensure that the performance obtained is representative and not overly dependent on the training data.

We used AUC as the primary metric for hyperparameter tuning, chosen because it represents the model's ability to correctly distinguish between classes, regardless of the classification threshold. AUC is particularly useful in medical contexts, where it is crucial to reduce both false positives and false negatives. A higher AUC indicates a more accurate model in predicting clinical events such as mortality and hospitalization, thereby improving the quality of therapeutic decision-making.

Formally, the optimization process aims to maximize AUC by selecting the optimal set of hyperparameters θ, and can be expressed as follows (Equation 7):

$$\theta^* = arg \max_{\theta \in \Theta} AUC(f(X_{train};\ \theta),\ y_{train}) \tag{7}$$

Where $\theta \in \Theta$ represents the set of hyperparameter combinations specified in the search grid, $f(X_{train};\ \theta)$ is the model's predictive function trained on the training data X_{train} with parameters θ, AUC is the evaluation metric that measures the area under the ROC curve, representing model performance relative to the true values y_{train}, θ^* is the combination of hyperparameters that maximizes AUC.

In Python, *GridSearchCV* applies cross-validation to each combination of hyperparameters θ, splitting the training set into k folds. The cross-validated mean AUC, denoted as $AUCcv$, for each fold can be expressed as (Equation 8):

$$AUC_{cv} = \frac{1}{k}\sum_{i=1}^{k} AUC(f(X_{train_i};\ \theta),\ y_{train_i}) \tag{8}$$

Where X_{train_i} and y_{train_i} represent the training data and targets for the i-th fold, respectively, k is the number of folds in the cross-validation. At the end of the procedure, *GridSearchCV* returns the combination of hyperparameters θ^*that maximizes the mean AUC across folds, providing an optimal configuration that represents the entire training set and minimizes the risk of overfitting. This approach ensures that the model is optimized for class discrimination, enhancing its generalizability to new data.

3.4 Selected ML models post-optimization

After the hyperparameter optimization process and using AUC as the primary metric to select the most effective model, Random Forest proved to be the best suited for predicting both the *Mortality* target (patient mortality event) and the *HF Hospitalization* target (patient hospitalization event). Model selection was based on comparing the average AUCs obtained through cross-validation for each model and target.

For predicting both the *Mortality* and *HF Hospitalization* targets, Random Forest showed optimal results. Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions to enhance the model's accuracy and robustness. The final prediction for each target using Random Forest, denoted as $f_{RF}(X)$, is obtained by averaging (for regression) or taking the majority vote (for classification) across the predictions from all trees, as shown in (Equation 9):

$$f_{RF}(X) = \frac{1}{N}\sum_{j=1}^{N} f_j(X) \tag{9}$$

Where N is the number of decision trees in the forest, $f_j(X)$ represents the prediction of the j-th tree for input X.

Each tree is trained on a randomly sampled subset of the training data with replacement, optimizing specifically for the *Mortality* and *HF Hospitalization* targets. The aggregation of predictions enhances the model's generalization ability, reducing the risk of overfitting and stabilizing its capacity to accurately predict both mortality and hospitalization events.

3.5 Data measurements

In our study, predictive models effectively differentiate between survival and mortality outcomes among HF patients. These models categorize observations based on their predictions: an outcome is identified as either an accurate mortality prediction (TP—true positive), an accurate survival prediction (TN—true negative), an incorrectly predicted mortality (FP -false positive), or a missed mortality (FN—false negative). This classification is vital for assessing the model's accuracy and utility in clinical settings.

The model's performance is evaluated using several metrics, which are crucial for ensuring accurate and reliable predictions:

- *AUC-ROC (Area Under the Curve—Receiver Operating Characteristics)*: Measures the model's discriminative ability between outcome classes. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) across varying thresholds u, and the AUC is calculated as (Equation 10):

$$AUC = \int_0^1 TPR\ [FPR^{-1}(u)]\ du \tag{10}$$

This integral covers all possible decision thresholds, providing a comprehensive measure of predictive accuracy.

- *Accuracy*: Represents the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances evaluated. It is defined by the following (Equation 11):

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

- *Precision*: Indicates the accuracy of positive predictions (mortality predictions), highlighting the model's ability to minimize false alarms, defined as (Equation 12):

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

- *Recall (Sensitivity)*: Reflects the model's ability to identify all actual positive instances (actual mortalitys), which is crucial for ensuring that no high-risk patients are overlooked, defined as (Equation 13):

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

- *F1 Score*: Combines precision and recall into a single metric, providing a balanced view of the model's overall predictive precision and sensitivity, defined as (Equation 14):

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision +\ Recall} \tag{14}$$

- *Matthews Correlation Coefficient (MCC):* A comprehensive measure that takes into account true and false positives and negatives, offering a balanced metric even for imbalanced datasets. The MCC is especially valuable as it ranges from −1 (total disagreement between predictions and actuals) to +1 (perfect prediction), with 0 indicating no predictive power, defined as (Equation 15):

$$MCC = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + \ FP)(TP + FN)(TN + \ FP)(TN + FN)}} \tag{15}$$

Utilizing these metrics ensures a thorough evaluation of the model's performance, facilitating improved clinical decision-making and patient management strategies in HF treatment. The integration of these diverse metrics, particularly AUC alongside precision, recall, and F1 score, supports the model's robustness, making it a valuable tool in clinical environments.

4 Results and discussion

In this section, we will discuss the selection of ML models used for risk prediction in patients with HF and thyroid dysfunctions, provide a detailed interpretation of the results for different thyroid subgroups, and introduce an experimental section on risk stratification. The objective is to explore the models' performance and evaluate their clinical applicability in the context of personalized risk management.

4.1 Performance of the selected predictive models

The results obtained from the optimized ML models for predicting mortality and hospitalization risks in patients with HF and thyroid dysfunctions are presented in Tables 2, 3. Each table includes a column labeled "Algorithm," which lists the ML algorithms considered in this study. Various algorithms known for their effectiveness in classification tasks were selected, including Random Forest, Stochastic Gradient Descent (SGD), Logistic Regression, Support Vector Machines (SVM), Gradient Boosting, AdaBoost, Naive Bayes, Neural Network, K-Nearest Neighbors (KNN), and Decision Tree. This variety of algorithms allows for a comprehensive comparison of performance, both in terms of predictive accuracy and the ability to balance key metrics such as precision, recall, and F1-score.

The performance of each algorithm was evaluated using metrics such as the AUC, accuracy, F1-score, precision, recall, and Matthews Correlation Coefficient (MCC). The AUC metric was particularly emphasized as the primary indicator of model performance, guiding the interpretation of results.

For mortality prediction, the Random Forest model achieved the best performance with an AUC of 0.797, an accuracy of 74.7%, and an F1-score of 0.685. These values indicate a good ability of the model to discriminate between high-risk and low-risk patients, balancing precision (0.768) and recall (0.618). The MCC for Random Forest was 0.485, further supporting its balanced performance across classes. This combination suggests that Random Forest is effective in identifying at-risk patients

TABLE 2 Model performance for mortality prediction.

Algorithm	AUC	Accuracy	F1	Precision	Recall	MCC
RandomForest	0.797	0.747	0.685	0.768	0.618	0.485
SGD	0.794	0.764	0.724	0.755	0.696	0.520
LogisticRegression	0.786	0.738	0.681	0.744	0.627	0.466
GradientBoosting	0.786	0.707	0.621	0.733	0.539	0.404
AdaBoost	0.762	0.721	0.660	0.721	0.608	0.430
SVM	0.759	0.729	0.667	0.738	0.608	0.448
NaiveBayes	0.753	0.690	0.585	0.725	0.490	0.369
NeuralNetwork	0.735	0.699	0.631	0.694	0.578	0.384
KNN	0.698	0.668	0.600	0.648	0.559	0.322
DecisionTree	0.608	0.624	0.522	0.603	0.461	0.227

TABLE 3 Model performance for HF hospitalization prediction.

Algorithm	AUC	Accuracy	F1	Precision	Recall	MCC
RandomForest	0.786	0.703	0.638	0.652	0.625	0.387
NeuralNetwork	0.785	0.725	0.659	0.685	0.635	0.430
LogisticRegression	0.784	0.729	0.687	0.667	0.708	0.449
SVM	0.779	0.725	0.683	0.660	0.708	0.442
NaiveBayes	0.769	0.690	0.643	0.621	0.667	0.370
SGD	0.763	0.712	0.673	0.642	0.708	0.418
GradientBoosting	0.746	0.681	0.597	0.635	0.563	0.336
KNN	0.727	0.664	0.645	0.579	0.729	0.342
AdaBoost	0.721	0.690	0.632	0.629	0.635	0.364
DecisionTree	0.641	0.659	0.606	0.588	0.625	0.307

while maintaining a low rate of false positives, making it particularly suitable for mortality prediction.

For hospitalization risk prediction, the Random Forest model again demonstrated the best performance, with an AUC of 0.786, an accuracy of 70.3%, and an F1-score of 0.638. With a precision of 0.652, recall of 0.625, and an MCC of 0.387, Random Forest effectively identifies patients at risk of hospitalization, maintaining a favorable balance between accuracy and sensitivity. This model's reliability for predicting hospitalization risk makes it a valuable tool for clinical applications where capturing at-risk patients is essential, even if it involves a slightly higher rate of false positives.

In summary, the results in Tables 2, 3 indicate that the Random Forest model is particularly promising for predicting both mortality and hospitalization risks. The AUC metric, used as the primary indicator, confirms the effectiveness of this model in providing robust decision support in clinical settings. Its application could significantly improve risk stratification and personalize treatments for patients with HF and thyroid dysfunctions, contributing to more precise and patient-centered medicine.

Figure 2 presents the confusion matrices for the top-performing ML model in predicting mortality and hospitalization risks, both achieved using the Random Forest algorithm: mortality prediction (left) and hospitalization prediction (right). These matrices are displayed in percentages, offering a comprehensive view of model performance regarding correct classifications and error rates. In the mortality prediction matrix (left), the Random Forest model correctly identified 85.04% of low-risk patients (class 0), while 14.96% of these patients were incorrectly classified as high-risk. For the high-risk group (class 1), the model correctly classified 61.76% of patients but misclassified 38.24% as low-risk. These results indicate that, while the Random Forest model has high precision for predicting low-risk patients, its sensitivity in identifying high-risk cases is moderate.

For hospitalization prediction (right), the Random Forest model accurately classified 75.94% of patients not at risk (class 0), with 24.06% misclassified as at-risk. In the at-risk group (class 1), 62.50% of patients were correctly identified, while 37.50% were classified as false negatives. This performance shows that the Random Forest model is effective in predicting hospitalization risk, maintaining a reasonable balance between precision and recall for at-risk patients.

Figure 2 illustrates the strengths and limitations of the Random Forest model in both predictive tasks. The model shows high accuracy for the low-risk mortality class but misses a significant portion of high-risk cases. Similarly, it performs well in predicting hospitalization risk but also exhibits some false negatives within the high-risk group. The model demonstrates a satisfactory balance between accuracy and sensitivity, reinforcing its clinical applicability for risk stratification.

Figure 3 shows the Receiver Operating Characteristic (ROC) curves for the Random Forest model in predicting mortality and hospitalization risks: mortality prediction (left) and hospitalization prediction (right). The ROC curve illustrates the model's ability to distinguish between classes, plotting the relationship between True Positive Rate (Sensitivity) and False Positive Rate. The Area Under the Curve reflects model performance, where values closer to 1 indicate greater discriminatory power. For mortality prediction, the Random Forest model achieved an AUC of 0.797, as depicted in the left ROC curve, demonstrating a strong capability to differentiate between high and low mortality risk. The ROC curve remains well above the reference line (indicating random classification) across thresholds, showcasing the Random Forest model's ability to sustain a high True Positive Rate while minimizing False Positives. For hospitalization prediction, the Random Forest model achieved an AUC of 0.786, as shown in the right ROC curve. Although slightly lower than the AUC for mortality

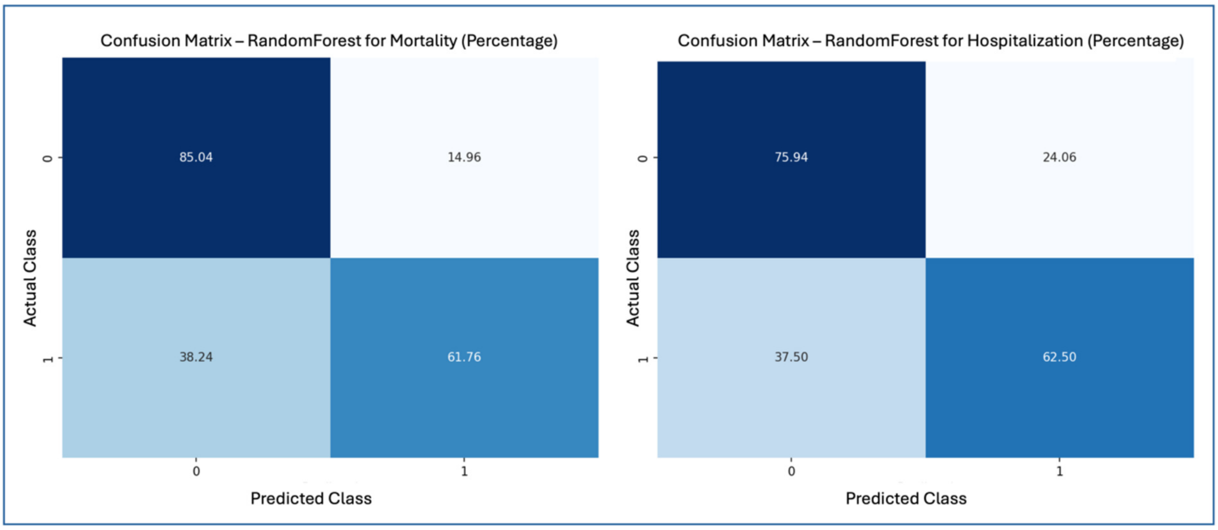


FIGURE 2
Confusion matrices—random forest for mortality and hospitalization.

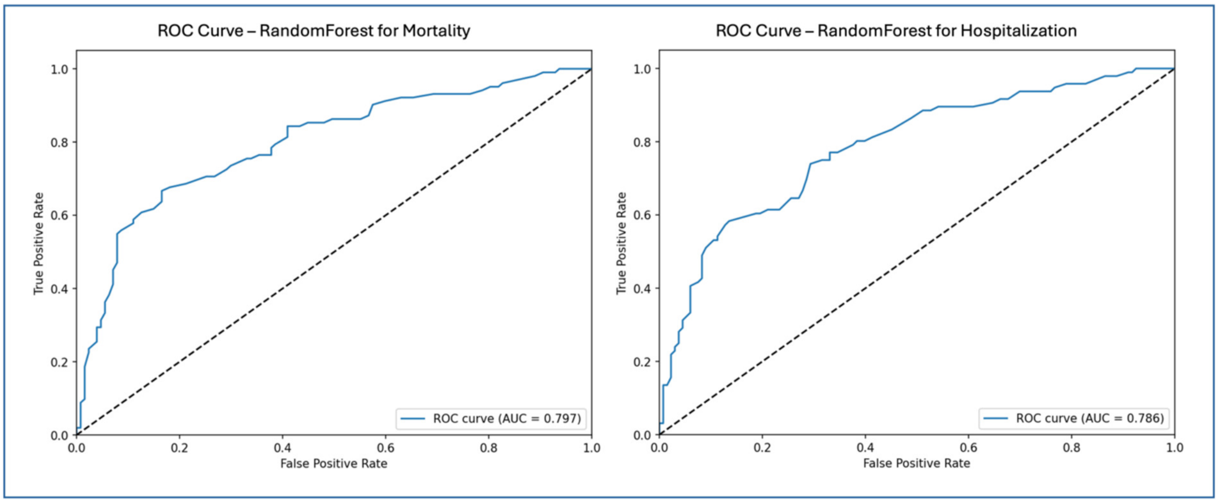


FIGURE 3
ROC curves—random forest for mortality and hospitalization.

prediction, this value still reflects strong performance in identifying hospitalization risk. The ROC curve for the Random Forest model stays above the reference line, indicating good model sensitivity and specificity in distinguishing hospitalized from non-hospitalized patients.

Figure 3 highlights the effective performance of the Random Forest model in both prediction tasks. The AUC values for mortality and hospitalization predictions confirm the model's suitability for clinical risk stratification. The ROC curves emphasize the model's capacity to balance True Positive and False Positive rates, reinforcing its utility as a reliable tool for clinical decision-making in managing patients with HF and thyroid dysfunction.

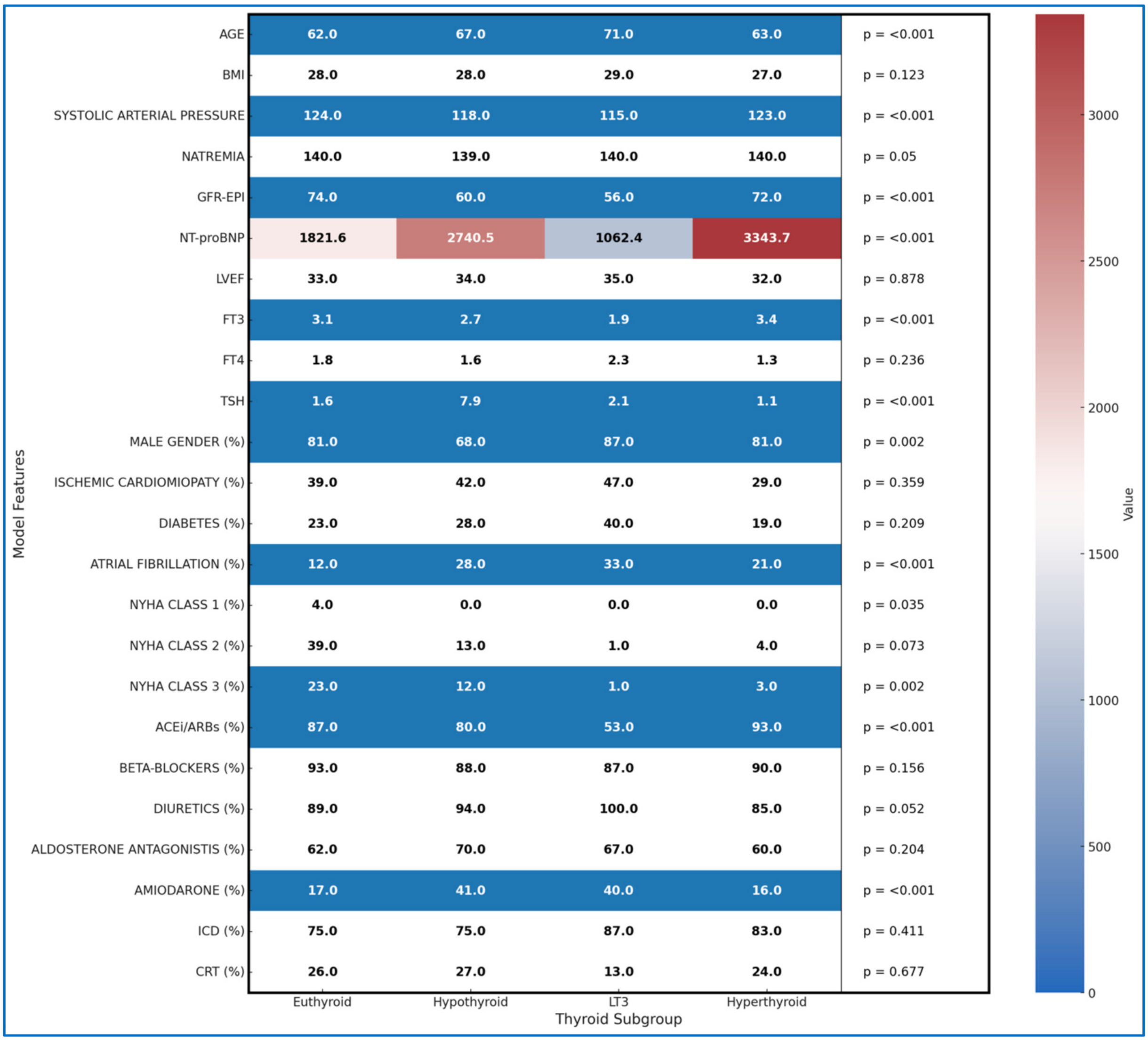
4.2 Analysis of clinical and statistical differences among thyroid subgroups

Among the 762 patients analyzed, 187 were affected by hypothyroidism; of these, 93 had a prior history of hypothyroidism, while in 94 cases, hypothyroidism was diagnosed during the initial or subsequent evaluations at our center. LT3 syndrome was diagnosed in 15 patients, while a total of 58 patients had hyperthyroidism, with 46 having a prior history and 12 diagnosed at the time of the first evaluation or during follow-up.

Figure 4 presents the statistical characteristics of the patients, divided into subgroups based on the presence or absence of thyroid disorders, providing a detailed overview of demographic variables, risk factors, and ongoing therapies for each subgroup. This arrangement allows for an in-depth comparison of clinical differences among patients with various thyroid dysfunctions. Among the patients, 175 were on amiodarone therapy at the time of the initial evaluation: 63 for secondary prevention of supraventricular tachycardia or flutter/atrial fibrillation, 73 for secondary prevention of sustained ventricular tachycardia/ventricular fibrillation, 24 for both, and 15 for control of frequent supraventricular or ventricular ectopic beats. To compare characteristics across the different thyroid groups, the Kruskal–Wallis test was used, a non-parametric test suitable for variables that do not follow a normal distribution. This statistical method allows significant differences to be detected among multiple groups without assuming normality, which is particularly useful given the nature of clinical variables, which are both continuous and categorical. In the heatmap (Figure 4), significant differences ($p < 0.005$) are visually highlighted using a blue background with white text, allowing immediate identification of key variables. Additionally, NT-proBNP values are color-coded using a gradient that reflects their magnitude in relation to the scale shown in the accompanying color bar, facilitating intuitive comparison across subgroups.

The results indicate that the mean age differs significantly between groups ($p < 0.001$), with patients with LT3 syndrome being older on average (71 years) than euthyroid patients (62 years). Systolic blood pressure and renal function, measured by GFR-EPI, also show significant differences ($p < 0.001$); hypothyroid and LT3 patients have lower average values, suggesting possible involvement of cardiovascular and renal function. NT-proBNP levels, an indicator of HF severity, are significantly higher in hypothyroid and hyperthyroid patients compared to euthyroid patients, reflecting a higher degree of clinical impairment ($p < 0.001$).

Thyroid function parameters, such as FT3 and TSH, also differ significantly among the groups. LT3 patients have the lowest average FT3 levels compared to the other subgroups, while hypothyroid patients show elevated TSH levels ($p < 0.001$). Atrial fibrillation is more common in patients with thyroid dysfunctions, particularly among those with LT3 and hypothyroidism, with percentages of 33% and 28%, respectively, compared to euthyroid patients (12%), suggesting an increased predisposition to arrhythmic events in the presence of thyroid disorders ($p < 0.001$).



Model Features	Euthyroid	Hypothyroid	LT3	Hyperthyroid	
AGE	62.0	67.0	71.0	63.0	p = <0.001
BMI	28.0	28.0	29.0	27.0	p = 0.123
SYSTOLIC ARTERIAL PRESSURE	124.0	118.0	115.0	123.0	p = <0.001
NATREMIA	140.0	139.0	140.0	140.0	p = 0.05
GFR-EPI	74.0	60.0	56.0	72.0	p = <0.001
NT-proBNP	1821.6	2740.5	1062.4	3343.7	p = <0.001
LVEF	33.0	34.0	35.0	32.0	p = 0.878
FT3	3.1	2.7	1.9	3.4	p = <0.001
FT4	1.8	1.6	2.3	1.3	p = 0.236
TSH	1.6	7.9	2.1	1.1	p = <0.001
MALE GENDER (%)	81.0	68.0	87.0	81.0	p = 0.002
ISCHEMIC CARDIOMIOPATY (%)	39.0	42.0	47.0	29.0	p = 0.359
DIABETES (%)	23.0	28.0	40.0	19.0	p = 0.209
ATRIAL FIBRILLATION (%)	12.0	28.0	33.0	21.0	p = <0.001
NYHA CLASS 1 (%)	4.0	0.0	0.0	0.0	p = 0.035
NYHA CLASS 2 (%)	39.0	13.0	1.0	4.0	p = 0.073
NYHA CLASS 3 (%)	23.0	12.0	1.0	3.0	p = 0.002
ACEi/ARBs (%)	87.0	80.0	53.0	93.0	p = <0.001
BETA-BLOCKERS (%)	93.0	88.0	87.0	90.0	p = 0.156
DIURETICS (%)	89.0	94.0	100.0	85.0	p = 0.052
ALDOSTERONE ANTAGONISTIS (%)	62.0	70.0	67.0	60.0	p = 0.204
AMIODARONE (%)	17.0	41.0	40.0	16.0	p = <0.001
ICD (%)	75.0	75.0	87.0	83.0	p = 0.411
CRT (%)	26.0	27.0	13.0	24.0	p = 0.677

FIGURE 4
Heatmap of all clinical features by thyroid subgroup.

The distribution of patients across NYHA classes reveals further differences, with lower representation of thyroid dysfunction patients in the more advanced classes ($p = 0.002$), potentially reflecting a different severity of symptoms among groups. In terms of pharmacological therapies, hypothyroid and LT3 patients are more frequently treated with diuretics and amiodarone compared to euthyroid patients, with statistically significant differences for the use of ACEi/ARBs and amiodarone ($p < 0.001$), which may indicate specific therapeutic needs for these subgroups.

These differences between thyroid groups provide a deeper understanding of the distinctive clinical profiles associated with thyroid dysfunctions, highlighting how clinical risk and therapeutic needs may vary based on thyroid status. The detailed statistical breakdown in Figure 4, along with the Kruskal–Wallis test, provides valuable information for a better understanding of the clinical specificities of each group, supporting the implementation of more targeted therapeutic strategies.

4.3 Interpretation of model results with LIME for thyroid subgroups

This section applies the Local Interpretable LIME technique to interpret the Random Forest model results, focusing on specific subgroups within thyroid-related patient populations. LIME enables the interpretation of complex models by creating locally interpretable models around individual predictions, allowing us to examine the contribution of each variable to the model's final decisions. The LIME technique was applied uniformly across all thyroid-related subgroups to support the interpretability of the model predictions. For each subgroup, the approach enabled the identification of clinical variables such as atrial fibrillation, ischemic cardiomyopathy, pharmacological treatment, and thyroid hormone values, contributing to the estimated risks of mortality and hospitalization. Illustrative examples of these explanations are presented in Figures 4, 5–11, including euthyroid, hypothyroid, LT3 and Hyperthyroid patient groups, thus offering a consistent interpretation framework across the cohort. In the graphical representations (Figures 5–11, 12), the impact of each clinical feature is visually represented through color-coded horizontal bars. Specifically, green bars indicate features that contribute to an increase in the predicted probability of the outcome (e.g., mortality or hospitalization), suggesting a higher risk associated with those variables. Conversely, red bars represent features that reduce the predicted probability, thus being protective factors associated with a lower risk. This visual distinction enhances interpretability by allowing a quick understanding of whether each feature pushes the model prediction toward or away from a critical outcome.

For each thyroid subgroup, LIME was applied to generate explanations that illustrate how key clinical factors modulate the model's predictions vary based on key clinical features, such as TSH levels, T3 and T4 hormone concentrations, and patient demographics. By analyzing these explanations, we can gain a clearer understanding of which features drive the model's predictions for each thyroid subgroup, distinguishing between low and high-risk classifications for both mortality and hospitalization.

Figures 5, 6 present the LIME interpretation results for the mortality and hospitalization models, respectively, in euthyroid patients. These figures list the main clinical features that impact the model's predictions. The impact values reflect the influence of each feature on the predicted probability, with positive values indicating features that contribute toward the outcome (e.g., mortality or hospitalization), while negative values indicate protective associations.

In Figure 5, titled "LIME Explanations for Mortality Prediction Model in Euthyroid Patients," the model shows a 53%predicted probability for "YES MORTALITY" vs. 47% for "NO MORTALITY," suggesting a slight inclination toward mortality for this subgroup. Among the influential features, the absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) shows a protective effect with an impact of −0.0311, lowering the mortality probability. Conversely, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) slightly increases the likelihood of mortality, with an impact value of 0.0275. Other features contribute with varying, though smaller, effects. For instance, the absence of the medication Amiodarone (AMIODARONE = 0) and of cardiac resynchronization therapy (CRT = 0) display minor positive impacts of 0.0077 and 0.0067, respectively, indicating an association with increased mortality when these treatments are not administered. Lower levels of TSH (≤0.96) reduce the probability of mortality with an impact of −0.0055, while the absence of diabetes (DIABETES = 0) has a similarly protective effect, with an impact of −0.0048. Minimal impacts are observed for free T4 levels (FT4 ≤1.08), BMI (23), and the use of diuretics and beta-blockers, with values ranging between 0.0005 and 0.0007, suggesting a more subtle influence on mortality risk in this model.

Figure 6, "LIME Explanations for Hospitalization Prediction Model in Euthyroid Patients," presents results for hospitalization prediction with identical predicted probabilities to the mortality model (53% for "YES HOSPITALIZATION" and 47% for "NO HOSPITALIZATION"), indicating a similar risk profile in this patient subgroup.

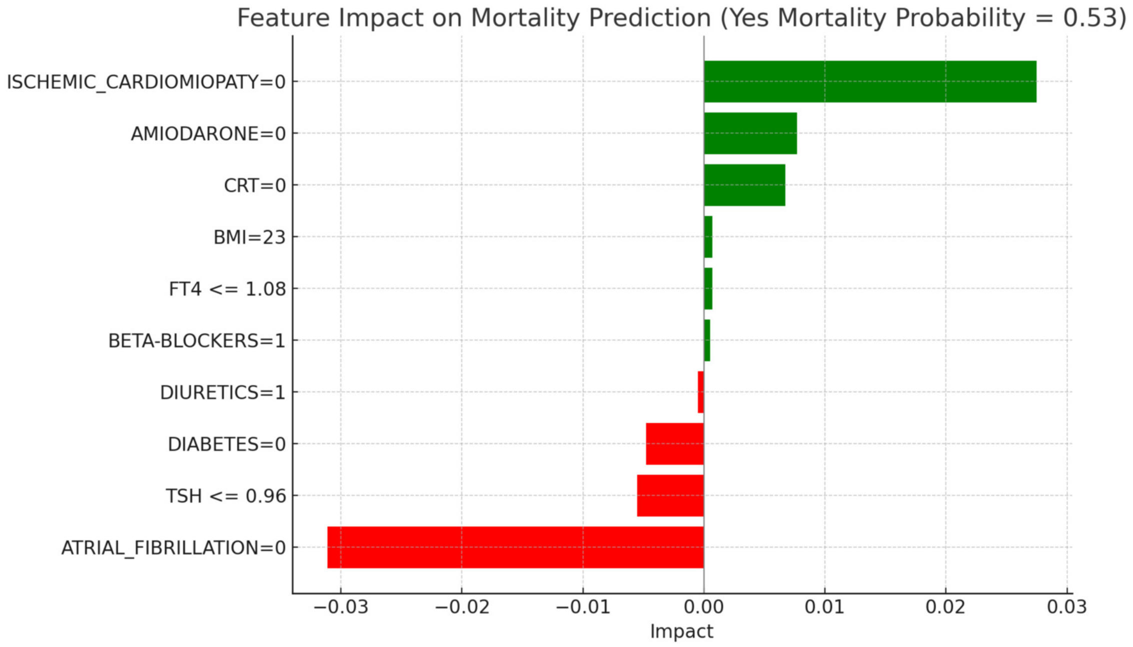


FIGURE 5
LIME explanations for mortality prediction model in euthyroid patients.

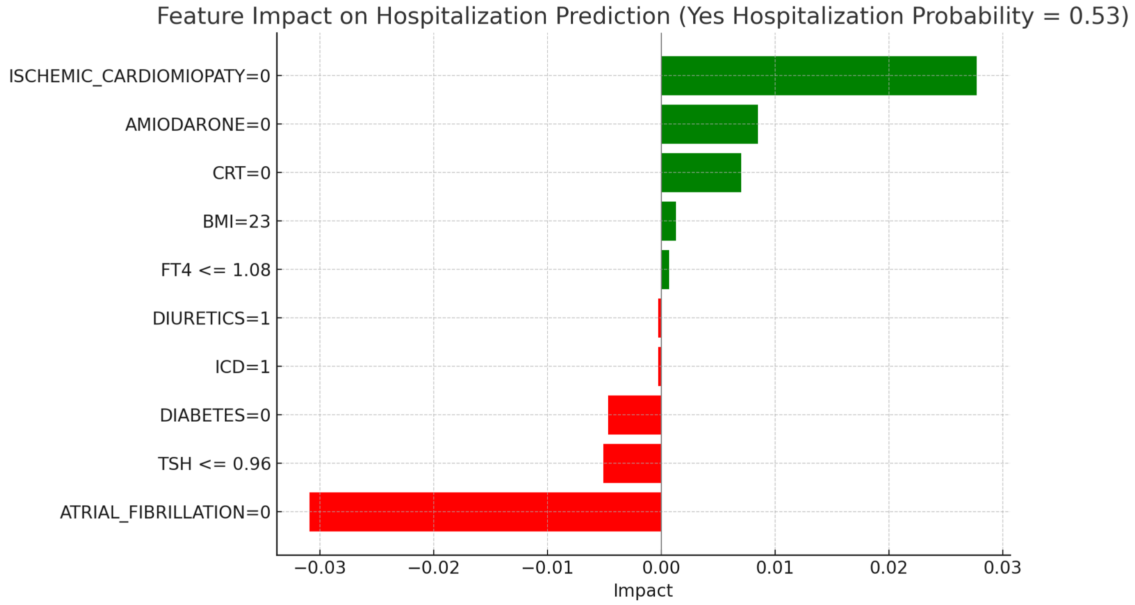


FIGURE 6
LIME explanations for hospitalization prediction model in euthyroid patients.

The absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) has a protective impact, reducing the likelihood of hospitalization with an impact value of −0.0309. Conversely, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) slightly increases the risk, showing a positive impact of 0.0277. The absence of Amiodarone (AMIODARONE = 0) and CRT (CRT = 0) also contribute to an increased hospitalization probability, with impact values of 0.0085 and 0.0070, respectively. Lower TSH levels (≤0.96) provide a protective influence with an impact of −0.0051, while the absence of diabetes (DIABETES = 0) similarly reduces the likelihood of hospitalization, reflected by an impact of −0.0047. BMI of 23 has a minor positive influence of 0.0013, indicating a slightly increased hospitalization probability for patients with this BMI value. Additional features with minimal impacts include free T4 levels (FT4 ≤1.08), presence of an ICD (ICD = 1), and the use of diuretics (DIURETICS = 1), each with values of 0.0007, −0.0003, and −0.0003 respectively. These factors suggest a nuanced, though limited, influence on the overall hospitalization prediction compared to the primary variables in this model.

Figure 7, “LIME Explanations for Mortality Prediction Model in Hypothyroid Patients,” presents the model’s interpretation results for the mortality prediction in hypothyroid patients, with 54% predicted probability for “YES MORTALITY” and 46% for “NO MORTALITY,” indicating a slight inclination toward mortality in this group.

In this model, the absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) serves as a protective factor, reducing the mortality probability with an impact of −0.0305. On the other hand, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) slightly increases the mortality risk, with a positive impact of 0.0213. The lack of Amiodarone (AMIODARONE = 0) and CRT (CRT = 0) also contribute to an elevated mortality probability, with impacts of 0.0076 and 0.0032, respectively. Other clinical variables influence mortality predictions to a lesser degree. The absence of diabetes (DIABETES = 0) decreases mortality risk, with an impact of −0.0024, while a BMI of 19 has a slight positive effect of 0.0023, indicating a marginal association with increased mortality. The use of diuretics (DIURETICS = 1) and beta-blockers (BETA-BLOCKERS = 1) exert small impacts, with values of 0.0012 and −0.0005, respectively, highlighting their limited role in influencing mortality predictions. Additional factors, such as NYHA class (NYHA_CLASS = 3) and FT3 levels within the range 2.70 < FT3 ≤ 3.00, have minimal impacts of 0.0004 each, suggesting a nuanced but relatively insignificant influence on the model’s overall prediction for mortality. In hypothyroid patients, the predicted probability of mortality was 54 percent. The absence of atrial fibrillation emerged as the most protective factor, aligning with its recognized clinical relevance in heart failure prognosis. Conversely, the absence of ischemic cardiomyopathy contributed to a moderate increase in predicted mortality, potentially reflecting the influence of alternative etiologies. Other variables, such as the lack of amiodarone therapy, absence of CRT, and a low BMI value, were associated with slightly elevated risk. FT3 values within borderline ranges and NYHA class exerted minor effects, confirming the multifactorial nature of mortality risk in this subgroup.

Figure 8, “LIME Explanations for Hospitalization Prediction Model in Hypothyroid Patients,” outlines the hospitalization prediction for hypothyroid patients, with 54% probability for “YES HOSPITALIZATION” and 46% for “NO HOSPITALIZATION,” again indicating a slight model tendency towards predicting hospitalization.

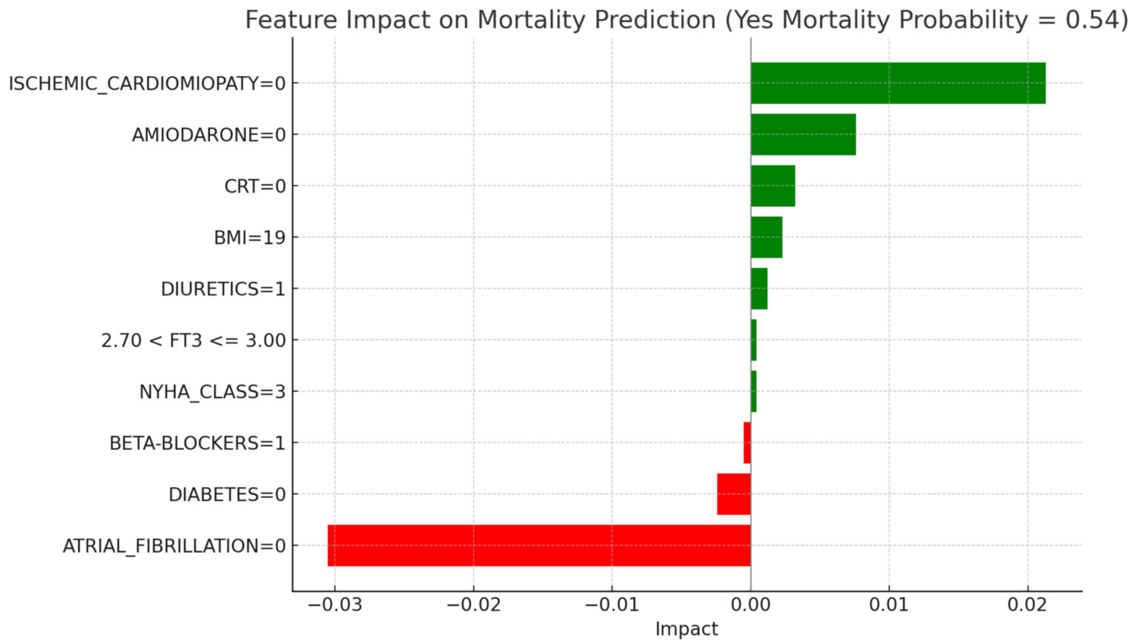


FIGURE 7
LIME explanations for mortality prediction model in hypothyroid patients.

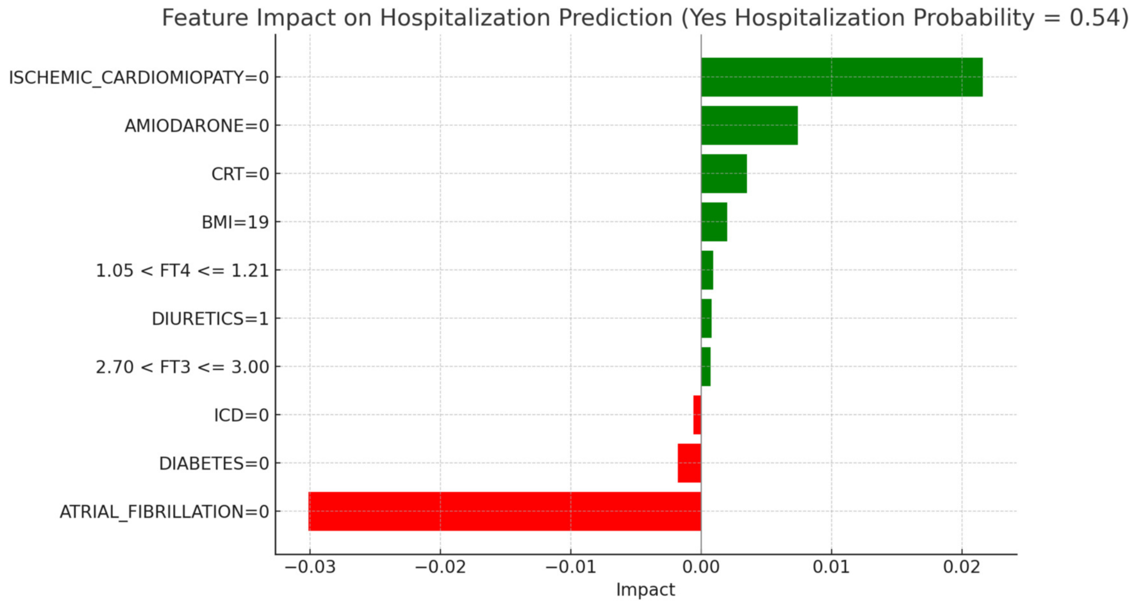


FIGURE 8
LIME explanations for hospitalization prediction model in hypothyroid patients.

Key protective factors include the absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0), which reduces the hospitalization risk with an impact of −0.0301. Meanwhile, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) slightly increases the hospitalization likelihood, with an impact of 0.0216. The absence of Amiodarone (AMIODARONE = 0) and CRT (CRT = 0) contribute positively, with impacts of 0.0074 and 0.0035, respectively, indicating that their absence may slightly increase hospitalization risk. Further influencing factors include BMI of 19, which has a minor positive impact of 0.0020 on hospitalization probability, and the absence of diabetes (DIABETES = 0), which has a small protective effect with an impact of −0.0018. Free T4 levels within the range 1.05 < FT4 ≤ 1.21 and FT3 levels within 2.70 < FT3 ≤ 3.00 add slight positive contributions, with impacts of 0.0009 and 0.0007, respectively. Finally, the presence of an ICD (ICD = 0) serves as a minor protective factor, with an impact of −0.0006, while diuretic usage (DIURETICS = 1) has a modest positive effect of 0.0008. These features, though present, exert relatively small effects in comparison to the more influential clinical factors impacting hospitalization predictions in this subgroup. In hypothyroid patients, the LIME interpretation results suggest a moderate increase in hospitalization risk, with a predicted probability of 54%. The absence of atrial fibrillation emerged as the most protective factor, consistent with its known adverse prognostic role in heart failure populations. Conversely, the absence of ischemic cardiomyopathy contributed positively to the predicted probability, potentially indicating the clinical impact of non-ischemic HF phenotypes in this subgroup. The absence of amiodarone and CRT therapy also showed modest positive contributions, aligning with the established utility of these interventions in selected HF patients. A lower BMI (19) was associated with a slight increase in predicted hospitalization, in line with the "obesity paradox" described in HF literature. Additionally, borderline FT4 and FT3 values exerted limited but noticeable effects, confirming the relevance of thyroid hormone levels in influencing short-term outcomes in this subgroup.

Figure 9, "LIME Explanations for Mortality Prediction Model in LT3 Patients," shows the model's interpretation results for mortality prediction in LT3 patients, with a predicted probability of 52% for "YES MORTALITY" and 48% for "NO MORTALITY," indicating a slight inclination towards predicting mortality for this group. Among the significant features, the absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) reduces the likelihood of mortality, acting as a protective factor with an impact of −0.0288. Conversely, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) slightly increases the probability of mortality, with a positive impact of 0.0226. Additionally, the use of Amiodarone (AMIODARONE = 1) appears to lower the mortality risk, indicated by an impact of −0.0094. BMI at 23 also has a slight protective influence, with an impact of −0.0034, while free T3 (FT3) levels in the range 2.00 < FT3 ≤ 2.10 contribute positively to mortality risk, showing an impact of 0.0031. The absence of CRT (CRT = 0) adds a minor positive influence with an impact of 0.0029, suggesting a potential association with increased mortality in LT3 patients when CRT is not in place. Other features play smaller roles: the absence of diabetes (DIABETES = 0) has a slight protective effect on mortality with an impact of −0.0015, and high levels of FT4 (>1.52) further reduce the probability of mortality with an impact of −0.0011. Additional factors, such as TSH levels between 2.30 and 2.75 and LVEF (Left Ventricular Ejection Fraction) values within 24.50–34.75, contribute minimally to the model's mortality predictions, with impacts of 0.0006 and −0.0005 respectively. Among LT3 patients, the model indicated a 52% probability of mortality. The strongest protective effect was associated with the absence of atrial fibrillation, while the absence of ischemic cardiomyopathy slightly increased predicted risk. The presence of amiodarone was linked to a lower mortality probability, possibly reflecting its therapeutic role in rhythm control. Hormonal indicators such as FT3 in the range 2.00–2.10 and higher FT4 levels provided subtle but consistent contributions. Overall, the results illustrate the complex

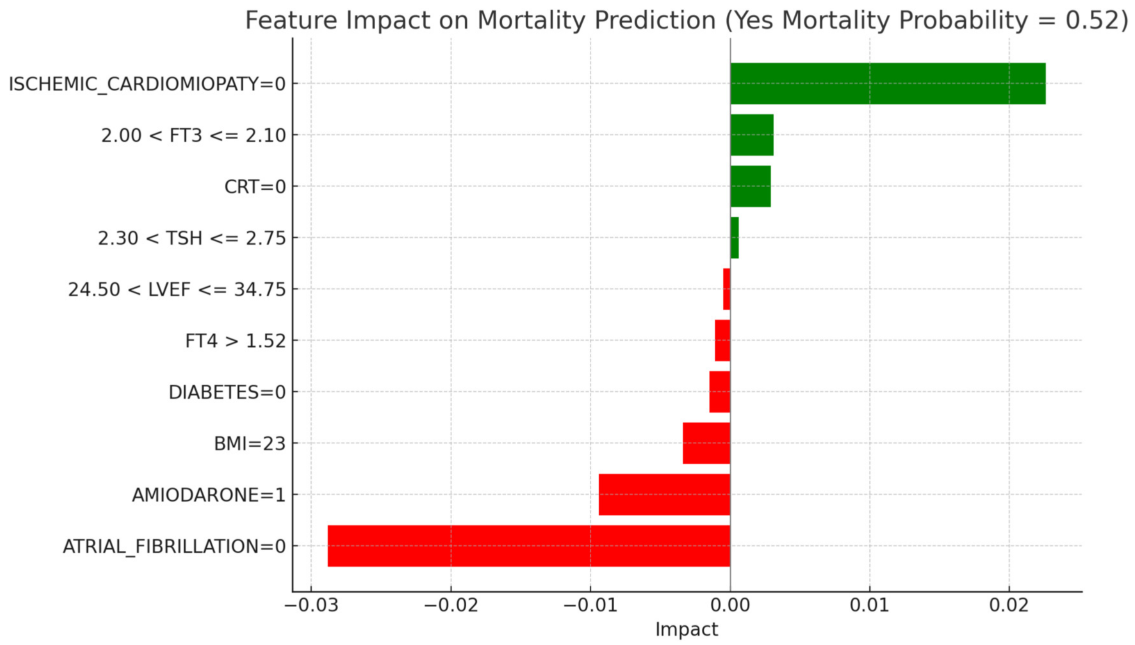


FIGURE 9
LIME explanations for mortality prediction model in LT3 patients.

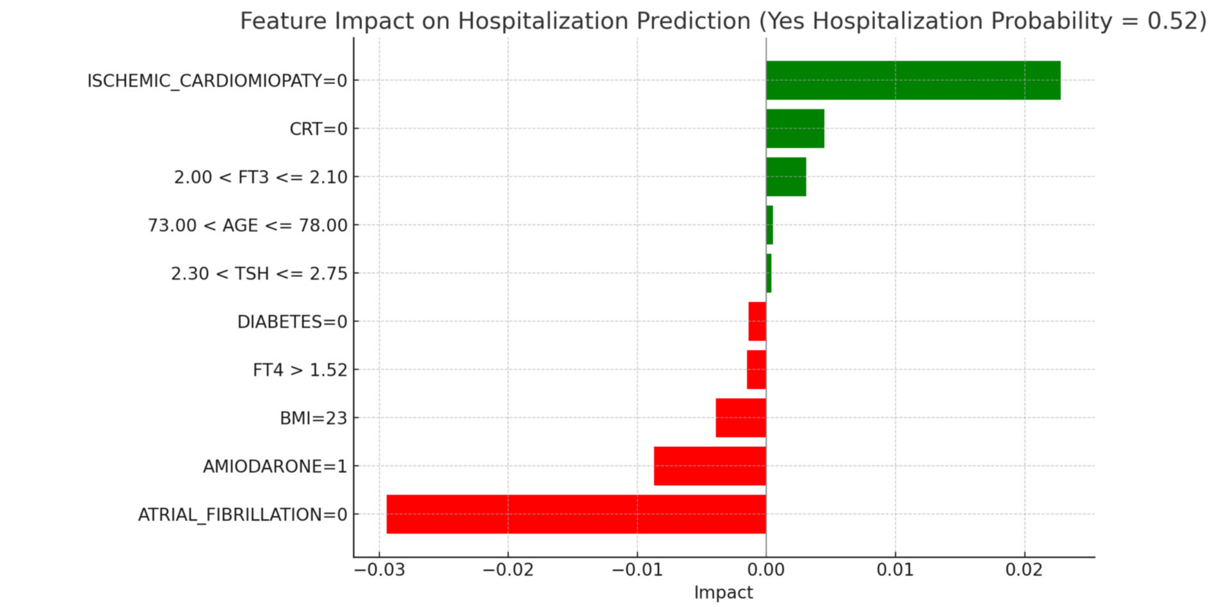


FIGURE 10
LIME explanations for hospitalization prediction model in LT3 patients.

interplay between metabolic, structural, and treatment-related factors in shaping risk within this distinct population.

Figure 10, "LIME Explanations for Hospitalization Prediction Model in LT3 Patients," provides insights into the model's predictions for hospitalization within this group. The model shows a 52% predicted probability for "YES HOSPITALIZATION" and 48% for "NO HOSPITALIZATION," again reflecting a slight tendency towards hospitalization risk. The absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) has the strongest protective effect, reducing the probability of hospitalization with an impact of −0.0294. In contrast, the absence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 0) is associated with a slight increase in hospitalization likelihood, with an impact of 0.0228. Use of Amiodarone (AMIODARONE = 1) similarly lowers the hospitalization risk, shown by an impact of −0.0087. The absence of CRT (CRT = 0) shows a positive influence on hospitalization probability with an impact of 0.0045, while BMI at 23 has a protective impact with a value of −0.0039. Free T3 levels within $2.00 < \text{FT3} \leq 2.10$ contribute a minor positive influence on hospitalization, with an impact of 0.0031, indicating a small association with increased risk for patients in this range. Other variables include FT4 levels greater than 1.52, which lower hospitalization probability with an impact of −0.0015, and the absence of diabetes (DIABETES = 0), which also acts protectively with an impact of −0.0014. Age within 73.00–78.00 years and TSH levels in the range $2.30 < \text{TSH} \leq 2.75$ exert minimal positive influences on hospitalization, with impacts of 0.0005 and 0.0004, respectively, suggesting limited yet present contributions in the model's hospitalization prediction. In LT3 syndrome patients, the predicted probability of hospitalization was 52%, indicating a subtle shift towards higher risk in this group. The absence of atrial fibrillation was again the most significant protective variable. Notably, the presence of amiodarone was associated with a lower predicted risk, which may reflect its therapeutic role in arrhythmia management among patients with compromised metabolic status. The absence of CRT demonstrated a minor positive impact on hospitalization probability, in line with its potential benefits in patients with advanced HF and electrical dyssynchrony. BMI at 23 appeared to exert a small protective influence, while FT3 values in the 2.00–2.10 range were associated with a mild increase in risk, consistent with reduced metabolic activity typical of LT3. Other features, including elevated FT4, absence of diabetes, and mid-range TSH values, showed marginal impacts, reinforcing the multifactorial nature of hospitalization risk in this complex subgroup.

Figure 11, "LIME Explanations for Mortality Prediction Model in Hyperthyroid Patients," shows the model's interpretation results for mortality prediction in hyperthyroid patients, with a predicted probability split evenly at 50% for "YES MORTALITY" and 50% for "NO MORTALITY," indicating no strong inclination towards either outcome in this group.

Key protective factors include the absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0), which reduces the mortality probability with an impact of −0.0303, and the presence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 1), which surprisingly acts as a protective factor in this model, with an impact of −0.0271. Conversely, the absence of Amiodarone (AMIODARONE = 0) contributes positively to mortality risk, with an impact of 0.0097. The absence of diabetes (DIABETES = 0) provides a protective effect with an impact of −0.0070, while TSH levels between 0.23 and 1.02 slightly increase the risk, with an impact of 0.0031. The presence of CRT (CRT = 1) also reduces the mortality probability, with an impact of −0.0027, indicating a marginal protective role. Other variables, such as a BMI of 18 and the use of ACE inhibitors or ARBs (ACEi/ARBs = 1), exert minor

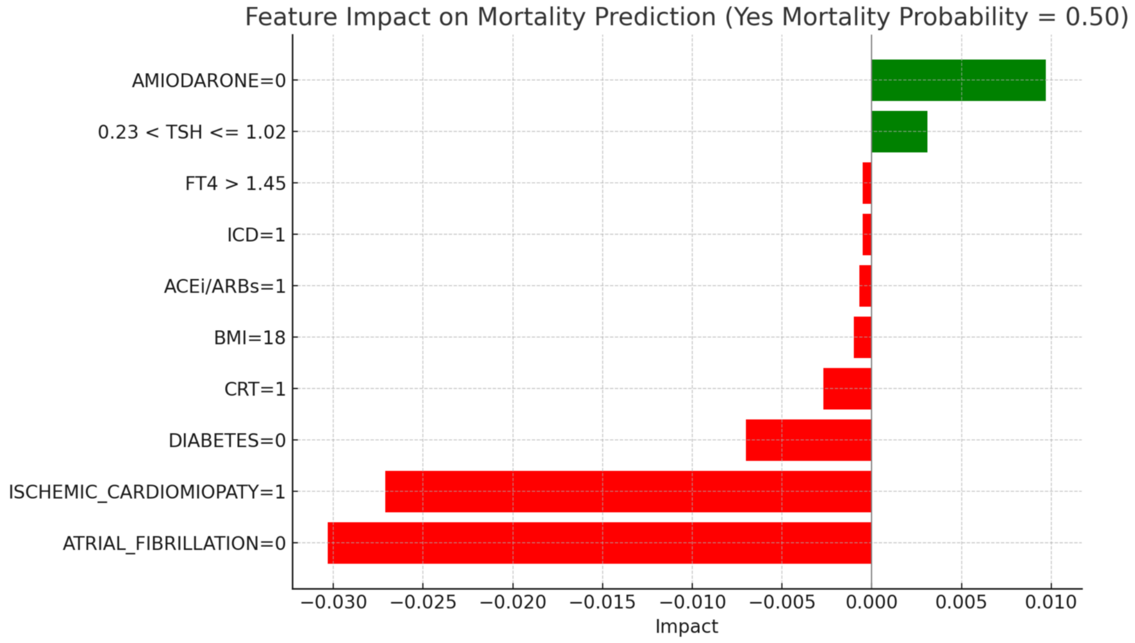


FIGURE 11
LIME explanations for mortality prediction model in hyperthyroid patients.

protective effects, with impacts of −0.0010 and −0.0007, respectively. Finally, the presence of an ICD (ICD = 1) and FT4 levels above 1.45 contribute minimally to reducing mortality, each with an impact of −0.0005.

Figure 12, "LIME Explanations for Hospitalization Prediction Model in Hyperthyroid Patients," provides insights into the model's predictions for hospitalization. Here, the predicted probabilities are also evenly split, with 50% for "YES HOSPITALIZATION" and 50% for "NO HOSPITALIZATION," indicating no dominant prediction tendency within this patient group.

The absence of atrial fibrillation (ATRIAL_FIBRILLATION = 0) serves as the strongest protective factor, reducing the hospitalization probability with an impact of −0.0309. Similarly, the presence of ischemic cardiomyopathy (ISCHEMIC_CARDIOMYOPATHY = 1) reduces hospitalization likelihood, with an impact of −0.0270. On the other hand, the absence of Amiodarone (AMIODARONE = 0) slightly increases the risk, with an impact of 0.0100. The absence of diabetes (DIABETES = 0) has a protective impact of −0.0076 on hospitalization probability. TSH levels in the range $0.23 < \text{TSH} \leq 1.02$ contribute a slight positive influence on hospitalization risk, with an impact of 0.0035. The presence of CRT (CRT = 1) also has a minor protective effect, with an impact of −0.0021, while a BMI of 18 provides additional protection with an impact of −0.0017. Other features exerting limited impacts include LVEF levels within 26.79–30.77, which slightly increase hospitalization likelihood (impact of 0.0005), while the use of ACE inhibitors or ARBs (ACEi/ARBs = 1) adds a minimal positive impact of 0.0004. Age over 70 (AGE >70) serves as a slight protective factor, with an impact of −0.0004, indicating a very marginal influence on hospitalization predictions. These features, though impactful to some extent, play a relatively small role in the overall predictions for mortality and hospitalization in hyperthyroid patients, highlighting the model's balanced treatment of features in predicting outcomes for this group.

4.4 Experimental risk stratifications

In this section, we present an experimental approach to risk stratification, where we evaluate and combine the probabilities of mortality and hospitalizations for patients across different thyroid classes and in various optimization scenarios. This approach aims to develop a risk stratification framework that can identify patients at high risk, facilitating targeted interventions. The process utilizes a multi-objective optimization strategy with four scenarios, ultimately visualized in a combined heatmap to summarize risk levels across groups. Our goal is to analyze and combine the risk of Mortality and Hospitalization across four thyroid classes: Euthyroid, Hypothyroid, LT3, and Hyperthyroid. This analysis is performed under four scenarios:

1. Maximize Mortality and Maximize Hospitalization: This scenario identifies conditions that maximize both risks.
2. Maximize Mortality and Minimize Hospitalization: This scenario targets patients with high risk of Mortality but lower risk of Hospitalization.
3. Minimize Mortality and Maximize Hospitalization: This scenario focuses on minimizing Mortality risk while maintaining a higher Hospitalization risk.
4. Minimize Mortality and Minimize Hospitalization: This scenario seeks to minimize both risks, representing the lowest overall risk profile.

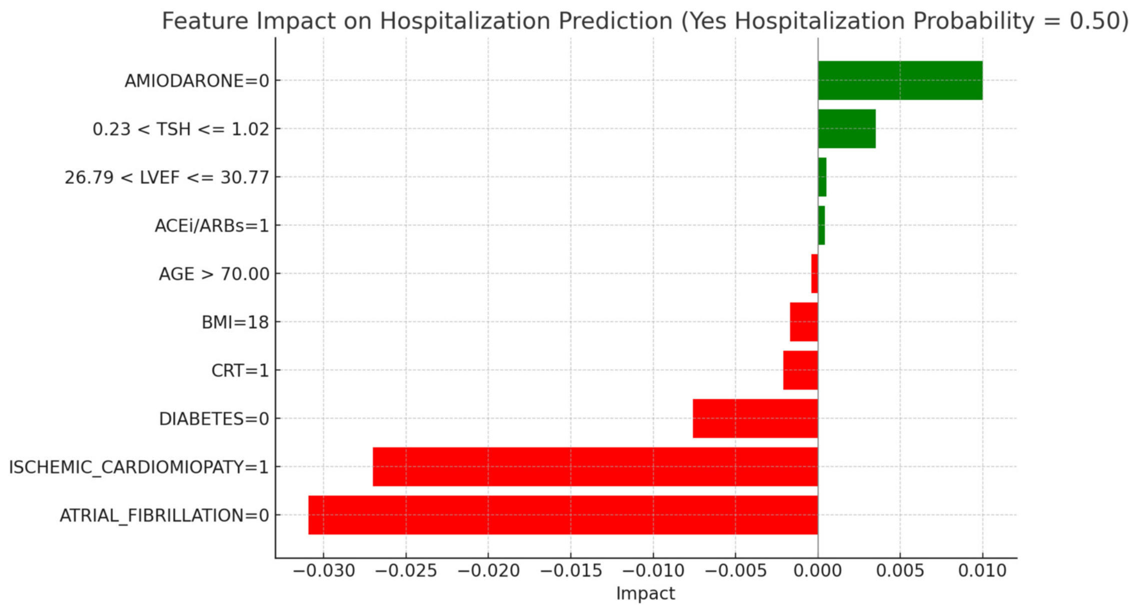


FIGURE 12
LIME explanations for hospitalization prediction model in hyperthyroid patients.

Each scenario provides insight into how the balance of Mortality and Hospitalization risks varies across patient classes, highlighting distinct risk profiles for targeted interventions.

To handle these dual objectives—Mortality and Hospitalization—we use a weighted sum approach. This approach is common in multi-objective optimization, where conflicting objectives must be simultaneously optimized. In our context, each objective is calculated based on the probability of Mortality (p_{Death}) and the probability of Hospitalization (p_{Hosp}), derived from pre-trained ML models. The weighted sum method allows us to combine these objectives into a single metric for easier comparison. The weighted sum method can be represented mathematically as (Equation 16):

$$Combined\ Risk = w1 \cdot Objective1 + w2 \cdot Objective2 \qquad (16)$$

where $w1$ and $w2$ are weights for each objective. In this analysis, we have set $w1 = 0.5$ and $w2 = 0.5$, giving equal importance to both Mortality and Hospitalization. The equal weighting provides a balanced assessment of the risks without favoring one over the other.

The optimization problem is structured around the four scenarios described above. Each scenario is defined by specific objective functions for Mortality and Hospitalization:

- Maximize Mortality & Maximize Hospitalization: $Objective1 = p_{Death}$, $Objective2 = p_{Hosp}$
- Maximize Mortality & Minimize Hospitalization: $Objective1 = p_{Death}$, $Objective2 = 1 - p_{Hosp}$
- Minimize Mortality & Maximize Hospitalization: $Objective1 = 1 - p_{Death}$, $Objective2 = p_{Hosp}$
- Minimize Mortality & Minimize Hospitalization: $Objective1 = 1 - p_{Death}$, $Objective2 = 1 - p_{Hosp}$

The predicted probabilities (p_{Death} and p_{Hosp}) are derived from pre-trained ML models, such as Random Forest, which estimate the likelihood of Mortality and Hospitalization for each patient. These formulations enable the analysis of specific combinations of high and low risks, tailoring the optimization to address varying clinical priorities and patient profiles. By utilizing these probabilities in the optimization framework, we ensure that the risk stratification process is directly linked to model outputs, providing actionable insights that align with predicted patient outcomes.

The optimization is performed for each thyroid class, and the results are summarized by calculating representative points—average values of Follow-up for Mortality (Mortality_FU) and Follow-up for Hospitalization (Hospi_FU). For each thyroid class and scenario, we compute the mean Hospi_FU and Mortality_FU values, which summarize the overall risk level under the specified conditions. These average values serve as the basis for comparison in the subsequent heatmap analysis. To create a single, interpretable measure of risk, we calculate a Combined Risk Score by averaging the Mortality_FU and Hospi_FU scores, as (Equation 17):

$$Combined\ Risk = w1 \cdot Death\_FU + w2 \cdot Hospi\_FU \qquad (17)$$

where $w1 = 0.5$ and $w2 = 0.5$. This balanced weighting helps identify thyroid classes and scenarios with higher overall risk, simplifying the complex multi-objective results into a single metric. We assigned equal weights ($w1 = w2 = 0.5$) to combine mortality and hospitalization risks, ensuring a balanced approach that reflects the clinical importance of both factors. Mortality represents the most severe outcome, while hospitalization significantly impacts quality of life and healthcare costs. By using identical weights, we ensure an unbiased analysis, avoiding distortions and providing an easily interpretable combined risk score. This exploratory approach, aligned with the experimental nature of the study, provides a robust foundation for future research that could explore customized weights based on emerging clinical priorities. Finally, the combined risk is normalized into a percentage for easier interpretation, as (Equation 18):

$$\textit{Combined Risk}\ (\%) = \textit{Combined Risk x}\ 100 \qquad (18)$$

The final output of this analysis is a heatmap representing the Combined Risk Levels across thyroid classes and scenarios, as shown in Figure 13. Each cell in the heatmap corresponds to a thyroid class-scenario combination, with color intensity indicating the level of combined risk. Darker colors represent higher combined risk scores, highlighting groups with elevated risks for Mortality and/or Hospitalization. The heatmap is generated as follows:

- Data Preparation: The representative points (mean Hospi_FU and Mortality_FU values) are organized into a pivot table with thyroid classes as rows and scenarios as columns. The Combined Risk Score is calculated for each combination.
- Heatmap Visualization: Using *seaborn*, we create a heatmap where each cell is colored according to the Combined Risk Score. Annotations show the exact risk level within each cell, and a color bar to the side provides a legend for interpreting the colors.

The heatmap provides an intuitive visualization of risk distribution across thyroid classes and scenarios:

- High-Risk Cells: Dark red cells indicate thyroid classes and scenarios with higher combined risks. For example, LT3 in the Max-Mortality & Max-Hospitalization scenario shows high risk, suggesting a need for close monitoring in this subgroup.
- Moderate-Risk Cells: Cells with medium color intensity represent scenarios with balanced risks. Hypothyroid and Hyperthyroid classes in the Max-Mortality & Min-Hospitalization and Min-Mortality & Max-Hospitalization scenarios display moderate risk, which may require tailored interventions.
- Low-Risk Cells: Blue cells, particularly in the Min-Mortality & Min-Hospitalization scenario, show the lowest combined risk. These groups may require less intensive follow-up.

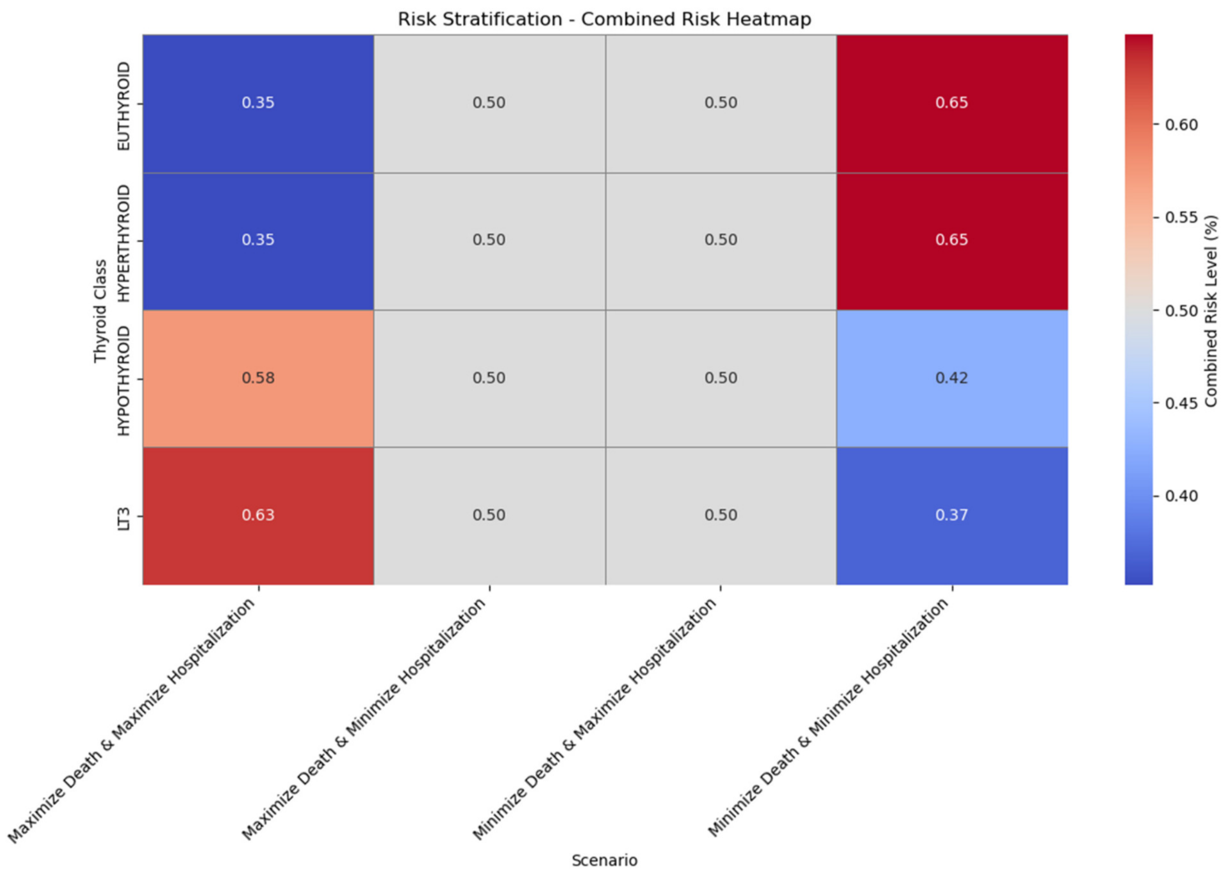


FIGURE 13
Risk stratification—combined risk heatmap across thyroid classes and scenarios.

The Figure 12 heatmap offers an intuitive visualization of risk distribution, highlighting clear differences between thyroid classes and optimization scenarios. This stratification serves as a basis for personalized clinical decision-making, identifying high-priority groups for intervention.

The analysis of combined risk levels across thyroid classes and scenarios reveals notable variations in risk profiles based on different optimization configurations. For the Euthyroid class, the combined risk is 0.35 when both mortality and hospitalization risks are maximized, indicating that Euthyroid patients exhibit a relatively low level of risk even under high-risk conditions for both factors. When mortality risk is maximized and hospitalization risk minimized, the combined risk rises to 0.50, suggesting a moderate risk level. Similarly, the combined risk remains at 0.50 when mortality risk is minimized and hospitalization risk maximized, indicating that reducing the mortality risk while maintaining high hospitalization risk does not significantly change the overall risk level. Surprisingly, when both risks are minimized, the combined risk increases to 0.65, suggesting that reducing both risks may increase the overall risk profile for Euthyroid patients.

For the Hyperthyroid class, the pattern of combined risk closely mirrors that of the Euthyroid class. With the maximization of both risks, the combined risk is also 0.35, suggesting that Hyperthyroid patients, like Euthyroid patients, maintain a relatively low risk level even under high-risk conditions. When mortality risk is maximized and hospitalization minimized, the combined risk reaches 0.50, a moderate level identical to that of the Euthyroid class. The same combined risk level of 0.50 is observed when mortality risk is minimized and hospitalization maximized. However, when both risks are minimized, the combined risk increases to 0.65, the highest value for this class, indicating a significant rise in overall risk under these conditions.

The Hypothyroid class demonstrates a distinct risk profile. When both mortality and hospitalization risks are maximized, the combined risk reaches 0.58, the highest observed so far, suggesting that for Hypothyroid patients, maximizing both risks considerably increases the overall risk level. In the scenario where mortality risk is maximized and hospitalization minimized, the combined risk reduces to a moderate level of 0.50, which remains unchanged even when mortality risk is minimized and hospitalization risk maximized. However, in a context where both risks are minimized, the combined risk further drops to 0.42, indicating that minimizing both risks has a more pronounced risk-reducing effect for the Hypothyroid class compared to high-risk conditions.

Finally, for the LT3 class, the maximization of both mortality and hospitalization risks results in the highest combined risk of all classes, at 0.63. This finding suggests that LT3 patients are particularly vulnerable in conditions of high mortality and hospitalization risk. When mortality risk is maximized and hospitalization minimized, the combined risk drops to 0.50, representing a moderate risk level consistent with other classes in this scenario. Similarly, when mortality risk is minimized and hospitalization maximized, the combined risk remains stable at 0.50. However, when both risks are minimized, the combined risk falls to the lowest level observed at 0.37, indicating that reducing both risks is associated with a very low overall risk level for the LT3 class.

These findings, illustrated in Figure 4, clearly demonstrate how combined risk levels vary across thyroid classes and scenarios. The Euthyroid and Hyperthyroid classes maintain relatively low risk levels across scenarios, while the Hypothyroid and LT3 classes show greater sensitivity to changes in risk scenarios, with higher combined risk levels in specific configurations of risk maximization or minimization. This analysis provides valuable insights for tailored interventions based on the unique risk profiles of each thyroid class.

4.5 Implications, limits and future perspectives

The ML models developed in this study offer significant potential to improve the clinical management of patients with HF and thyroid dysfunctions. By accurately identifying individuals at high risk of mortality and hospitalization, these models enable targeted interventions and personalized treatment strategies. For instance, the early identification of hypothyroid patients with a high likelihood of adverse events could lead to more frequent monitoring, adjustments in pharmacological therapy. Additionally, the interpretation of model outcomes using LIME provides valuable insights to guide clinical decision-making. By highlighting the specific factors contributing to a patient's individual risk, LIME allows clinicians to tailor treatment plans and focus interventions on areas of particular concern.

It is important to acknowledge the limitations of this study to properly interpret the results and guide future research. Although the ML-based approach has shown promising results, the generalizability of the models must be further assessed in larger and more diverse patient populations. The study was retrospective in nature, which introduces potential biases and limits the ability to establish causal relationships. Specifically, there is an inherent risk of selection bias, as patients were not randomly assigned, and the dataset reflects a single-center population with specific inclusion criteria. Information bias and residual confounding may also be present, despite efforts to include a comprehensive set of clinical variables and ensure complete case analysis. Moreover, since the data were not originally collected for predictive modeling purposes, the retrospective design may have introduced selection and information bias. Although only 0.2% of missing values were handled using model-based imputation—which is methodologically appropriate for such low levels of missingness—this approach could still introduce subtle distortions and affect model interpretability, particularly for clinically sensitive variables such as NT-proBNP or thyroid hormones, which may influence risk classification thresholds. These potential biases, related both to the study design and data handling procedures, should be carefully considered when interpreting the results. While the dataset was sizable and well-characterized, these limitations must be considered when interpreting the results. Furthermore, while the statistical analysis included comparisons across multiple variables and subgroups, no formal correction for multiple comparisons was applied. This may increase the risk of Type I error, particularly in exploratory analyses. Therefore, the results should be interpreted with appropriate caution. Future research should incorporate statistical correction techniques—such as Bonferroni or false discovery rate (FDR) adjustments—especially in studies involving formal hypothesis testing across large variable sets. In this study, missing data (accounting for only 0.2% of the dataset) were handled using model-based imputation with a simple decision tree, implemented via the "Impute" widget in Orange. While this approach ensures consistent and reliable estimation of missing values and minimizes information loss, we acknowledge that even low-level imputation may introduce subtle biases or influence model transparency. Future studies should consider comparing multiple imputation techniques to evaluate their impact on the reliability and interpretability of predictive models. Therefore, prospective and multicenter studies with external validation cohorts are strongly recommended to confirm the generalizability and clinical applicability of the proposed models. In this study, the dataset was split into a training set (70%) and a validation set (30%) using the *train_test_split* function from Python's *sklearn* library, with the aim of assessing model performance on unseen internal data and minimizing the risk of overfitting. Additionally, all models were subjected to 10-fold cross-validation to ensure internal consistency and robustness. While these approaches provide strong internal validation, they do not replace the use of independent external datasets. The absence of external validation limits the ability to assess the reproducibility of the model across different populations and healthcare settings. Future research should incorporate external, multicenter cohorts to confirm the generalizability and clinical utility of the proposed framework. Testing the model on broader and more clinically diverse populations will be essential to validate its real-world applicability and ensure its effectiveness in routine clinical practice. Moreover, the lack of prospective validation in the current study represents a significant limitation that further restricts the generalizability of the findings. Although cross-validation and internal testing were rigorously applied, these do not replace the need for validation in real-world, forward-looking clinical environments. Future research should prioritize prospective study designs to verify the model's robustness across diverse patient populations and clinical workflows. While the sample size was substantial, it may not be sufficient to capture the full range of complex interactions between HF and thyroid dysfunctions. Moreover, the demographic composition of the dataset reflects a predominance of male patients (78%), which may introduce gender bias into the model's predictions. This imbalance limits the ability to draw sex-specific conclusions and could impact the model's performance in female subpopulations. Future studies should aim to recruit gender-balanced cohorts to ensure the fairness and representativeness of AI-based risk stratification tools. Additionally, some clinically and socially significant variables—such as medication adherence, health literacy, and socioeconomic status—were not included in the model due to their absence from the structured electronic health records used in this retrospective study. The lack of these variables may limit the completeness and equity of the risk predictions. Future research should prioritize the integration of behavioral and contextual factors to develop more comprehensive and socially aware AI models that better reflect real-world complexities. Further studies in larger, ideally prospective, cohorts would strengthen the study's conclusions and validate its clinical application.

The insights derived from this study pave the way for promising directions in future research. Exploring the integration of additional clinical variables, such as genetic markers and advanced imaging data, could further enhance the predictive accuracy of the models. Incorporating these multidimensional factors could lead to a more comprehensive risk stratification and more precise personalized medicine. Developing ML models capable of predicting not only mortality and hospitalization but also other important patient outcomes, such as quality of life and disease progression, would improve the clinical value of these tools. Additionally, investigating the role of different ML algorithms and optimization techniques could lead to more robust and efficient models. Furthermore, it is essential to study the impact of targeted interventions guided by ML models on patient outcomes. Conducting randomized clinical trials to evaluate the effectiveness of personalized treatment strategies based on model predictions would provide definitive evidence of their clinical benefit. Finally, translating these research findings into practical and accessible clinical tools is essential to realize their full potential. Developing intuitive interfaces and integrating ML models into electronic health record systems would facilitate their widespread adoption and improve patient care. To promote clinical integration, the proposed model could be embedded into electronic health record (EHR) systems as a decision support tool. For example, automatically generated risk scores could trigger alerts for clinicians, prompting earlier intervention or closer monitoring of high-risk patients with thyroid dysfunction and HF. Moreover, the use of interpretable AI techniques such as LIME can help clinicians understand and trust the model's outputs, enhancing transparency and supporting more personalized treatment decisions.

To ensure real-world applicability, future studies should focus on prospective validation using independent and multicenter patient cohorts. This process should involve: (1) recruiting representative populations across different clinical sites; (2) integrating the model into electronic health record systems for real-time risk assessment; (3) comparing clinical decision-making and outcomes with and without model support; and (4) conducting prospective, pragmatic trials to assess the effectiveness of AI-assisted care in routine clinical workflows.

In conclusion, this study demonstrates the immense potential of ML in predicting the risk of mortality and hospitalization in patients with HF and thyroid dysfunctions. AI and ML are increasingly emerging as promising tools to support clinical decision-making and personalize therapeutic pathways, offering new perspectives in the integrated management of cardiovascular and endocrine comorbidities (25). By recognizing the limitations and pursuing future research directions, this field is poised to advance our understanding of this complex interaction and to guide personalized treatment strategies to improve patient outcomes.

5 Conclusions

This study highlights ML as a promising tool to enhance risk stratification and treatment personalization for patients with HF and thyroid dysfunctions. Leveraging a comprehensive set of clinical data, the study demonstrates that ML models, particularly the Random Forest algorithm, can accurately predict mortality and hospitalization risk in this patient population.

The good discriminative ability, evidenced by AUC values for mortality prediction (0.797) and hospitalization risk (0.786), underscores the effectiveness of the Random Forest model in distinguishing between high- and low-risk patients. The model's robust performance, evaluated through metrics such as accuracy, precision, recall, and F1 score, further reinforces its reliability for clinical decision support.

Model interpretation using LIME provides valuable insights into the factors contributing to an individual patient's risk. This information enables targeted interventions and personalized treatment strategies, tailored to the specific needs of each patient. For instance, identifying high-risk patients with clinical characteristics, such as the presence of atrial fibrillation or the absence of amiodarone therapy, could lead to more frequent monitoring, adjustments in pharmacological therapy, and careful consideration of interventions such as CRT.

The study analyzed 762 patients, divided into subgroups based on the presence or absence of thyroid dysfunctions. The results revealed significant clinical differences between groups, with LT3 and hypothyroid patients showing a higher risk of atrial fibrillation and elevated levels of NT-proBNP, an indicator of HF severity. These differences underscore the importance of considering thyroid status in risk assessment and treatment planning for patients with HF.

The risk stratification analysis, using a multi-objective optimization strategy, provided additional insights into the risk profiles of different thyroid classes. Hypothyroid and LT3 patients exhibited a higher combined risk in scenarios where both mortality and hospitalization risk were maximized, highlighting their vulnerability under high-risk conditions.

However, the study has certain limitations. Its retrospective nature introduces potential biases, and the generalizability of the findings should be assessed in larger, more diverse patient cohorts. Further prospective studies are needed to validate the study's findings and clinical applicability.

Despite these limitations, the study represents a significant step forward in applying ML to improve care for patients with HF and thyroid dysfunctions. Integrating additional clinical variables, such as genetic markers and advanced imaging data, could further enhance the predictive accuracy of these models. Exploring different ML algorithms and optimization techniques may lead to more robust and efficient models.

In conclusion, this study demonstrates the potential of ML in transforming the management of patients with HF and thyroid dysfunctions. By leveraging ML, clinicians can gain a deeper understanding of individual risk profiles, enabling targeted interventions and personalized treatment strategies to improve patient outcomes and promote more effective healthcare delivery.

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 humans were approved by Department of Medicine—University of Foggia. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. 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

MI: Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing. VS: Data curation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. AP: Writing – review & editing, Writing – original draft. AM: Formal analysis, Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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. De Luca M, D'Assante R, Iacoviello M, Triggiani V, Rengo G, De Giorgi A, et al. Subclinical hypothyroidism predicts outcome in heart failure: insights from the T.O.S.CA. Registry. *Intern Emerg Med.* (2024) 19(6):1667–74. doi: 10.1007/s11739-024-03665-w

2. Xu Y, Derakhshan A, Hysaj O, Wildisen L, Ittermann T, Pingitore A, et al. The optimal healthy ranges of thyroid function defined by the risk of cardiovascular disease and mortality: systematic review and individual participant data meta-analysis. *Lancet Diabetes Endocrinol.* (2023) 11(10):743–54. doi: 10.1016/S2213-8587(23)00227-9

3. Yang G, Wang Y, Ma A, Wang T. Subclinical thyroid dysfunction is associated with adverse prognosis in heart failure patients with reduced ejection fraction. *BMC Cardiovasc Disord.* (2019) 19(1):83. doi: 10.1186/s12872-019-1055-x

4. Tian J, Yan J, Han G, Du Y, Hu X, He Z, et al. Machine learning prognosis model based on patient-reported outcomes for chronic heart failure patients after discharge. *Health Qual Life Outcomes.* (2023) 21(1):31. doi: 10.1186/s12955-023-02109-x

5. Marengo A, Pagano A, Santamato V. An efficient cardiovascular disease prediction model through AI-driven IoT technology. *Comput Biol Med.* (2024) 183:109330. doi: 10.1016/j.compbiomed.2024.109330

6. Nakamura K, Zhou X, Sahara N, Toyoda Y, Enomoto Y, Hara H, et al. Risk of mortality prediction involving time-varying covariates for patients with heart failure using deep learning. *Diagnostics.* (2022) 12(12):2947. doi: 10.3390/diagnostics12122947

7. Zhou X, Nakamura K, Sahara N, Asami M, Toyoda Y, Enomoto Y, et al. Exploring and identifying prognostic phenotypes of patients with heart failure guided by explainable machine learning. *Life.* (2022) 12(6):776. doi: 10.3390/life12060776

8. Triggiani V, Iacoviello M, Monzani F, Puzzovivo A, Guida P, Forleo C, et al. Incidence and prevalence of hypothyroidism in patients affected by chronic heart failure: role of amiodarone. *Endocr Metab Immune Disord Drug Targets.* (2012) 12(1):86–94. doi: 10.2174/187153012799278947

9. Iacoviello M, Parisi G, Gioia MI, Grande D, Rizzo C, Guida P, et al. Thyroid disorders and prognosis in chronic heart failure: a long-term follow-up study. *Endocr Metab Immune Disord Drug Targets.* (2020) 20(3):437–45. doi: 10.2174/1871530319666191018134524

10. Terlizzese P, Albanese M, Grande D, Parisi G, Gioia MI, Brunetti ND, et al. TSH Variations in chronic heart failure outpatients: clinical correlates and outcomes. *Endocr Metab Immune Disord Drug Targets.* (2021) 21(10):1935–42. doi: 10.2174/1871530321666210430131510

11. Garg P, Wood S, Swift AJ, Fent G, Lewis N, Rogers D, et al. Clinical predictors of all-cause mortality in patients presenting to specialist heart failure clinic with raised NT-proBNP and no heart failure. *ESC Heart Fail.* (2020) 7(4):1791–800. doi: 10.1002/ehf2.12742

12. Behnoush AH, Shariatnia MM, Khalaji A, Asadi M, Yaghoobi A, Rezaee M, et al. Predictive modeling for acute kidney injury after percutaneous coronary intervention in patients with acute coronary syndrome: a machine learning approach. *Eur J Med Res.* (2024) 29(1):76. doi: 10.1186/s40001-024-01675-0

13. Li D, Fu J, Zhao J, Qin J, Zhang L. A deep learning system for heart failure mortality prediction. *PLoS One.* (2023) 18(2):e0276835. doi: 10.1371/journal.pone.0276835

14. Sibilia B, Toupin S, Dillinger JG, Brette JB, Ramonatxo A, Schurtz G, et al. Machine learning to predict in-hospital outcomes in patients with acute heart failure. *Eur Heart J.* (2023) 44(Supplement_2):ehad655.1102. doi: 10.1093/eurheartj/ehad655.1102

15. Danieli MG, Brunetto S, Gammeri L, Palmeri D, Claudi I, Shoenfeld Y, et al. Machine learning application in autoimmune diseases: state of art and future prospectives. *Autoimmun Rev.* (2024) 23(2):103496. doi: 10.1016/j.autrev.2023.103496

16. Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, et al. An application of machine learning to etiological diagnosis of secondary hypertension: retrospective study using electronic medical records. *JMIR Med Inform.* (2021) 9(1):e19739. doi: 10.2196/19739

17. Bucholc M, Bradley D, Bennett D, Patterson L, Spiers R, Gibson D, et al. Identifying pre-existing conditions and multimorbidity patterns associated with in-hospital mortality in patients with COVID-19. *Sci Rep.* (2022) 12(1):17313. doi: 10.1038/s41598-022-20176-w

18. Ponomartseva DA, Derevitskii IV, Kovalchuk SV, Babenko AY. Prediction model for thyrotoxic atrial fibrillation: a retrospective study. *BMC Endocr Disord.* (2021) 21(1):150. doi: 10.1186/s12902-021-00809-3

19. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail.* (2020) 8(1):12–21. doi: 10.1016/j.jchf.2019.06.013

20. Inoue K, Ritz B, Brent GA, Ebrahimi R, Rhee CM, Leung AM. Association of subclinical hypothyroidism and cardiovascular disease with mortality. *JAMA Netw Open.* (2020) 3(2):e1920745. doi: 10.1001/jamanetworkopen.2019.20745

21. Segar MW, Hall JL, Jhund PS, Powell-Wiley TM, Morris AA, Kao D, et al. Machine learning–based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure. *JAMA Cardiol.* (2022) 7(8):844–54. doi: 10.1001/jamacardio.2022.1900

22. Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. *Risk prediction of cardiovascular disease using machine learning classifiers. Open Med.* (2022) 17(1):1100–13. doi: 10.1515/med-2022-0508

23. Sinha I, Aluthge DP, Chen ES, Sarkar IN, Ahn SH. Machine learning offers exciting potential for predicting postprocedural outcomes: a framework for developing random forest models in IR. *J Vasc Interv Radiol.* (2020) 31(6):1018–24.e4. doi: 10.1016/j.jvir.2019.11.030

24. Santamato V, Esposito D, Tricase C, Faccilongo N, Marengo A, Pange J. Assessment of public health performance in relation to hospital energy demand, socio-economic efficiency and quality of services: an Italian case study. In: Gervasi O, Murgante B, Rocha AMAC, Garau C, Scorza F, Karaca Y, et al., editors. *Computational Science and Its Applications—ICCSA 2023 Workshops*; 2023 Jul 3–6; Athens, Greece. Cham: Springer Nature Switzerland (2023). pp. 505–22. doi: 10.1007/978-3-031-37111-0_35

25. Santamato V, Tricase C, Faccilongo N, Iacoviello M, Marengo A. Exploring the impact of artificial intelligence on healthcare management: a combined systematic review and machine-learning approach. *Appl Sci.* (2024) 14(22):10144. doi: 10.3390/app142210144

26. Santamato V, Tricase C, Faccilongo N, Iacoviello M, Pange J, Marengo A. Machine learning for evaluating hospital mobility: an Italian case study. *Appl Sci.* (2024) 14(14):6016. doi: 10.3390/app14146016

27. Santamato V, Tricase C, Faccilongo N, Marengo A, Pange J. Healthcare performance analytics based on the novel PDA methodology for assessment of efficiency and perceived quality outcomes: a machine learning approach. *Expert Syst Appl.* (2024) 252:124020. doi: 10.1016/j.eswa.2024.124020

28. Toki EI, Tsoulos IG, Santamato V, Pange J. Machine learning for predicting neurodevelopmental disorders in children. *Appl Sci.* (2024) 14(2):837. doi: 10.3390/app14020837

29. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol.* (2017) 69(21):2657–64. doi: 10.1016/j.jacc.2017.03.571

30. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med.* (2009) 150(9):604–12. doi: 10.7326/0003-4819-150-9-200905050-00006

31. Braune K, Boss K, Schmidt-Herzel J, Gajewska KA, Thieffry A, Schulze L, et al. Shaping workflows in digital and remote diabetes care during the COVID-19 pandemic via service design: prospective, longitudinal, open-label feasibility trial. *JMIR Mhealth Uhealth.* (2021) 9(4):e24374. doi: 10.2196/24374

32. Shamshirband S, Fathi M, Dehzangi A, Chronopoulos AT, Alinejad-Rokny H. A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. *J Biomed Inform.* (2021) 113:103627. doi: 10.1016/j.jbi.2020.103627

33. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med Inform Decis Mak.* (2020) 20(11):295. doi: 10.1186/s12911-020-01318-4

34. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One.* (2016) 11(5):e0155705. doi: 10.1371/journal.pone.0155705

35. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Acad Pathol.* (2019) 6:2374289519873088. doi: 10.1177/2374289519873088

TYPE Original Research
PUBLISHED 03 November 2025
DOI 10.3389/fcvm.2025.1652467

Machine learning-driven exploration of therapeutic targets for atrial fibrillation-joint analysis of single-cell and bulk transcriptomes and experimental validation

Yicheng Wang[1,2,3†], Hong-Yi Yang[1,2,3†], Zi-Ao Fan[1,2,3†] and Jian-Cheng Zhang[1,2,3*]

[1]Shengli Clinical Medicine College of Fujian Medical University, Fuzhou, Fujian, China, [2]Fuzhou University Affiliated Provincial Hospital, Fuzhou, Fujian, China, [3]Department of Cardiology, Fujian Provincial Hospital, Fuzhou, Fujian, China

OPEN ACCESS

EDITED BY
Ulrich Parlitz,
Max Planck Institute for Dynamics and Self-Organization, Germany

REVIEWED BY
Carlo Lavalle,
Sapienza University of Rome, Italy
Xianglin Long,
Central South University, China

*CORRESPONDENCE
Jian-Cheng Zhang
✉ fjzhangjiancheng@126.com

†These authors have contributed equally to this work

RECEIVED 23 June 2025
ACCEPTED 16 October 2025
PUBLISHED 03 November 2025

CITATION
Wang Y, Yang H-Y, Fan Z-A and Zhang J-C (2025) Machine learning-driven exploration of therapeutic targets for atrial fibrillation-joint analysis of single-cell and bulk transcriptomes and experimental validation.
Front. Cardiovasc. Med. 12:1652467.
doi: 10.3389/fcvm.2025.1652467

Background: To explore new therapeutic targets and strategies for atrial fibrillation (AF) by analyzing gene expression profiles of AF patients using machine learning techniques combined with transcriptomic data, and to uncover the potential molecular mechanisms underlying AF.

Methods: Transcriptomic datasets associated with AF were obtained from the GEO database. After batch effect removal and normalization, differential gene expression analysis was performed to identify differentially expressed genes (DEGs). Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Disease Ontology (DO) enrichment analyses were conducted to explore the functions and pathways of these DEGs. Three machine learning algorithms, Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine—Recursive Feature Elimination (SVM-RFE), and random forest (RF), were applied to screen key genes related to AF. A nomogram model was developed based on the identified key genes, and its diagnostic performance was evaluated. Single-cell transcriptome analysis was performed to investigate the cell-type-specific expression patterns of these key genes. Finally, Real-time PCR (RT-qPCR) and western blot (WB) analyses was performed on right auricular tissue from patients with atrial fibrillation and control samples.

Results: A total of 64 DEGs were identified, including 27 upregulated and 37 downregulated genes. Enrichment analyses revealed that these genes were involved in biological processes such as positive regulation of muscular systemic processes, immune responses, and calcium signaling pathways. Three machine learning algorithms identified six key genes for AF. The nomogram model based on these six genes demonstrated excellent diagnostic performance with an AUC of 0.97. Single-cell transcriptome analysis showed specific expression patterns of these key genes in different cell types. Additionally, immune infiltration analysis indicated changes in the immune microenvironment in AF patients. qPCR and WB analyses also indicated that the differences in mRNA and protein expression levels of these six molecules between the control group and the atrial fibrillation group were consistent with the results of transcriptome analysis.

Conclusion: This study provides new insights into the molecular mechanisms of AF and offers potential non-invasive biomarkers for AF diagnosis. The identified key genes and constructed model may facilitate the development of targeted therapies for AF.

KEYWORDS

atrial fibrillation, machine learning, single-cell, bulk transcriptomes, therapeutic targets, network physiology

Introduction

Atrial fibrillation is one of the most common sustained arrhythmias in clinical practice, with its prevalence showing a steady upward trend (1, 2). AF not only significantly impairs patients' quality of life but also markedly increases the risk of severe complications such as stroke and heart failure, imposing a substantial economic burden on patients, families, and society (3).

The pathophysiology of AF involves multiple processes, including cardiac electrophysiological remodeling, structural remodeling, aberrant neural regulation, and inflammatory responses (4, 5). Interactions among ion channel dysfunction in atrial myocytes, alterations in intercellular connexins, progression of myocardial fibrosis, and autonomic nervous system imbalance collectively contribute to the initiation and maintenance of AF (6). However, the understanding of these mechanisms remains incomplete, which limits the development of targeted therapeutic strategies (7).

Current treatment options for AF primarily include pharmacological therapy, catheter ablation, and surgical intervention (8, 9). While pharmacological therapy is effective in controlling ventricular rate and preventing thromboembolism, long-term use is often associated with adverse effects, and some patients exhibit poor responsiveness to medication (10). Catheter ablation, as a curative approach, has limited success rates and carries a risk of recurrence. Surgical treatment, being highly invasive, is applicable only to specific patient populations (11). Overall, existing therapies fail to fully meet the clinical needs of AF patients, highlighting the urgent need to explore novel therapeutic targets and strategies (12).

The rapid advancement of high-throughput omics technologies has enabled comprehensive systemic analysis of biological samples, thereby uncovering disease-related molecular signatures and potential mechanisms (13, 14). Transcriptomics, in particular, plays a critical role in elucidating the relationship between gene expression changes and disease progression, providing a rich resource for cardiovascular research (15, 16).

Continuous progress in machine learning and bioinformatics has provided effective tools for processing and interpreting large-scale omics datasets (13, 17–19). Machine learning algorithms can identify patterns, select key features, and construct predictive models from complex datasets, facilitating the discovery of potential biomarkers and therapeutic targets (20–22).

Furthermore, our study aligns with the emerging framework of Network Physiology, which emphasizes the integration of multi-level biological networks to understand complex physiological systems and disease states. In the context of atrial fibrillation, we explore not only gene-level interactions through protein-protein interaction networks but also cell-type-specific expression patterns and immune microenvironment crosstalk, thereby uncovering the network-based mechanisms underlying AF pathogenesis. The application of machine learning further enables the identification of key network hubs that drive AF progression, highlighting the central role of network analysis in bridging molecular features with clinical phenotypes.

This study aims to systematically analyze the gene expression profiles of AF patients using transcriptomic data and machine learning techniques, with the goal of identifying key genes closely associated with AF pathogenesis and therapeutic responses. Through in-depth investigation of these genes, we aim to uncover the potential molecular mechanisms underlying AF.

Materials and methods

Data acquisition

Transcriptomic datasets associated with atrial fibrillation were obtained from the Gene Expression Omnibus (GEO) database. For the discovery phase, we selected three datasets (GSE41177, GSE115574, and GSE79768) based on the following criteria: (1) sample type consisted of human atrial tissue, which is directly relevant to AF pathophysiology; (2) each dataset contained a sufficient number of both AF and control samples to ensure analytical robustness; (3) they were generated using comparable high-throughput platforms (Affymetrix or Illumina) to minimize technical batch effects. Other AF-related datasets in GEO were excluded if they had a small sample size ($n < 5$ per group), were derived from non-cardiac tissues for the discovery analysis, or lacked clear phenotyping. The dataset GSE2240, which is an independent atrial tissue dataset not used in the discovery process, was utilized for external validation of the machine-learning-identified feature genes. Furthermore, the dataset GSE255612, which contains right auricular tissue samples from 18 AF patients and 16 non-AF individuals, was downloaded for subsequent single-cell transcriptomic analysis to explore cell-type-specific expression patterns. The specific distribution of sample sizes for each dataset is shown in Table 1.

Batch effect removal

Before performing the difference analysis, we merged the three AF datasets (GSE41177, GSE115574, GSE79768). We then

TABLE 1 Distribution of sample sizes in each dataset.

Dataset	Platform	Country	Tissue origin	Anatomical location	AF (*n*)	Control (*n*)
GSE41177	GPL570	Taiwan	left atrial appendage	LA free wall	32	6
GSE115574	GPL570	Turkey	left/right atrial appendage	LA/RA free wall	15	15
GSE79768	GPL570	Taiwan	right atrial appendage	LA/RA free wall	13	13
GSE2240	GPL96	Germany	left/right atrial appendage	LA/RA free wall	20	10

corrected for batch effects using the "sva" package of the R language. To assess the effectiveness of this correction, we compared data quality before and after batch removal using principal component analysis (PCA).

Differential expression analysis

Differential gene expression analysis of the sequencing data was performed using the "limmaa" package in R software to compare samples from the control and experimental groups, thereby identifying DEGs. The criteria for screening DEGs were set as |log2FC| > 0.5 and a *P*-value < 0.05. The results of the differential analysis were visualized using the "ggplot2" package to generate volcano plots and heatmaps. The volcano plot clearly illustrates the distribution of DEGs, including upregulated genes, downregulated genes, and genes with no significant difference in expression.

GO and KEGG enrichment analysis

GO annotation from the org.Hs.eg.db package (version 3.1.0) in R software was used as the background. Genes were mapped to this background, and GO analysis was subsequently performed using the clusterProfiler package (version 3.14.3). The GO analysis covered three aspects: biological processes (BP), molecular functions (MF), and cellular components (CC), aiming to detect enriched pathways and thereby reveal the cellular functions, signaling pathways, and disease-related differentially expressed gene pathways primarily affected by the candidate target genes. KEGG was used to annotate gene pathways. Enrichment was considered statistically significant when $P < 0.05$.

DO enrichment analysis

DO enrichment analysis was performed using the org.Hs.eg.db R package (version 3.1.0) to obtain gene annotation information for the gene set. These genes were linked to the DO background dataset to ensure each gene was associated with disease classifications in the DO system. This approach aimed to identify disease processes related to atrial fibrillation treatment responses.

Machine learning algorithm applications

LASSO regression was employed to identify key genes associated with atrial fibrillation. After preprocessing the candidate differentially expressed genes, LASSO regression was implemented using the glmnet function, treating the data as a binary classification problem. The response variable was extracted from sample names using regular expressions. The model was evaluated by plotting the model object and performing cross-validation via cv.glmnet to determine the optimal lambda value. Finally, genes with non-zero coefficients corresponding to the optimal lambda value were identified as key genes related to the disease status of atrial fibrillation and were output. SVM-RFE analysis was conducted using the "e1071", "kernlab" and "caret" packages in R. The number of genes corresponding to the minimized cross-validation error in the analysis results was used to determine the count of potential biomarkers identified by SVM-RFE machine learning. Genes with average rankings corresponding to the SVM-RFE analysis were selected as potential biomarkers for AF. Random forest analysis was performed using the "randomForest" package in R. The importance scores of differentially expressed genes were obtained at the point of minimized error on the cross-validation curve. Genes with importance scores exceeding 1 were selected as potential biomarkers for AF. A venn diagram was used to identify the intersection of genes obtained from LASSO, SVM-RFE, and Random Forest analyses. The final set of potential AF biomarkers was derived from the overlapping genes identified by these three machine learning methods.

Construction of protein-protein interaction (PPI) networks

Protein-protein interaction (PPI) networks were constructed using the GeneMANIA database (http://www.string-db.org/) to explore the regulatory interactions between genes and predict potential regulatory factors. This approach facilitated a deeper understanding of gene relationships and their regulatory mechanisms in the context of atrial fibrillation.

Development and validation of nomogram

The integrated dataset from GSE41177, GSE115574, and GSE79768 (after batch effect correction) was used as the training set to construct the diagnostic model. To ensure a rigorous evaluation and avoid data leakage, the validation process was strictly separated. In this study, the "rms" package in R software was employed to develop a nomogram model for identifying diagnostic genes in AF. Each candidate gene was assigned a specific score, with the total score being the sum of these individual gene scores. The model's performance was first evaluated internally on the training set. To evaluate the model's accuracy, calibration curves were plotted to assess the consistency between predicted probabilities and actual outcomes. Furthermore, decision curve analysis (DCA) was conducted to evaluate the clinical utility of the model. The diagnostic efficacy of six key genes was assessed through receiver operating characteristic (ROC) curve analysis. Finally, the robustness of the model was validated using the independent external validation set (GSE2240), which was not involved in any prior steps of differential expression analysis or machine learning feature selection.

Single-cell transcriptome analysis

In the single-cell RNA-sequencing (scRNA-seq) analysis pipeline, data normalization is first carried out via the LogNormalize method to guarantee the comparability of gene expression levels across different cells. Then, the FindVariableFeatures method is employed to select highly variable genes (top 2,000). To further eliminate batch effects, the Harmony algorithm is applied for batch correction, enhancing the comparability of data from different experimental batches. Subsequently, dimensionality reduction is performed using principal component analysis (PCA). For cell clustering, the non-linear dimensionality reduction method of *t*-distributed stochastic neighbor embedding (t-SNE) is utilized for analysis. Cell grouping is conducted using the FindClusters function, and the clustering results are optimized by adjusting different resolution parameters. The entire quality control workflow comprises steps such as normalization, batch correction, and dimensionality reduction to ensure the accuracy and reliability of data analysis. With the thresholds of $P < 0.05$ and log2FC > 0.25, "FindAllMarkers" is used to identify differentially expressed genes in each cluster. Based on the unique marker genes in the study, the expression of these marker genes in different clusters is analyzed to annotate the cells.

Quantitative Rt-PCR analysis

Total RNA was extracted from cardiac tissue using TRIzol Reagent (Invitrogen, CA, USA), and reverse-transcribed into cDNA via the Novo Protein Reverse Transcription Kit (Suzhou, China). Real-time PCR was performed on a Roche LightCycler® 480 Real-Time PCR Apparatus (Bio-Rad, Basel, Switzerland) to detect the expression of C1orf105, DHRS9, CHGB, PDE8B, CSRP3,

TABLE 2 The sequences of the primers for qPCR.

Gene symbol	Species	Forward primer	Reverse primer
C1orf105	Human	ATTCACTACAGACTGCCCATTCT	CGTTGTCTTGCCTATTGGTTCC
DHRS9	Human	GGCTTTGGAAACTTGGCAGC	TCGGTCACATCCAGAAGCAC
CHGB	Human	GCCAGATCGGAAACACATGC	CGTCGTTTGTCCACCTCAGA
PDE8B	Human	CAAACTCAGAACTTCGATGCAGA	CTTCATGGTCATCCGATACTCG
CSRP3	Human	GTGCTATGGGCGCAGATATGG	CTCGGACTCTCCAAACTTCGC
FCER1G	Human	CTCCAGCCCAAGATGATTCCA	CTTTCGCACTTGGATCTTCAGTC

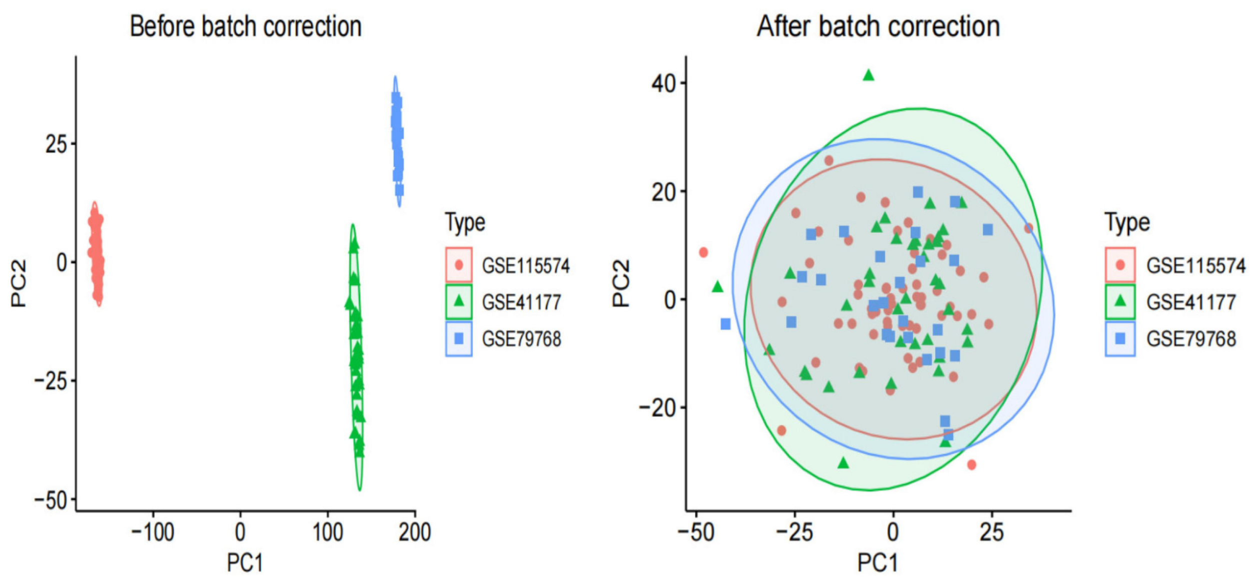


FIGURE 1
PCA of three original AF datasets prior to batch effect correction and PCA of integrated AF dataset after batch effect correction.

FCER1G, and β-actin (as a normalization control). The relative expression levels of these hub genes were calculated using the 2 −ΔΔCT method. Statistical analysis was conducted with GraphPad Prism, and t-tests were applied for two groups of data following a normal distribution. A significance level of $P < 0.05$ was adopted. The primer sequences for C1orf105, DHRS9, CHGB, PDE8B, CSRP3, and FCER1G are listed in Table 2.

Western blot analysis

Total protein was extracted from right auricular tissues of AF patients and non-AF controls using RIPA lysis buffer containing protease and phosphatase inhibitors. Protein concentrations were determined using a BCA Protein Assay Kit (Beyotime, China). Equal amounts of protein (20 µg per lane) were separated by 10% SDS-PAGE and transferred onto PVDF membranes (MeilunBio, China). After blocking with 5% non-fat milk for 1 h at room temperature, the membranes were incubated overnight at 4°C with primary antibodies against C1orf105 (1:2,000, Abmart, PH13497), DHRS9 (1:2,000, immunoway, YN0639), CHGB (1:2,000, immunoway, YT6192), PDE8B (1:2,000, Proteintech, 30708-1-AP), CSRP3 (1:2,000, immunoway, YN6528), FCER1G (1:2,000, Abmart, TD13263), and β-actin (1:10,000, immunoway, YM8343) as a loading control. After washing, the membranes were incubated with HRP-conjugated secondary antibodies (1:5,000, Proteintech) for 1 h at room temperature. Protein bands were visualized using an ECL detection system (Tanon, China). The grayscale values of protein bands were analyzed using ImageJ software (National Institutes of Health, USA), and the relative expression levels were normalized to β-actin. Statistical analysis and graph generation for WB data were performed using GraphPad Prism software (version 9.5, USA).

Statistical analysis

All statistical analyses and gene expression data were processed using R (version 4.4.3). When the data were normally distributed, we compared the two groups using an independent two-sample *t*-test. If the data were not normally distributed, we used the Wilcoxon rank-sum test for intergroup comparisons. A *p*-value of less than 0.05 was set as the threshold for statistical significance.

Results

Identification of differentially expressed genes

Raw AF and control transcriptome data were obtained from the GEO database, integrated after batch effect removal, and normalized to generate 58 AF cases and 65 control treatment cohorts (Figure 1).

Identifying of differentially expressed associated with AF

We performed differential analysis of the AF cohort to reveal differential genes for AF. A total of 64 deg were identified, of which 27 were upregulated and 37 were downregulated (Figure 2).

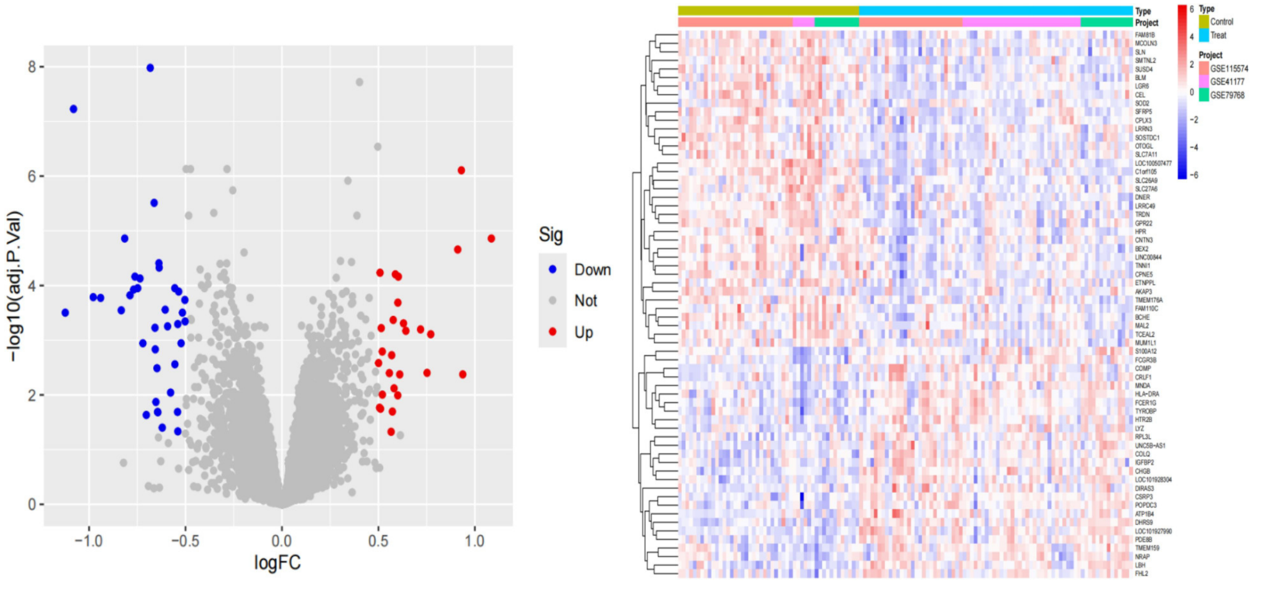


FIGURE 2
Volcano and Heatmap plots depicting DECs between AF and healthy controls.

Functional enrichment analysis of AF differential genes

Differential gene PPI networks were constructed through the GeneMANIA database (Figure 3) and analyzed for functional enrichment using GO, KEGG, and DO to identify potential mechanisms of action.

The results of the enrichment analysis are shown in Figure 4. In the biological process, AF-related DEGs are enriched in positive regulation of muscular systemic processes. This includes positive regulation of muscle cell development; negative regulation of myocardial fiber assembly; regulation of immune response; positive regulation of muscle tissue development; and negative regulation of muscle cell differentiation. For cellular components, these genes are mainly enriched in cellular structures such as myofibers, sarcoplasmic reticulum, nucleus pulposus lumen, autophagosomal membranes, I-bands, Z-discs, and myogenic fibers. For molecular function, these genes are enriched in a variety of molecular binding activities: cytokine binding; immunoglobulin receptor activity; glycosaminoglycan binding; immunoglobulin binding; BMP binding; heparin binding. These functions are involved in the regulation of the heart and the immune system, suggesting that AF may be closely related to the interaction and signaling of these molecules.

KEGG pathway analysis further revealed significant enrichment of AF-related genes in several biological processes. Specifically, pathways such as pancreatic secretion, salivary secretion, and myocardial contraction, which are closely related to the regulation of cardiac function and the digestive system,

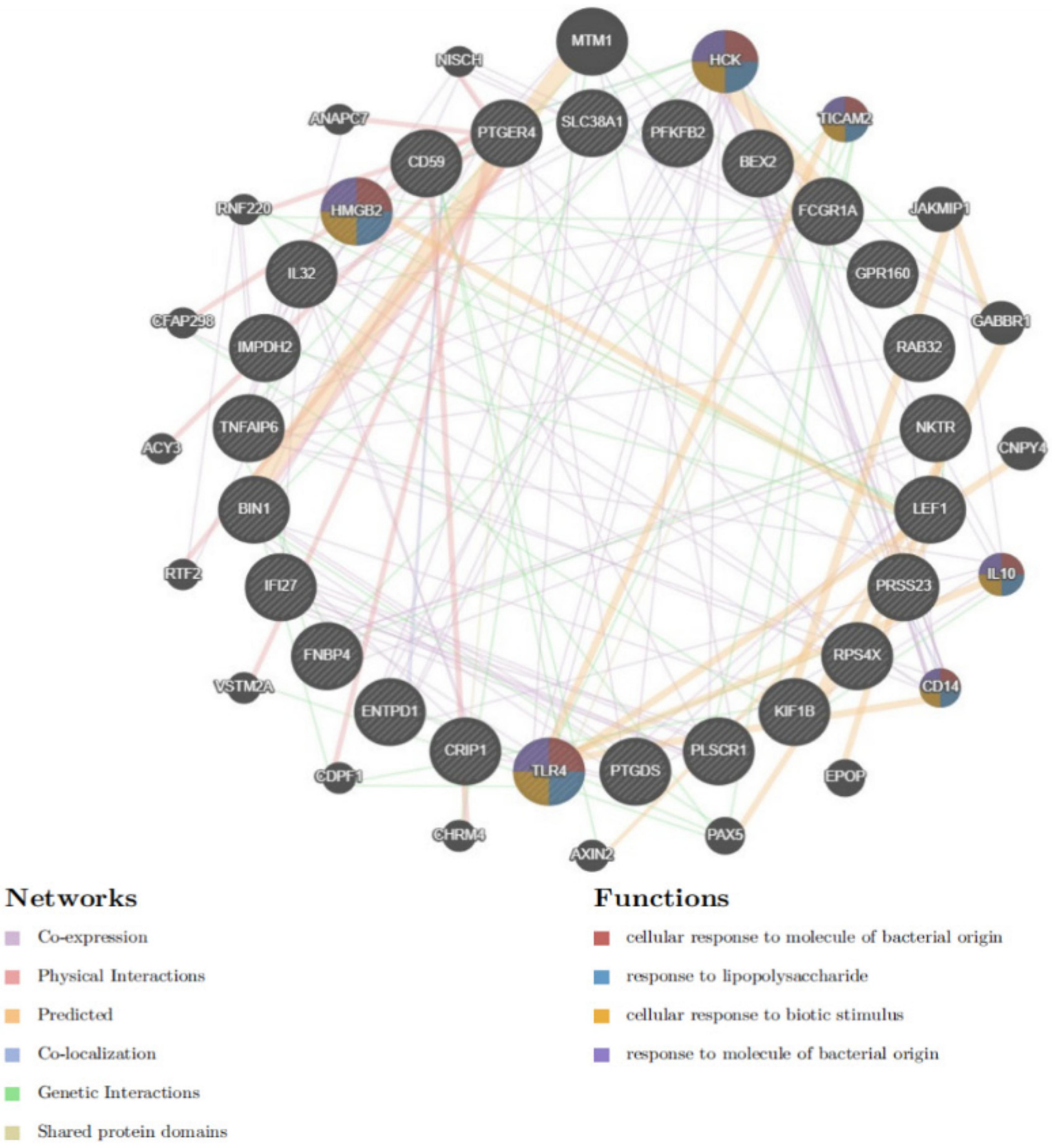


FIGURE 3
PPI network analysis for differential genes.

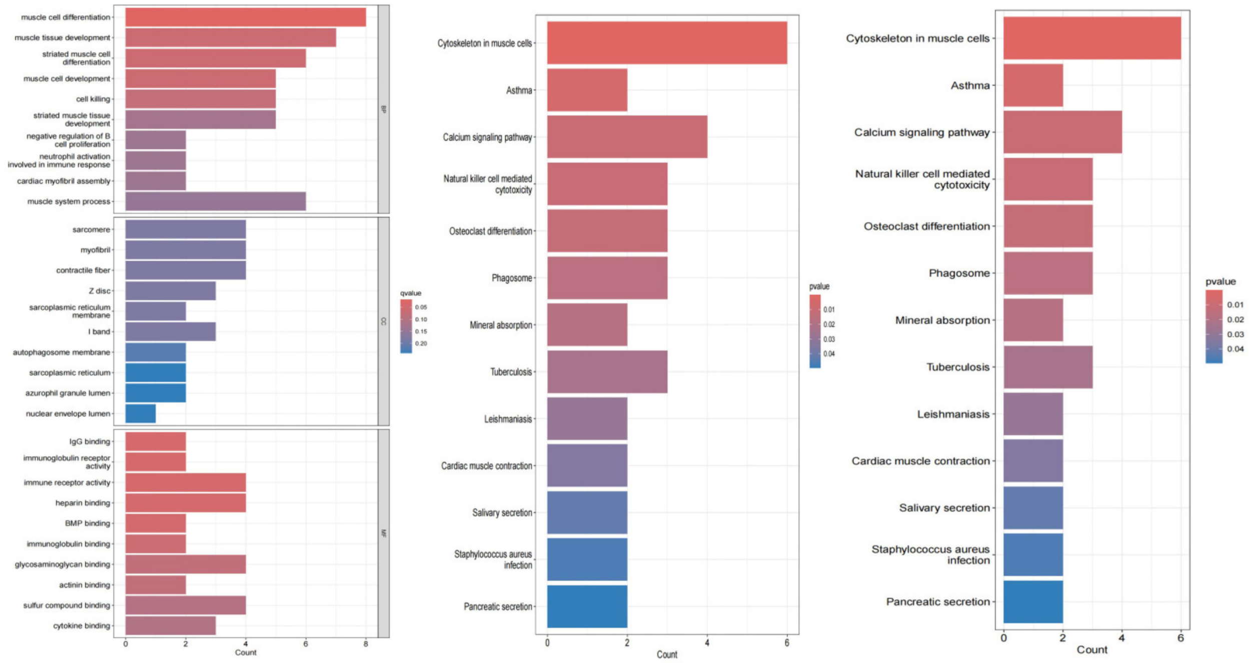


FIGURE 4
Barplots of GO, KEGG, D0 enrichment analysis results.

showed significant enrichment. At the same time, we also observed enrichment of pathways related to infectious diseases such as Staphylococcus aureus infection and tuberculosis, which may be related to the activation of inflammatory responses in patients with atrial fibrillation. In addition, the enrichment of pathways such as mineral uptake and natural killer cell-mediated cytotoxicity suggests possible immune and metabolic mechanisms involved in AF. The significant enrichment of the calcium signaling pathway is particularly noteworthy because this pathway plays a central role in cardiac electrophysiology and contractile function, and its abnormalities may be directly associated with the development of AF. Finally, the enrichment of cytoskeletal pathways in muscle cells emphasizes the importance of cardiac muscle structure and function in AF.

Disease ontology semantic and enrichment analyses revealed significant associations of AF with multiple biological processes. Specifically, AF was significantly associated with processes such as pancreatic secretion, Staphylococcus aureus infection, salivation, myocardial contraction, leishmaniasis, tuberculosis, mineral uptake, phagolysosomes, osteoclast differentiation, natural killer cell-mediated cytotoxicity, calcium signaling pathways, asthma, and cytoskeleton in muscle cells.

Analysis of immune cell infiltration in AF

Single-sample gene set enrichment analysis (ssGSEA) results for atrial fibrillation revealed functions and pathways associated with immune cell subsets. ssGSEA was used to depict the relative abundance of immune cell subsets in the AF cohort. Samples from the AF cohort showed activated B cells, activated CD4+ T cells, activated CD8+ T cells, activated dendritic cells, CD56bright natural killer cells, CD56dim natural killer cells, eosinophils, γ δ T cells, immature B cells, immature dendritic cells, myeloid-derived suppressor cells (MDSC), as compared to controls, macrophages, mast cells, monocytes, natural killer T cells, natural killer cells, neutrophils, plasmacytoid dendritic cells, regulatory T cells, follicular helper T cells, type 1 helper T cells, type 17 helper T cells, and type 2 helper T cells were enriched. The box line plot further demonstrates that the proportions of macrophages, endothelial cells, and activated dendritic cells were elevated in the atrial fibrillation cohort, whereas the abundance of effector memory CD8+ T cells was reduced compared with the control group. These results suggest changes in the immune microenvironment in the AF cohort, particularly in the composition of specific immune cell subsets (Figures 5A,B).

Identification of hub genes *via* machine learning

We used three machine learning algorithms, LASSO, RF, and SVM-RFE, to further screen Hub genes for AF. We identified 24 potential candidate biomarkers by the LASSO algorithm (Figure 6). The RF algorithm ranked the genes based on the importance calculation of each gene, and we selected the top 30 as potential candidates for AF (Figure 7). To establish the optimal number of Hub genes, we selected the top 30 genes for the SVM-RFE algorithm results as candidate genes (Figure 8). By intersecting the results of all three algorithms, we identified six Hub genes for AF: C1orf105, DHRS9, CHGB, PDE8B, CSRP3 and FCER1G. The visualization results were shown in Figure 9.

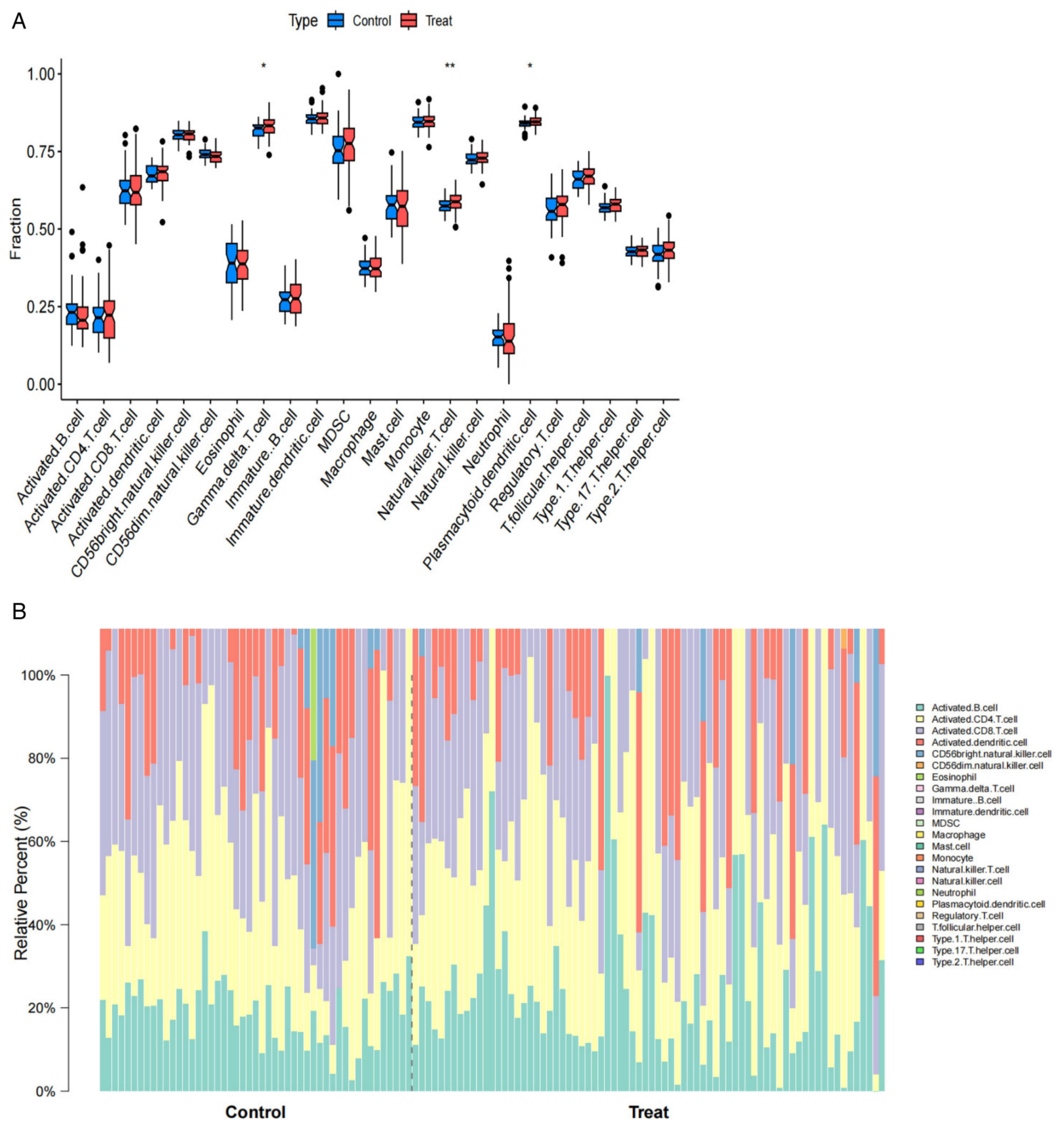


FIGURE 5
(A) Boxplots comparing immune cell abundances between AF vs. controls. (B) Barplots comparing imumme cell abundances between AF vs. controls.

Diagnostic value assessment

We constructed a nomogram model based on the six gene signature. This model demonstrated excellent diagnostic performance, with an AUC of 0.97. Calibration curves validated its accurate predictive capacity for AF. Moreover, DCA results confirmed the clinical applicability of the nomogram model. Collectively, these findings indicate that the nomogram model exhibits robust predictive performance (Figure 10). Additionally, we generated a differential expression box plot of the Hub gene. Finally, we validated the hub genes in GSE2240 by ROC curve analysis. The differential expression results showed that the expression of DHRS9, CHGB, PDE8B, and CSRP3 was up-regulated, and the expression of FCER1G and C1orf105 was down-regulated compared to the control (Figure 11). In the external validation set (Figure 12), the expression of DHRS9, CHGB,

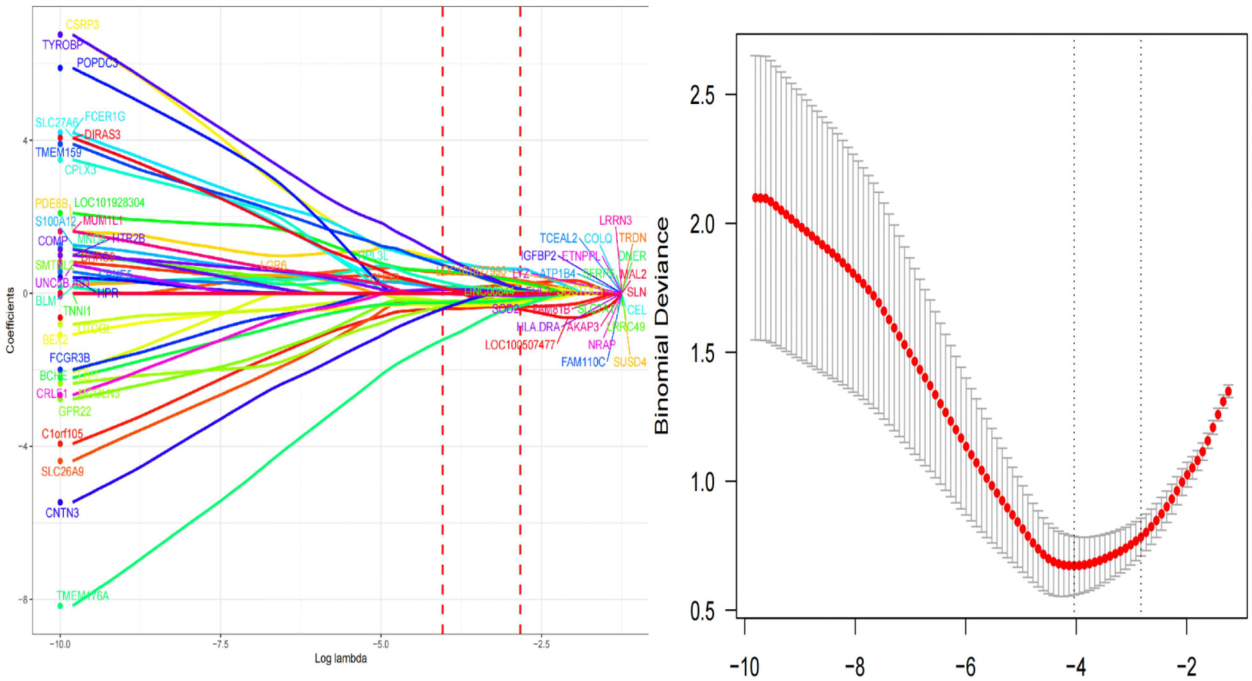


FIGURE 6
Biomarkers screening and optimal parameter (lambda) in the Lasso model.

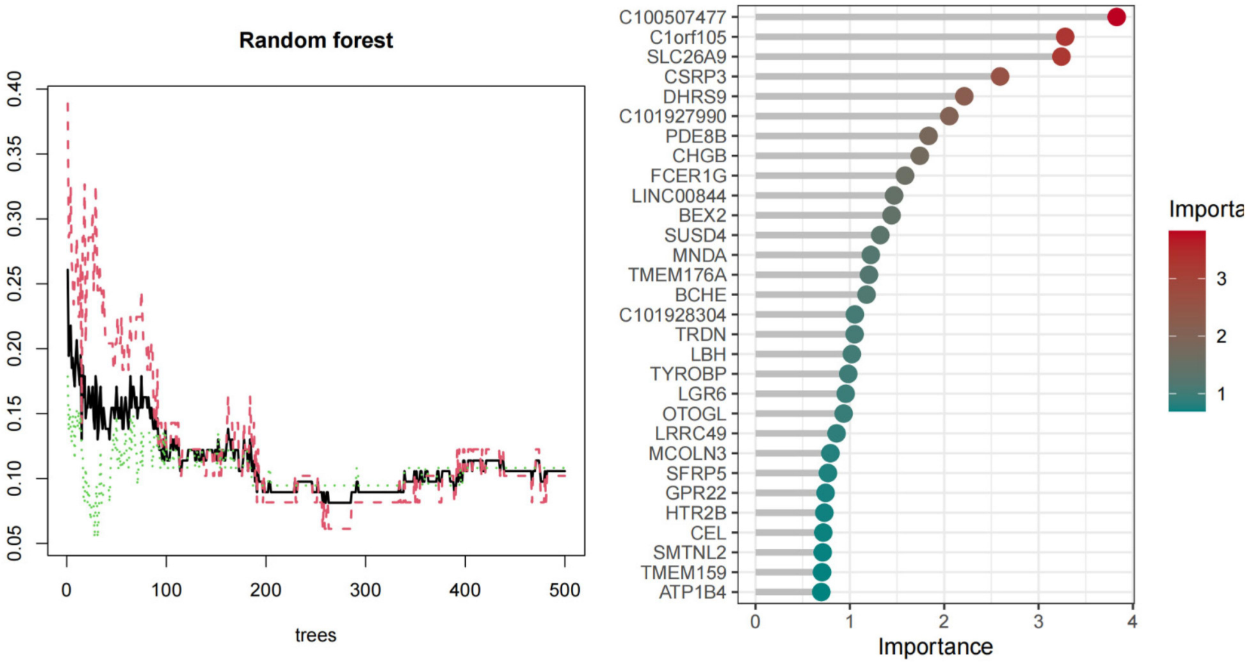


FIGURE 7
Biomarkers screening relative importance of overlapping candidate top 20 genes calculate& the RF model.

PDE8B, CSRP3 and FCER1G was up-regulated, whereas that of C1orf105 was down-regulated. The ROC curve analysis results showed that the AUC of each gene exceeded 0.75, indicating significant diagnostic value. The visualization results are shown in Figure 13. Similarly, in the external validation set (Figure 14), each gene showed great diagnostic value.

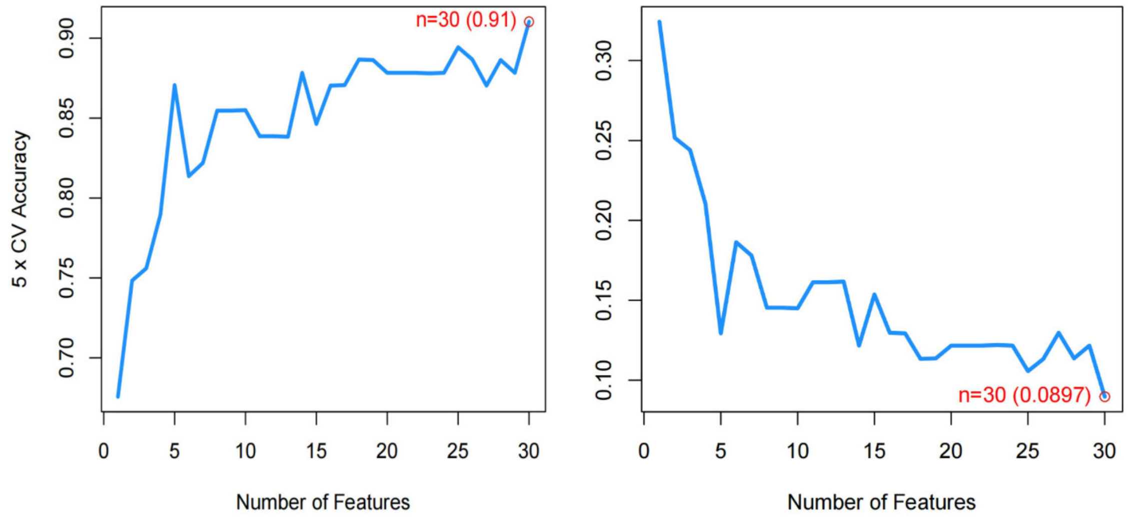


FIGURE 8
The curve with the highest and lowest biomarker screening accuracy in the SVM-RFE model.

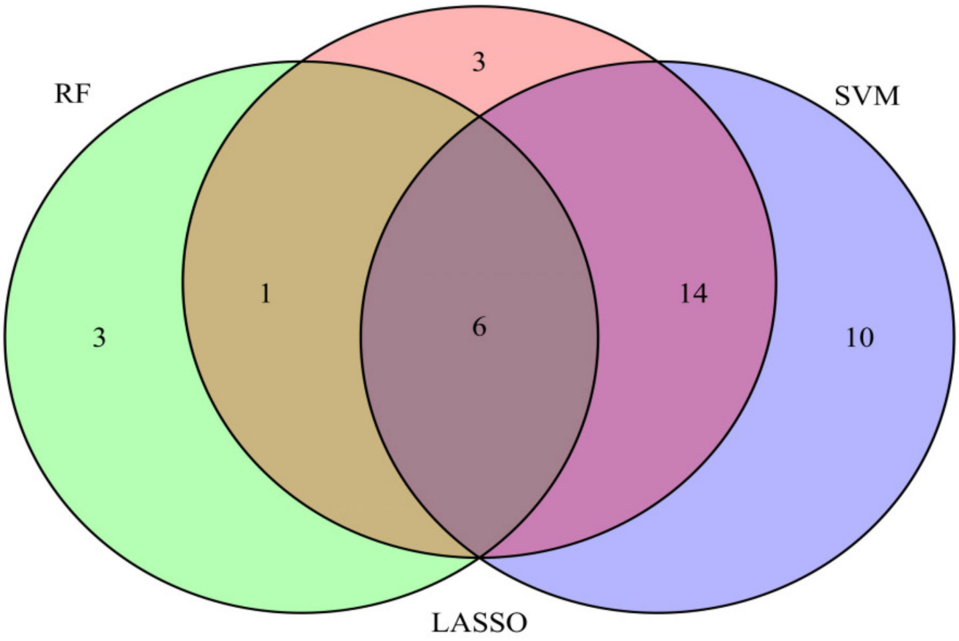


FIGURE 9
Venn diagram of six candidate genes screened by three machine learning algorithms.

Expression levels in single-cell transcriptome data

To further explore the relationship between the six key genes and atrial fibrillation, we downloaded right auricular tissue samples from 18 AF patients and 16 non-AF individuals in the GSE255612 dataset of the GEO database. After data pre-processing, normalization, scaling, and cell clustering, 12 distinct clusters were identified in the dataset. Upon cell annotation, these clusters were categorized into 12 cell types, namely Fibroblasts, Cardiomyocytes, Macrophages, Endothelial Cells, Pericytes, Adipocytes, Smooth Muscle Cells, T Cells, Neuroendocrine Cells, Mast Cells, Mesenchymal Stem Cells, and Proliferating Cells (Figure 15). Further analysis revealed that

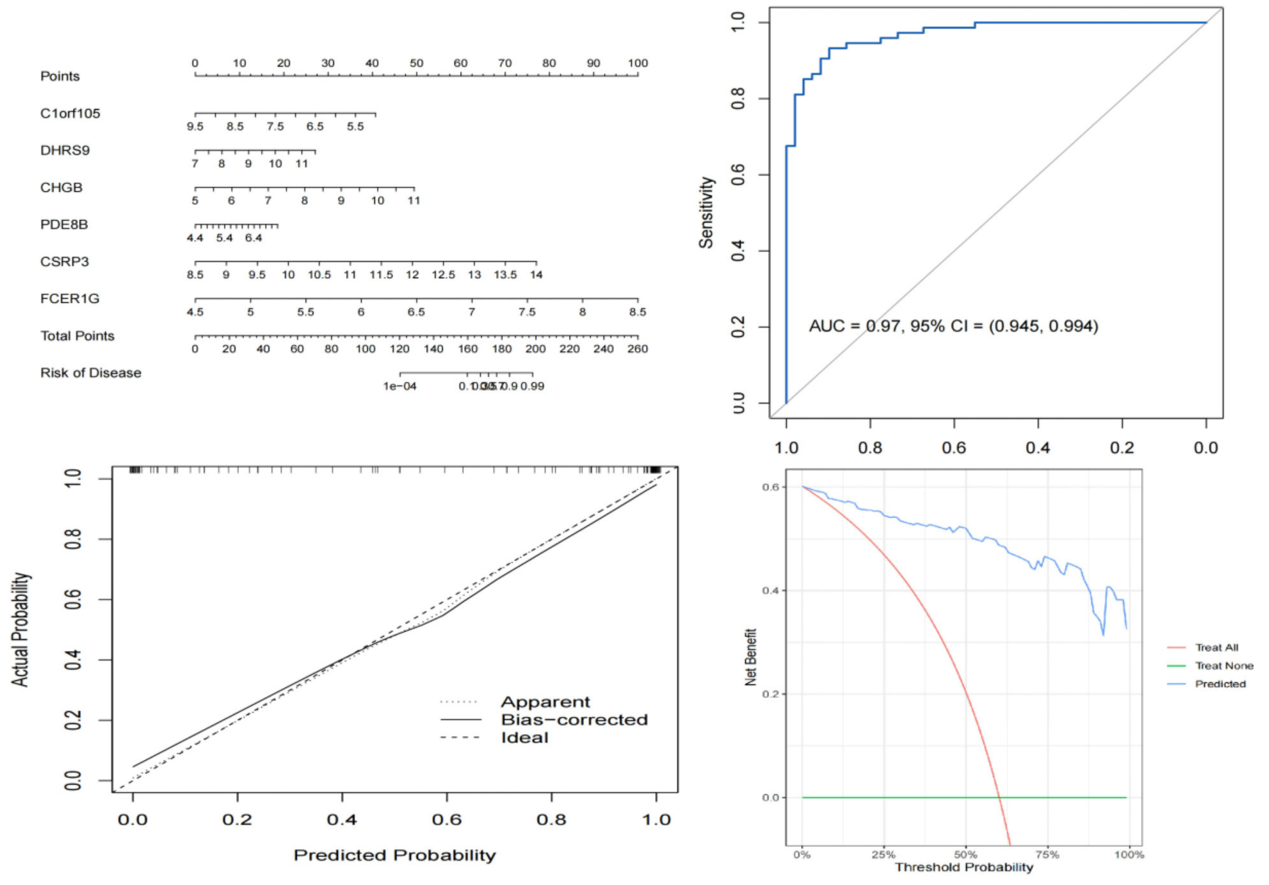


FIGURE 10
The visible nomogram, ROC curve, calibration curve, DCA curve for diagnosing AF.

DHRS9 and CSRP3 were predominantly expressed in cardiomyocytes, PDE8B in Adipocytes and cardiomyocytes, and FCER1G in macrophages (Figures 16, 17).

To determine if the hub genes were differentially expressed within specific cell types, we performed comparative analysis between AF and control samples for each major cell population, including fibroblasts, cardiomyocytes and macrophages. Violin plots illustrating the expression distribution of the six hub genes in fibroblasts are presented in Supplementary Figures S1, S2. Notably, none of these genes exhibited significant differential expression at the single-cell level within these populations. This indicates that their identification as differentially expressed genes in the bulk tissue analysis is likely attributable to AF-associated changes in the cellular composition of the atrial tissue, such as the expansion of fibroblast and macrophage populations, rather than substantial changes in their expression level within individual cells.

qRT-PCR experimental validations of the hub genes

First, we collected right auricular tissues from 4 AF patients and 4 non-AF patients. qRT-PCR results showed that mRNA levels of DHRS9, CHGB, PDE8B, CSRP3, and FCER1G were downregulated in right auricular tissues of patients with AF and upregulated in C1orf105 compared with non-lesional control tissues (Figure 18).

Western blot experimental validations of the hub genes

To further validate the protein expression levels of the six hub genes, we performed Western blot analysis on right auricular tissues from 3 AF patients and 3 non-AF controls. Consistent with the mRNA results, the protein levels of DHRS9, CHGB, PDE8B, CSRP3, and FCER1G were significantly downregulated in AF tissues, whereas C1orf105 protein expression was upregulated compared to controls (Figure 19).

Discussion

This study has systematically revealed the key molecular mechanisms and potential therapeutic targets in the development of atrial fibrillation by integrating single-cell and bulk transcriptomic data with machine learning algorithms.

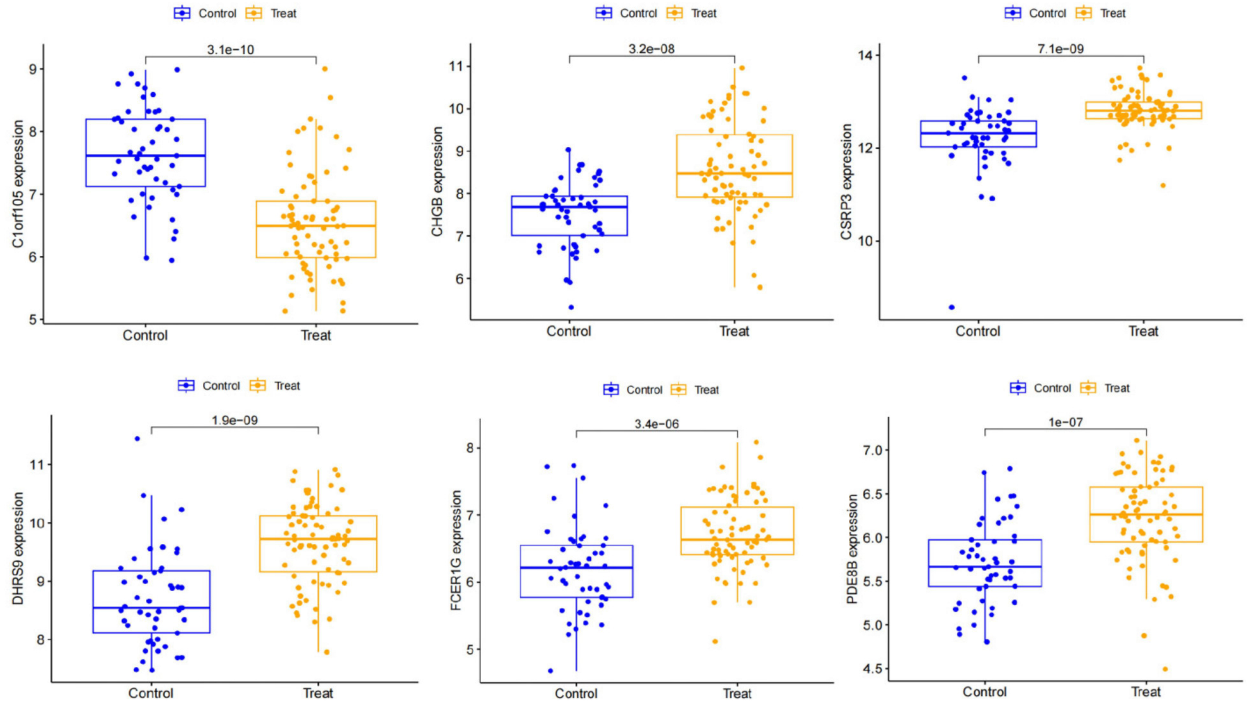


FIGURE 11
Expression of Hub genes in AF patients compared to normal controls in the training set.

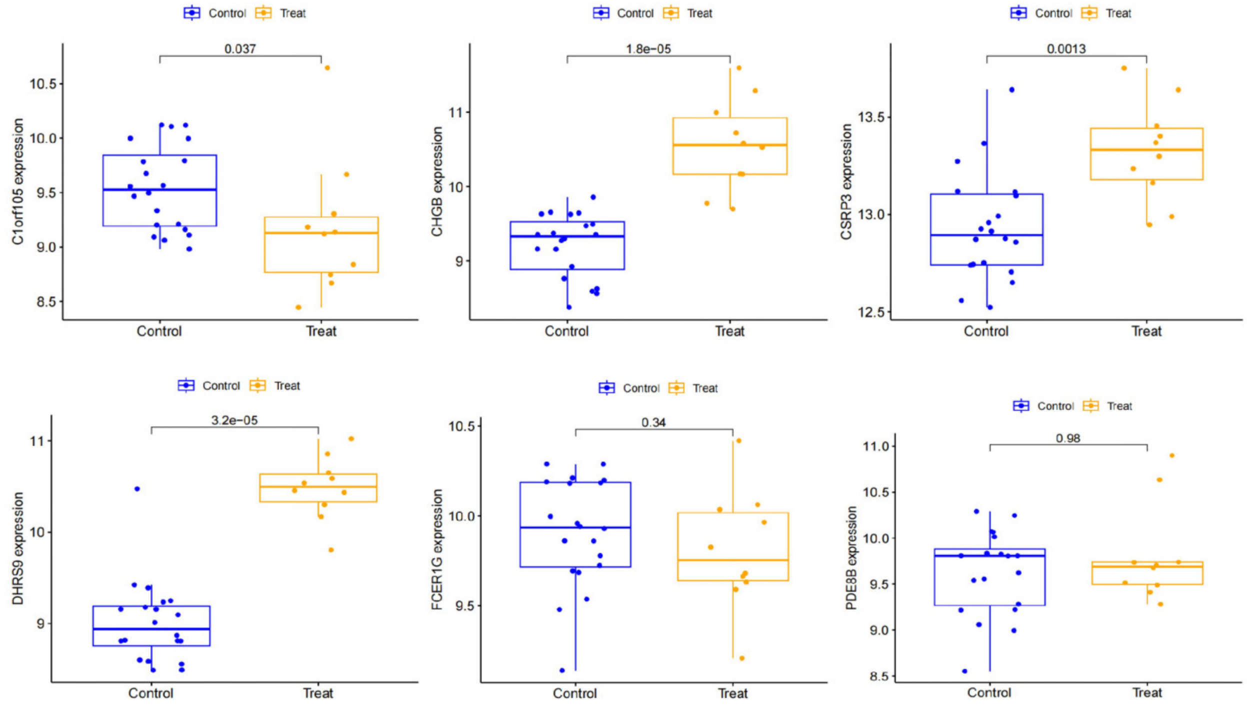


FIGURE 12
Expression of Hub genes in AF patients compared to normal controls in the validation set.

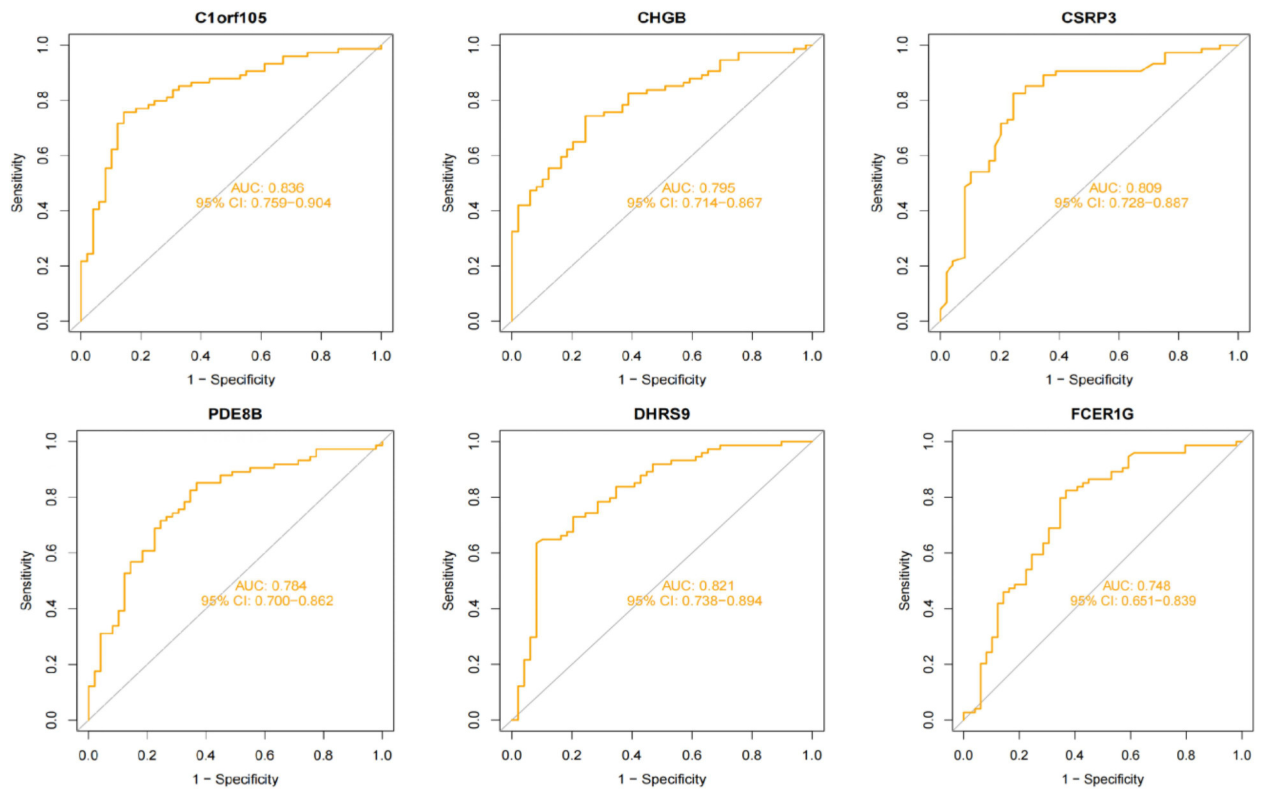


FIGURE 13
The ROC curve of each candidate genes in the training set.

Our application of machine learning to dissect the molecular underpinnings of AF aligns with a growing trend in cardiovascular medicine, particularly in electrophysiology, to leverage artificial intelligence (AI) for enhanced disease understanding and patient management. For instance, recent advances demonstrate the powerful role of AI and machine learning in electrophysiology, ranging from analyzing electrocardiograms for improved AF detection and classification to predicting ablation outcomes and optimizing patient-specific treatment strategies (23). Our study extends this paradigm by applying similar computational intelligence not to clinical signal data, but to high-dimensional transcriptomic data. This approach allows us to move beyond correlation towards identifying causative molecular features and cell-type-specific expressions that underlie the AF substrate. By integrating bulk and single-cell RNA sequencing with robust machine learning algorithms, we demonstrate how AI-driven bioinformatics can uncover novel, interpretable biomarker signatures that may inform both mechanistic biology and future precision medicine approaches in AF.

Several hub genes closely related to AF have been identified (C1orf105, DHRS9, CHGB, PDE8B, CSRP3, FCER1G). Functional enrichment analysis indicates that calcium signaling pathways, immune microenvironment imbalance, and myocardial structural remodeling play a central role in AF. Single-cell transcriptomic data further reveals the cell—type—specific expression patterns of these hub genes.

DHRS9 is specifically highly expressed in cardiomyocytes, suggesting it may play an important role in cardiomyocyte electrophysiology or structural remodeling (24). DHRS9 encodes a member of the dehydrogenase/reductase family 9 involved in retinoic acid metabolism, and retinoic acid signaling has been proven to be related to cardiac development and fibrosis regulation (25). In this study, the significant differential expression of DHRS9 may reflect myocardial cell metabolic reprogramming in AF patients, leading to abnormal calcium signaling pathways, thereby inducing arrhythmias. In addition, the association of DHRS9 with cardiomyocyte-related pathways, such as myocardial contraction and myofibril assembly, implies that it may be involved in AF progression by regulating the contractility of cardiomyocytes.

CSRP3 is highly expressed in cardiomyocytes, and its encoded protein is involved in sarcomere assembly and cytoskeletal stabilization (26, 27). This study shows that downregulated CSRP3 expression may be closely related to myocardial fibrosis and structural remodeling in AF patients. Previous studies have confirmed that CSRP3 deficiency can lead to the disruption of the Z-disc structure in cardiomyocytes, thereby inducing arrhythmias (28). The significant enrichment of CSRP3 in “myofibril” and “Z-disc” cell components further supports its key role in maintaining the structural integrity of cardiomyocytes. Moreover, the interaction of CSRP3 with calmodulin may indirectly influence the occurrence of AF by regulating calcium ion homeostasis.

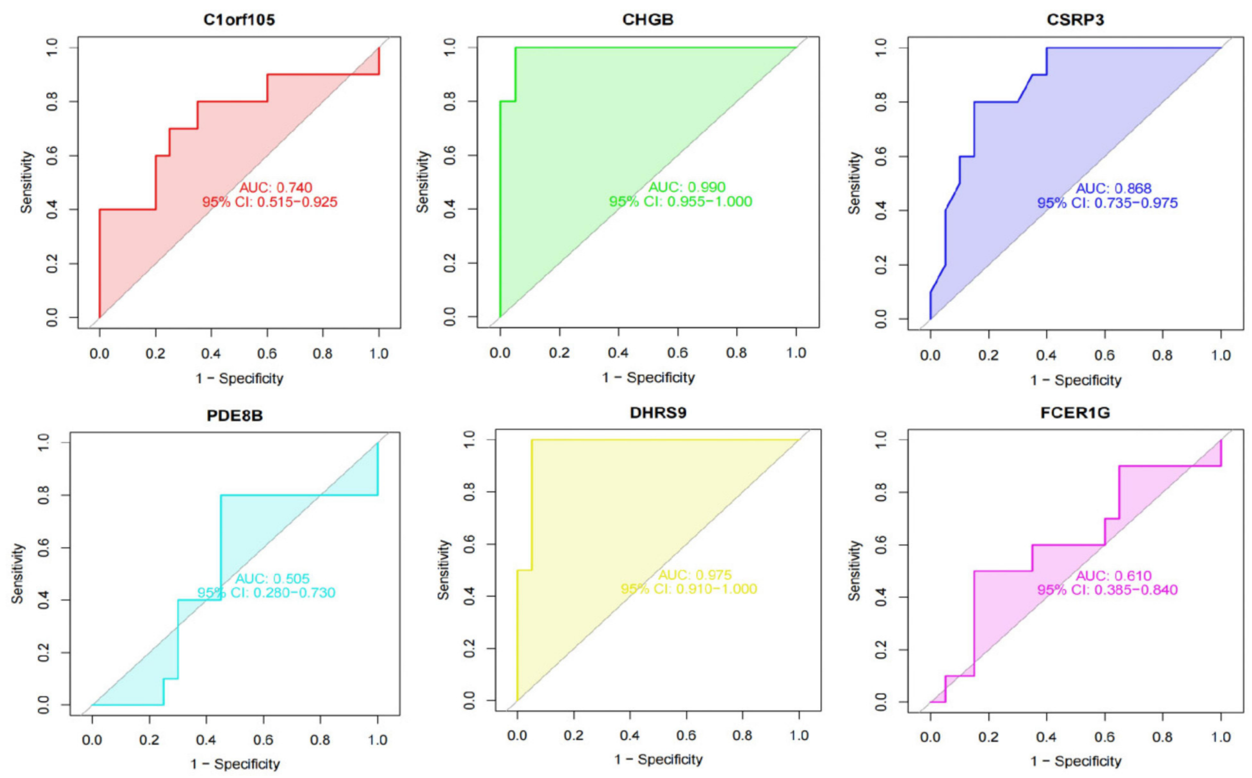


FIGURE 14
The ROC curve of each candidate genes in the validation set.

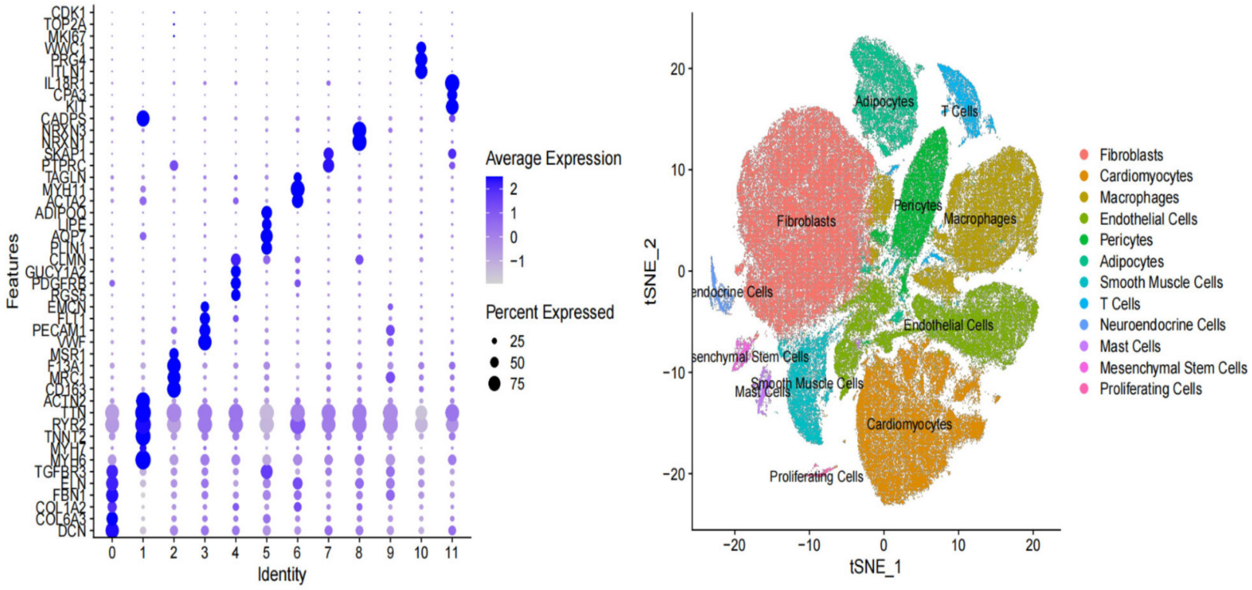


FIGURE 15
T-SNE clustering visualization for single-cell transcriptome data.

The dual-expression pattern of PDE8B in adipocytes and cardiomyocytes reveals the potential role of metabolic regulation in AF. PDE8B encodes phosphodiesterase 8B, which is involved in energy metabolism and signal transduction by degrading cAMP (29). This study finds that abnormal expression of PDE8B may lead to an imbalance in cAMP levels within cardiomyocytes, thereby

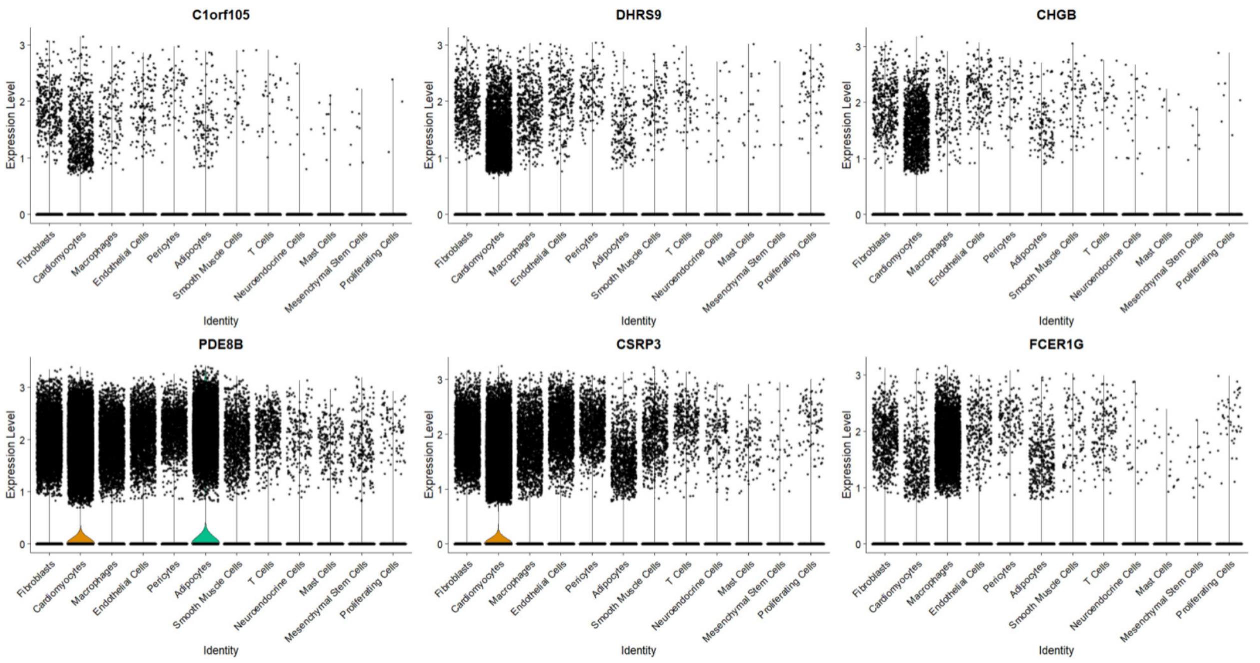


FIGURE 16
Expression levels of six genes in single-cell treatscriptome data.

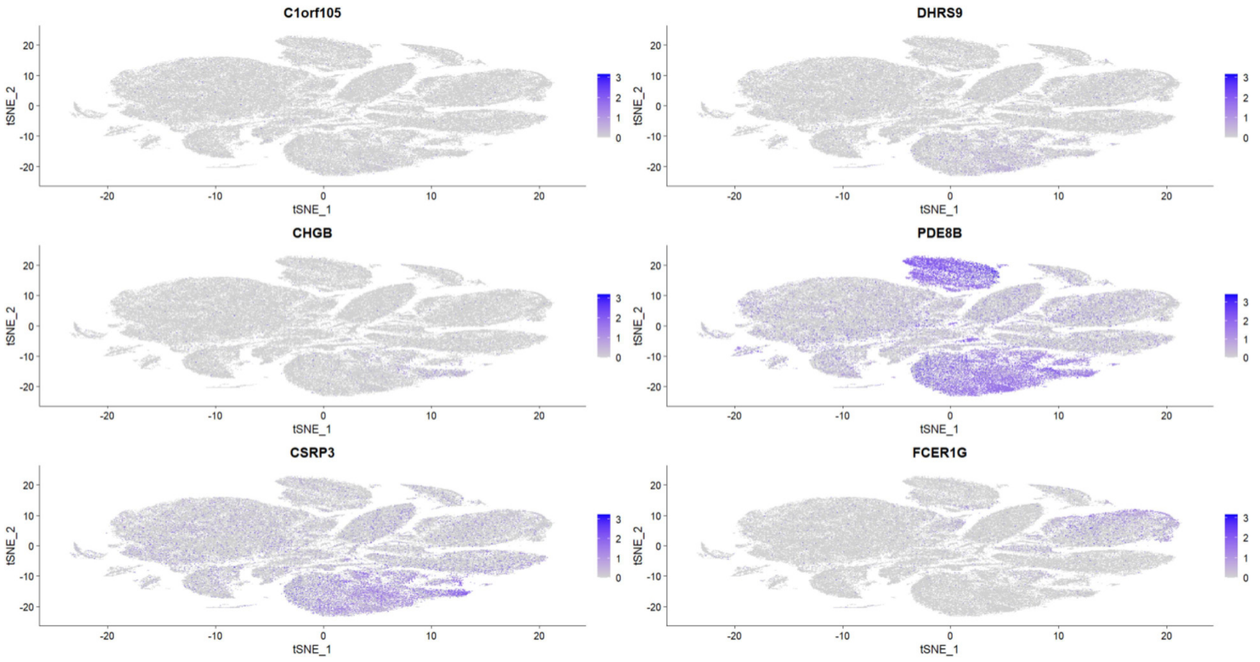


FIGURE 17
Distribution of six gene expressions in t-SNE space.

affecting calcium ion release. The KEGG-enriched "cardiac muscle contraction" pathway supports this finding. Additionally, the high expression of PDE8B in adipocytes may suggest that adipose tissue-derived factors can regulate myocardial electrical activity through a paracrine pathway, offering a new perspective on the metabolic-electrophysiological coupling mechanism of AF.

The specific high expression of FCER1G in macrophages suggests that it is involved in AF progression through immune-

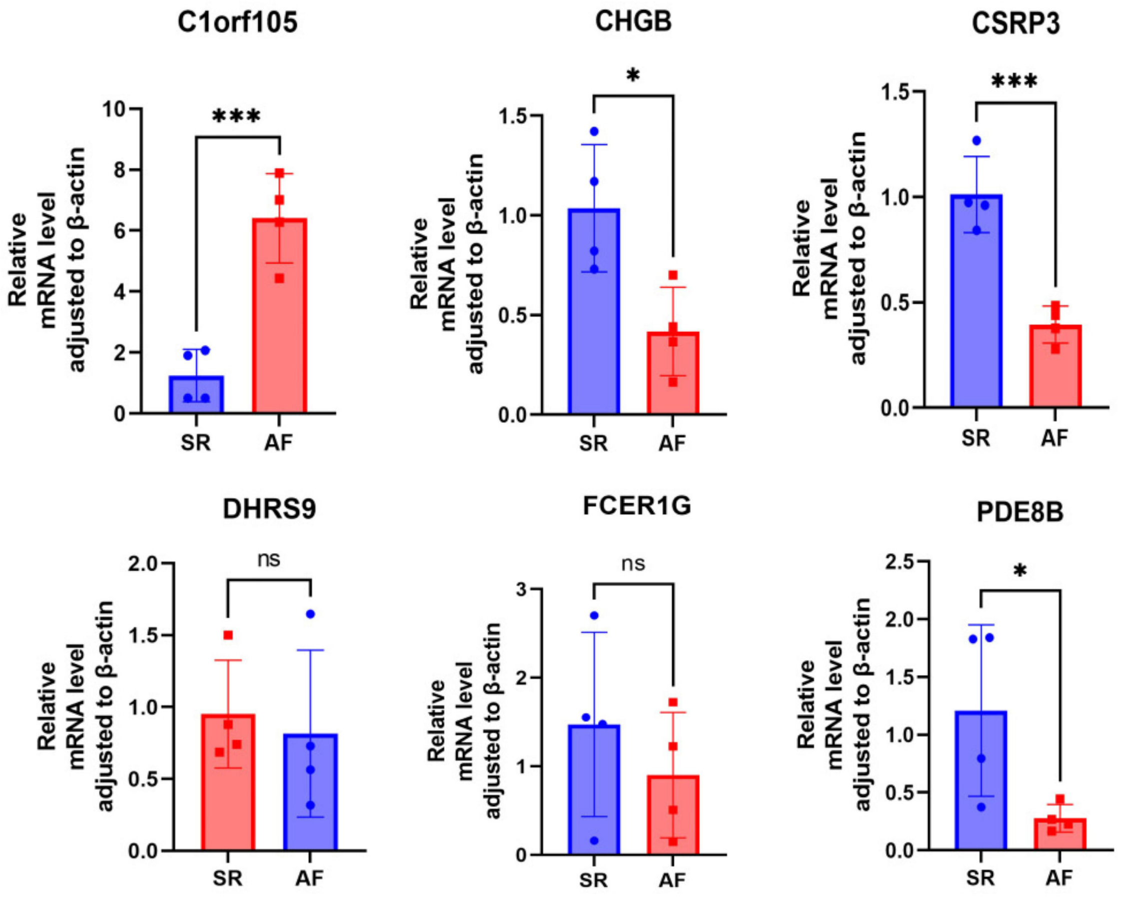


FIGURE 18
RT-qPCR analysis of six genes expression.

inflammatory pathways. FCER1G encodes the high-affinity IgE receptor γ-chain, a key molecule in the activation of mast cells and macrophages (30). This study shows an increase in macrophage infiltration in AF patients. FCER1G may promote the release of pro-inflammatory factors by activating the NF-κB pathway, thereby aggravating atrial fibrosis and electrical remodeling. Its association with "natural killer cell-mediated cytotoxicity" indicates that it may influence the AF microenvironment by regulating immune cell interactions, providing a potential target for targeted immunotherapy.

As newly-discovered AF-associated genes, the specific functions of C1orf105 and CHGB remain to be further elucidated. C1orf105 is widely expressed in single-cell data and may be involved in atrial remodeling by regulating cell proliferation or apoptosis. CHGB is commonly found in neuroendocrine cells, and its upregulated expression may reflect autonomic nervous system dysregulation in AF patients. This is consistent with previous reports that autonomic imbalance can trigger AF (31). Although the functions of these two genes are not yet clear, their association with "neuroendocrine regulation" and "cell proliferation" pathways suggests their potential role in AF, which needs to be verified through functional experiments.

The biomarkers identified in this study have distinct translational pathways depending on their primary source of expression. Tissue-based markers, such as CSRP3 and DHRS9 which are highly expressed in cardiomyocytes, directly reflect the pathophysiological processes of atrial remodeling, fibrosis, and electrophysiological dysfunction. They represent promising therapeutic targets for interfering with the core mechanisms of AF. However, their clinical application as diagnostic tools is limited by the invasiveness of obtaining cardiac tissue. In contrast, the detection of key genes like FCER1G and PDE8B in peripheral blood mononuclear cells (PBMCs), as revealed by our single-cell analysis, offers a promising avenue for non-invasive diagnosis. Blood-based biomarkers could be developed into liquid biopsies for AF screening, risk stratification, and potentially monitoring treatment response. It is important to note that while blood-based markers provide high clinical applicability, their expression levels may reflect systemic states such as inflammation or metabolic alterations, which could be influenced by comorbidities. Therefore, the integration of tissue-specific mechanistic insights with blood-based non-invasive detection methods could facilitate the development of a comprehensive strategy for managing AF.

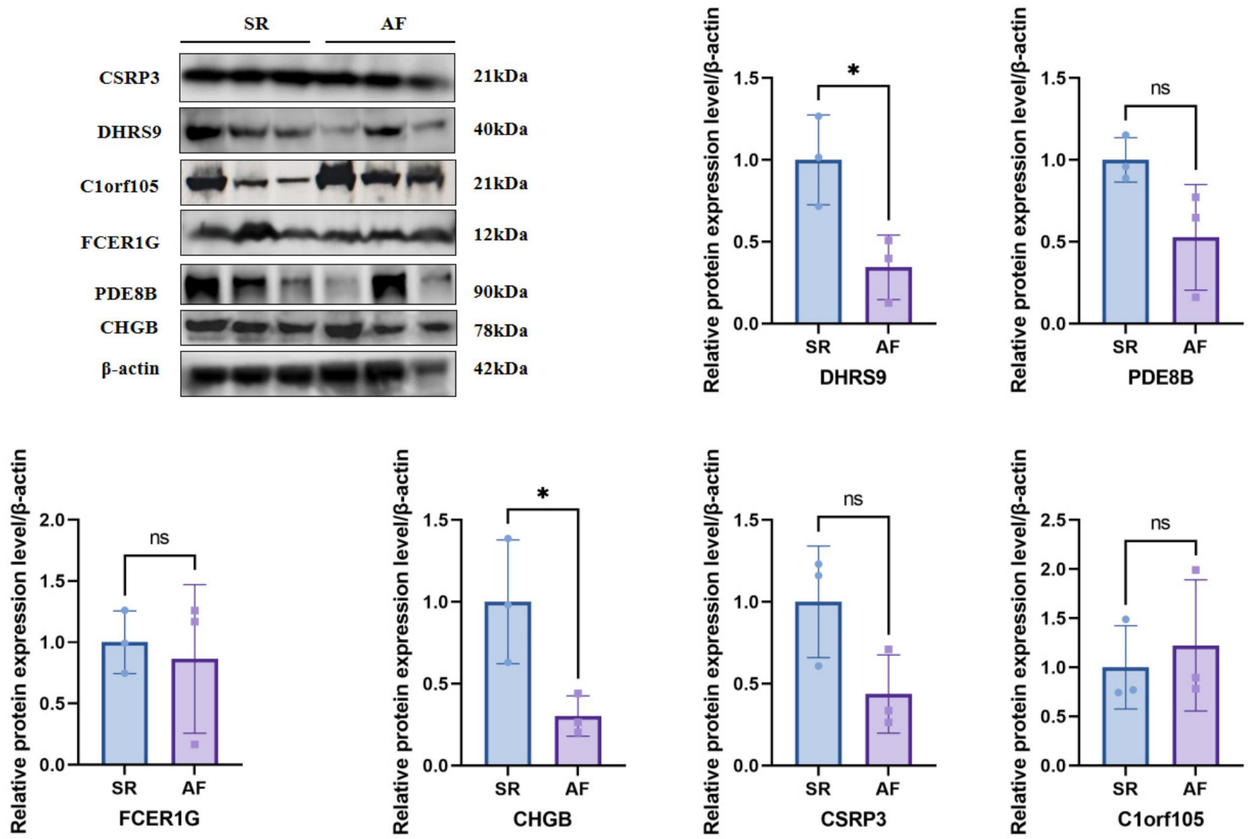


FIGURE 19
Western blotting and quantitative analysis of six genes expression.

Immune infiltration analysis shows that the proportion of macrophages and activated dendritic cells is increased in AF patients, while the number of effector memory CD8+ T cells is reduced. This is consistent with the characteristics of the chronic inflammatory state in AF patients. The nomogram model based on the five-gene signature shows excellent diagnostic performance, and its robustness has been validated in an independent dataset. This finding provides a theoretical basis for the development of non-invasive AF biomarker detection. However, the clinical application of the current model still needs further validation in prospective cohorts, and its value in AF subtype stratification or treatment-response prediction needs to be explored.

In addition, the identification of these hub genes and their expression patterns in specific cell types provides novel insights into the pathophysiology of AF. For instance, the high expression of DHRS9 in cardiomyocytes and its association with metabolic reprogramming highlight the importance of metabolic alterations in AF. This could lead to the development of therapeutic strategies targeting metabolic pathways to modulate cardiac electrophysiology and structure. Similarly, the role of CSRP3 in maintaining cardiomyocyte integrity and its link to fibrosis suggest that preserving or restoring its function might mitigate AF progression. Moreover, the dual expression of PDE8B in adipocytes and cardiomyocytes underscores the complex interplay between metabolic tissues and cardiac function, indicating that targeting adipocyte-derived factors could be a novel approach to manage AF.

The immune-related findings, particularly the overexpression of FCER1G in macrophages and the increased infiltration of macrophages in AF patients, emphasize the inflammatory nature of AF. This supports the potential of immunotherapeutic strategies in AF management. The association of FCER1G with immune cell interactions and its role in promoting pro-inflammatory cytokines through the NF-κB pathway offer specific targets for intervention. Modulating the immune response in AF could not only reduce inflammation but also prevent adverse structural remodeling.

Overall, this study bridges the gap between transcriptomic data and functional insights, providing a comprehensive view of AF mechanisms. It highlights the importance of integrating multi-omics data with advanced analytical techniques to uncover disease mechanisms and identifies potential therapeutic targets. Future research should focus on validating these findings in larger, diverse cohorts and exploring the functional roles of these genes through experimental models to translate these insights into clinical applications. Despite the limitations of this study, including its retrospective design and the need for further experimental validation, the identified genes and pathways present promising avenues for developing novel diagnostic tools and personalized treatment strategies for AF.

Conclusion

This study reveals key molecular mechanisms and potential therapeutic targets for AF. It identifies six genes closely related to AF and demonstrates their specific expression patterns in different cell types. The constructed nomogram model shows excellent diagnostic performance and provides a basis for developing non-invasive biomarkers for AF. However, further experimental validation is needed for clinical application.

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 approved by the Ethics Committee of Provincial Hospital Affiliated to Fuzhou University [Approval No: K2025-07-003]. We confirmed that all experiments were performed in accordance with the regulations. Written informed consent was obtained from all participants for this study.

Author contributions

YW: Data curation, Investigation, Methodology, Software, Writing – original draft. H-YY: Conceptualization, Data curation, Methodology, Writing – review & editing. Z-AF: Writing – review & editing. J-CZ: Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by grants obtained from the National Natural Science Foundation of China [Nos: 82070341], the Natural Science Foundation of Fujian Province [Nos: 2019J01189, 2023J011189], and the Fujian Province Science and Technology Innovation Joint Fund Project [Nos: 2023Y9350].

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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/fcvm.2025.1652467/full#supplementary-material

References

1. Treewaree S, Lip GYH. An updated global perspective of atrial fibrillation: trends, risk factors, and socioeconomic disparities. *CJC Open*. (2025) 7(3):259–61. doi: 10.1016/j.cjco.2024.12.003

2. Joglar JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: a report of the American college of cardiology/American heart association joint committee on clinical practice guidelines. *Circulation*. (2024) 149(1):e1–e156. doi: 10.1161/CIR.0000000000001193

3. Rush KL, Seaton CL, Burton L, Loewen P, O'Connor BP, Moroz L, et al. Quality of life among patients with atrial fibrillation: a theoretically-guided cross-sectional study. *PLoS One*. (2023) 18(10):e0291575. doi: 10.1371/journal.pone.0291575

4. Murakata Y, Yamagami F, Murakoshi N, Xu D, Song Z, Li S, et al. Electrical, structural, and autonomic atrial remodeling underlies atrial fibrillation in inflammatory atrial cardiomyopathy. *Front Cardiovasc Med*. (2022) 9:1075358. doi: 10.3389/fcvm.2022.1075358

5. Iwamiya S, Ihara K, Nitta G, Sasano T. Atrial fibrillation and underlying structural and electrophysiological heterogeneity. *Int J Mol Sci*. (2024) 25(18), 10193. doi: 10.3390/ijms251810193

6. Fakuade FE, Hubricht D, Möller V, Sobitov I, Liutkute A, Döring Y, et al. Impaired intracellular calcium buffering contributes to the arrhythmogenic substrate in atrial myocytes from patients with atrial fibrillation. *Circulation*. (2024) 150(7):544–59. doi: 10.1161/CIRCULATIONAHA.123.066577

7. Zhu X, Lv M, Cheng T, Zhou Y, Yuan G, Chu Y, et al. Bibliometric analysis of atrial fibrillation and ion channels. *Heart Rhythm*. (2024) 21(7):1161–9. doi: 10.1016/j.hrthm.2024.01.032

8. Chander S, Kumari R, Luhana S, Shiwlani S, Parkash O, Sorath F, et al. Antiarrhythmic drug therapy and catheter ablation in patients with paroxysmal or persistent atrial fibrillation: a systematic review and meta-analysis. *BMC Cardiovasc Disord*. (2024) 24(1):321. doi: 10.1186/s12872-024-03983-z

9. Boersma L, Rienstra M, de Groot JR. Therapeutic options for patients with advanced atrial fibrillation: from lifestyle and medication to catheter and surgical ablation. *Neth Heart J.* (2020) 28(Suppl 1):13–8. doi: 10.1007/s12471-020-01447-5

10. Shen NN, Zhang C, Wang N, Wang JL, Gu ZC, Han H. Effectiveness and safety of under or over-dosing of direct oral anticoagulants in atrial fibrillation: a systematic review and meta-analysis of 148909 patients from 10 real-world studies. *Front Pharmacol.* (2021) 12:645479. doi: 10.3389/fphar.2021.645479

11. Aldian FM, Visuddho V, Witarto BS, Witarto AP, Hartanto JO, Sembiring YE, et al. Effectiveness and safety of different types of ablation modalities in patients with atrial fibrillation: a Bayesian network meta-analysis from randomized controlled trials. *J Cardiothorac Surg.* (2025) 20(1):225. doi: 10.1186/s13019-025-03460-4

12. Camm AJ, Naccarelli GV, Mittal S, Crijns H, Hohnloser SH, Ma CS, et al. The increasing role of rhythm control in patients with atrial fibrillation: JACC state-of-the-art review. *J Am Coll Cardiol.* (2022) 79(19):1932–48. doi: 10.1016/j.jacc.2022.03.337

13. Morabito A, De Simone G, Pastorelli R, Brunelli L, Ferrario M. Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review. *J Transl Med.* (2025) 23(1):425. doi: 10.1186/s12967-025-06446-x

14. Dai X, Shen L. Advances and trends in omics technology development. *Front Med (Lausanne).* (2022) 9:911861. doi: 10.3389/fmed.2022.911861

15. Seeler S, Arnarsson K, Dreßen M, Krane M, Doppler SA. Beyond the heartbeat: single-cell omics redefining cardiovascular research. *Curr Cardiol Rep.* (2024) 26(11):1183–96. doi: 10.1007/s11886-024-02117-3

16. Miranda AMA, Janbandhu V, Maatz H, Kanemaru K, Cranley J, Teichmann SA, et al. Single-cell transcriptomics for the assessment of cardiac disease. *Nat Rev Cardiol.* (2023) 20(5):289–308. doi: 10.1038/s41569-022-00805-7

17. Lin M, Guo J, Gu Z, Tang W, Tao H, You S, et al. Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice. *J Transl Med.* (2025) 23(1):388. doi: 10.1186/s12967-025-06425-2

18. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. *Genes (Basel).* (2019) 10(2):87. doi: 10.3390/genes10020087

19. Vijayakumar S, Magazzù G, Moon P, Occhipinti A, Angione C. A practical guide to integrating multimodal machine learning and metabolic modeling. *Methods Mol Biol.* (2022) 2399:87–122. doi: 10.1007/978-1-0716-1831-8_5

20. Anjum F, Alsharif A, Bakhuraysah M, Shafie A, Hassan MI, Mohammad T. Discovering novel biomarkers and potential therapeutic targets of amyotrophic lateral sclerosis through integrated machine learning and gene expression profiling. *J Mol Neurosci:.* (2025) 75(2):61. doi: 10.1007/s12031-025-02340-9

21. Ahmadieh-Yazdi A, Mahdavinezhad A, Tapak L, Nouri F, Taherkhani A, Afshar S. Using machine learning approach for screening metastatic biomarkers in colorectal cancer and predictive modeling with experimental validation. *Sci Rep.* (2023) 13(1):19426. doi: 10.1038/s41598-023-46633-8

22. Ayubi E, Farashi S, Tapak L, Afshar S. Development and validation of a biomarker-based prediction model for metastasis in patients with colorectal cancer: application of machine learning algorithms. *Heliyon.* (2025) 11(1):e41443. doi: 10.1016/j.heliyon.2024.e41443

23. Cersosimo A, Zito E, Pierucci N, Matteucci A, La Fazia VM. A talk with ChatGPT: the role of artificial intelligence in shaping the future of cardiology and electrophysiology. *J Pers Med.* (2025) 15(5):205. doi: 10.3390/jpm15050205

24. Yang Y, Wang S, Xie X, Li J, Zhang R. Change of gene expression profiles in human cardiomyocytes and macrophages infected with SARS-CoV-2 and its significance. *Zhong nan da xue xue bao. Yi xue ban, J Cent South Uni Med Sci.* (2021) 46(11):1203–11. doi: 10.11817/j.issn.1672-7347.2021.210221

25. Wiesinger A, Boink GJJ, Christoffels VM, Devalla HD. Retinoic acid signaling in heart development: application in the differentiation of cardiovascular lineages from human pluripotent stem cells. *Stem Cell Rep.* (2021) 16(11):2589–606. doi: 10.1016/j.stemcr.2021.09.010

26. Marian AJ. Molecular genetic basis of hypertrophic cardiomyopathy. *Circ Res.* (2021) 128(10):1533–53. doi: 10.1161/CIRCRESAHA.121.318346

27. Vafiadaki E, Arvanitis DA, Sanoudou D. Muscle LIM protein: master regulator of cardiac and skeletal muscle functions. *Gene.* (2015) 566(1):1–7. doi: 10.1016/j.gene.2015.04.077

28. Vafiadaki E, Arvanitis DA, Papalouka V, Terzis G, Roumeliotis TI, Spengos K, et al. Muscle lim protein isoform negatively regulates striated muscle actin dynamics and differentiation. *FEBS J.* (2014) 281(14):3261–79. doi: 10.1111/febs.12859

29. Tsai LC, Chan GC, Nangle SN, Shimizu-Albergine M, Jones GL, Storm DR, et al. Inactivation of Pde8b enhances memory, motor performance, and protects against age-induced motor coordination decay. *Genes Brain Behav.* (2012) 11(7):837–47. doi: 10.1111/j.1601-183X.2012.00836.x

30. Yang R, Chen Z, Liang L, Ao S, Zhang J, Chang Z, et al. Fc fragment of IgE receptor ig (FCER1G) acts as a key gene involved in cancer immune infiltration and tumour microenvironment. *Immunology.* (2023) 168(2):302–19. doi: 10.1111/imm.13557

31. Zhang K, Rao F, Rana BK, Gayen JR, Calegari F, King A, et al. Autonomic function in hypertension; role of genetic variation at the catecholamine storage vesicle protein chromogranin B. *Circ. Cardiovasc Genet.* (2009) 2(1):46–56. doi: 10.1161/CIRCGENETICS.108.785659

frontiers | Frontiers in Network Physiology

TYPE Original Research
PUBLISHED 25 November 2025
DOI 10.3389/fnetp.2025.1701758

OPEN ACCESS

EDITED BY
Ulrich Parlitz,
Max Planck Institute for Dynamics and Self-Organization, Germany

REVIEWED BY
Calin Corciova,
Grigore T. Popa University of Medicine and Pharmacy, Romania
Gerald Marx,
Boston Children's Hospital and Harvard Medical School, United States

*CORRESPONDENCE
Silvia Corona,
✉ silviacoro89@gmail.com
Walid Ben Ali,
✉ dr.walidbenali@gmail.com

RECEIVED 09 September 2025
REVISED 29 October 2025
ACCEPTED 10 November 2025
PUBLISHED 25 November 2025

CITATION
Corona S, Godefroy T, Tastet O, Corbin D, Modine T, von Bardeleben S, Lesage F and Ben Ali W (2025) Towards standardizing mitral transcatheter edge-to-edge repair with deep-learning algorithm: a comprehensive multi-model strategy.
Front. Netw. Physiol. 5:1701758.
doi: 10.3389/fnetp.2025.1701758

Towards standardizing mitral transcatheter edge-to-edge repair with deep-learning algorithm: a comprehensive multi-model strategy

Silvia Corona[1]*, Théo Godefroy[2], Olivier Tastet[2], Denis Corbin[2], Thomas Modine[3], Stephan von Bardeleben[1], Frédéric Lesage[1,2] and Walid Ben Ali[1]*

[1]Structural Heart Valve Center, Montreal Heart Institute, Montreal, QC, Canada, [2]Biomedical Engineering Department, Polytechnique Montréal, Montreal, QC, Canada, [3]UMCV, Hôpital Haut-Lévêque, CHU Bordeaux, Bordeaux, France

Background: Severe mitral valve regurgitation requires comprehensive evaluation for optimal treatment. Initial screening uses transthoracic echocardiography (TTE), followed by transesophageal echocardiography (TEE) to determine eligibility for adequate intervention. Mitral Transcatheter Edge-to-Edge Repair (M-TEER) indications are based on detailed and quality valve and sub-valvular apparatus assessment, including anatomy and regurgitation pathophysiology.
Aim: To develop AI algorithms for standardizing M-TEER eligibility assessment using TTE and TEE echocardiograms, supporting all stages of mitral valve regurgitation evaluation to assist non-expert centers throughout the entire process, from severe mitral valve regurgitation diagnostic to M-TEER procedure.
Methods: Three deep learning algorithms were developed using echocardiographic data from M-TEER patients performed at Montreal Heart Institute (2018–2025). 1. ECHO-PREP was trained to identify key diagnostic views in TTE ($n = 530$) and diagnostic and procedural views in TEE ($n = 2,222$) examinations to determine the level of quality images needed to do a M-TEER. 2. 4D TEE segmentation with automated mitral valve area (MVA) quantification ($n = 221$), and 3. 2D TEE scallop-level segmentation of leaflets and sub-valvular structures ($n = 992$).
Results: Preliminary results on test sets showed 95.7% accuracy in TTE view classification and 91% accuracy for TEE view classification. The 4D segmentation module demonstrated excellent agreement with manual MVA measurements ($R = 0.84$, $p < 0.001$), successfully discriminating patients undergoing M-TEER from those referred for surgical replacement ($p = 0.046$ for AI predictions). The 2D scallop-level analysis achieved a mean Dice score of 0.534 across 11 anatomical structures, with better performance in commonly represented configurations (e.g., A2-P2, P1-A2-P3).
Conclusion: ECHO-PREP demonstrates the feasibility of an integrated AI-assisted workflow for MR assessment, combining quality control, dynamic 4D valve quantification, and scallop-level anatomy interpretation. These results

support the potential of AI to standardize M-TEER eligibility, reduce inter-observer variability, and provide decision support across centers with different levels of expertise.

KEYWORDS

mitral regurgitation, AI, M-TEER, deep-learning, segmentation, TEE, 4D

1 Introduction

Mitral regurgitation (MR) is a prevalent valvular heart disease, affecting approximately 2% of the general population and up to 10% of individuals over 75 years of age (Nkomo et al., 2006). In patients with severe MR and high or prohibitive surgical risk, transcatheter edge-to-edge repair (M-TEER) has emerged as an established therapeutic option that can reduce symptoms, hospitalizations, and improve quality of life (Stone et al., 2018; Maisano et al., 2013).

Successful M-TEER depends critically on detailed anatomic and functional characterization of the mitral valve apparatus. This complex apparatus is a dynamic interface between the left atrium and ventricle, composed of two leaflets attached to a saddle-shaped annulus and supported by a subvalvular network of chordae tendineae and papillary muscles. Transesophageal echocardiography (TEE) remains the cornerstone imaging modality for pre-procedural assessment and intra-procedural guidance (Zamorano et al., 2011), providing high-quality imaging of cardiac structures in 2D and 3D, enabling real-time dynamic assessment. In contrast, transthoracic echocardiography (TTE) is typically reserved for initial screening and post-procedural follow-up. Precise quantification of valvular morphology and kinematics from these images can also feed into computational models, such as finite element simulations, to replicate patient-specific biomechanics (Votta et al., 2008). Deriving this level of detail, particularly a pixel-wise annotation of valve substructures from 4D TEE data, is a formidable task. The automation of mitral valve segmentation and tracking is hindered by intrinsic challenges of echocardiography, such as artifacts from patient motion, variable image quality, and scarse availability of expertly annotated 4D datasets for training. However, conventional clinical workflows rely heavily on expert interpretation and manual measurements, which are time-consuming and subject to inter- and intra-observer variability (Hien et al., 2014; Thomas et al., 2008). Artificial intelligence (AI), particularly deep learning, offers an opportunity to overcome these limitations by providing rapid, reproducible, and quantitative analysis of echocardiographic images.

Convolutional neural networks (CNNs), particularly encoder-decoder architectures like U-Net and its 3D extensions, have demonstrated remarkable success over the last decade in automating tasks in cardiac ultrasound, including chamber segmentation, functional analysis, and valvular assessment (Leclerc et al., 2019; Ouyang et al., 2020). Clinical and technical precedents illustrate this trajectory. Vendor-integrated solutions such as Anatomic Intelligence in Ultrasound (AIUS) (Philips Healthcare) have implemented automated recognition and measurement of cardiac structures, showing the feasibility of integrating anatomy-aware algorithms into daily workflows. Academic initiatives and challenges (e.g., the Mitral Valve Segmentation challenge -MVSEG- at the International Conference on Medical Image Computing and Computer Assisted Intervention congress -MICCAI-) have provided standardized benchmarks to accelerate innovation and compare algorithmic performance. The winning model at MVSEG 2023 (Synapse, 2025), often leveraging advanced architectures like nnU-Net or vision transformers, achieved state-of-the-art Dice scores, showcasing an unprecedented ability to accurately delineate the thin, dynamic mitral leaflets and complex annular geometry.

Several research groups have contributed to this field. Costa et al. (2019) developed a 2D CNN for leaflet segmentation in 2D TTE, while Carnahan et al. (2021) and Aly et al. (2022) focused on 3D segmentation from TEE using a 3D Residual UNet and nnUNet, respectively. Chen et al. (2023) introduced a two-stage nnUNet approach, initializing it with a classifier pre-trained to identify the valve's open and closed states. Munafò et al. (2024) created a Multi-Decoder Residual UNet to segment the annulus and both leaflets separately at end-systole from 3D TEE. A critical limitation of these studies is their inability to perform frame-by-frame (4D) analysis of the entire valve apparatus throughout the cardiac cycle. Previous 4D efforts have been restricted to annulus-specific segmentation (Andreassen et al., 2019; Andreassen et al., 2022) or tracking (Taskén et al., 2023), or were confined to 2D imaging for leaflets and annulus, as seen in the work of Wifstad et al. (2024), who used a UNet with attention gates for 2D TTE. Recently, Munafò et al. (2025) proposed a semi-supervised training strategy using pseudo-labeling for MV segmentation in 4D TEE employing a Teacher-Student framework to ensure reliable pseudo-label generation. The Student model demonstrated reliable frame-by-frame MV segmentation on 120 4D TEE recordings from 60 candidates for MV repair, accurately capturing leaflet morphology and dynamics throughout the cardiac cycle, with a significant reduction in inference time compared to the ensemble. Despite these advances, several challenges persist. Generalizability across vendors and imaging protocols is limited, and a fully automated 4D MV segmentation with a scallop-level analysis, which is also able to provide automated measurements in complex anatomies to define the M-TEER eligibility, is difficult. The development of such a method is highly challenged by the labor-intensive manual annotation process needed to generate the extensive datasets required for the supervised training of CNNs.

In this context, we developed an integrated deep learning framework for the comprehensive pre-procedural assessment of the mitral valve in patients with severe mitral regurgitation. Our solution features a three-stage algorithmic pipeline designed to: 1. assess the quality of available TTE and TEE images, 2. perform segmentation of the mitral annulus, leaflets, and scallops, and 3. automatically compute the mitral valve area (MVA) from 4D-TEE volumes. By generating reproducible, clinically relevant measurements, this approach has the potential to standardize

Abbreviations: AI, Artificial intelligence; MR, Mitral regurgitation; TTE, Transthoracic Echocardiography; TEE, Transesophageal echocardiography; M-TEER, Mitral transcatheter edge-to-edge repair; DL, Deep learning.

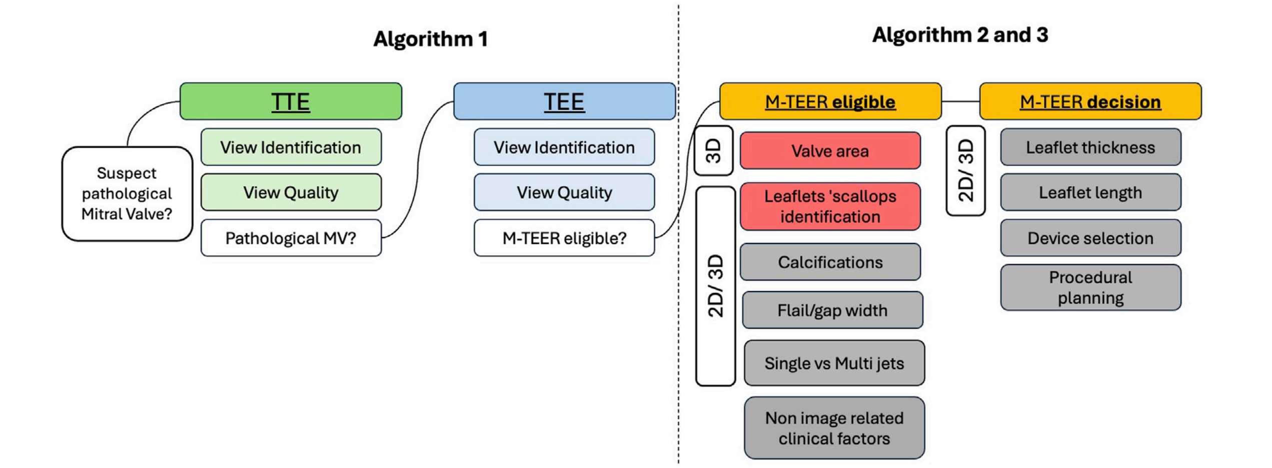


FIGURE 1
ECHO-PREP fully automatic clinical workflow for MV Assessment and M-TEER procedural planning. Grey boxes denote potential applications of the available algorithms that are currently under development or have not yet been validated. 2D = two-dimensional; 3D = three-dimensional; M-TEER = mitral transcatheter edge-to-edge repair; MV = mitral valve; TTE = transthoracic echocardiography; TEE = transesophageal echocardiography.

feasibility assessment and support heart team decision-making in transcatheter mitral interventions.

2 Methods

The proposed original multi-step workflow, called ECHO-PREP, consisting of three sequential algorithms for image quality assessment, mitral valve segmentation, including scallop-level analysis, and automated measurements, is illustrated in Figure 1.

TTE and TEE pre-procedural images from M-TEER (Mitraclip) and surgical mitral valve replacement (MVR) patients, performed at the Montreal Heart Institute from 1 January 2018, to 1 January 2025, were retrospectively collected. Both two-dimensional (2D) images, three-dimensional (3D), and four-dimensional (4D) volume images were used. 3D refers to single-volume acquisitions, whereas 4D refers to multi-volume datasets spanning the entire cardiac cycle. Our algorithm was primarily trained on 3D echo volumes to establish accurate segmentation performance. Once optimized in this setting, the model was subsequently extended and retrained to analyze sequences of 3D volumes across the cardiac cycle, thereby enabling full 4D assessment.

2.1 Automatic classification of 2D- TTE and TEE images: quality views assessment

2.1.1 Dataset processing and splitting

TEE and TTE video images were processed through a systematic pipeline. All frames were extracted from source videos using OpenCV, with each frame inheriting its parent video's label. Multi-label annotations were transformed to single labels using priority rules, removing technical artifacts such as 'delivery_system' and 'clip' tags. Dataset splitting was performed at the video level using instance_uid identifiers to prevent data leakage, ensuring no video appeared in multiple splits. Videos were stratified by label distribution and randomly assigned to training (50%), validation (25%), and test (25%) sets. This video-level splitting approach maintained temporal integrity while enabling robust model evaluation.

2.1.2 Model architecture and training

We employed MobileNetV3-Large (Elaziz et al., 2023) as our base architecture, initialized with pre-trained weights from ImageNet (Figure 2). For both quality assessment (binary classification) and view classification (multi-class single-label), only the final classification layer was modified to match the target classes, implementing a transfer learning approach. Training images underwent augmentation, including random horizontal and vertical flips, random rotation (+/−10°), resizing to 256 × 256 pixels, and random cropping to 224 × 224 pixels. Validation and test images underwent deterministic preprocessing, which included resizing to 256 × 256 pixels and center cropping to 224 × 224 pixels. To accelerate training, entire datasets were loaded into memory using a custom Dataset class. Models were trained using the Adam optimizer with a learning rate of 1 × 10ˆ (−4) and cross-entropy loss for 100 epochs, with a batch size of 128. Training utilized eight parallel data loading workers and CUDA acceleration. Model selection was based on validation accuracy, with the best-performing checkpoint saved for inference. Performance was monitored using both accuracy and the area under the receiver operating characteristic curve (AUROC). AUROC was computed using one-vs-rest methodology for multi-class view classification.

2.2 4D-TEE-based automatic MV segmentation and MVA measurements

2.2.1 Dataset and data preparation

This study utilized the MVSEG2023 public dataset (Synapse, 2025), a standardized collection of TEE volumes acquired using the Philips EPIQ cardiac ultrasound system. The dataset contains

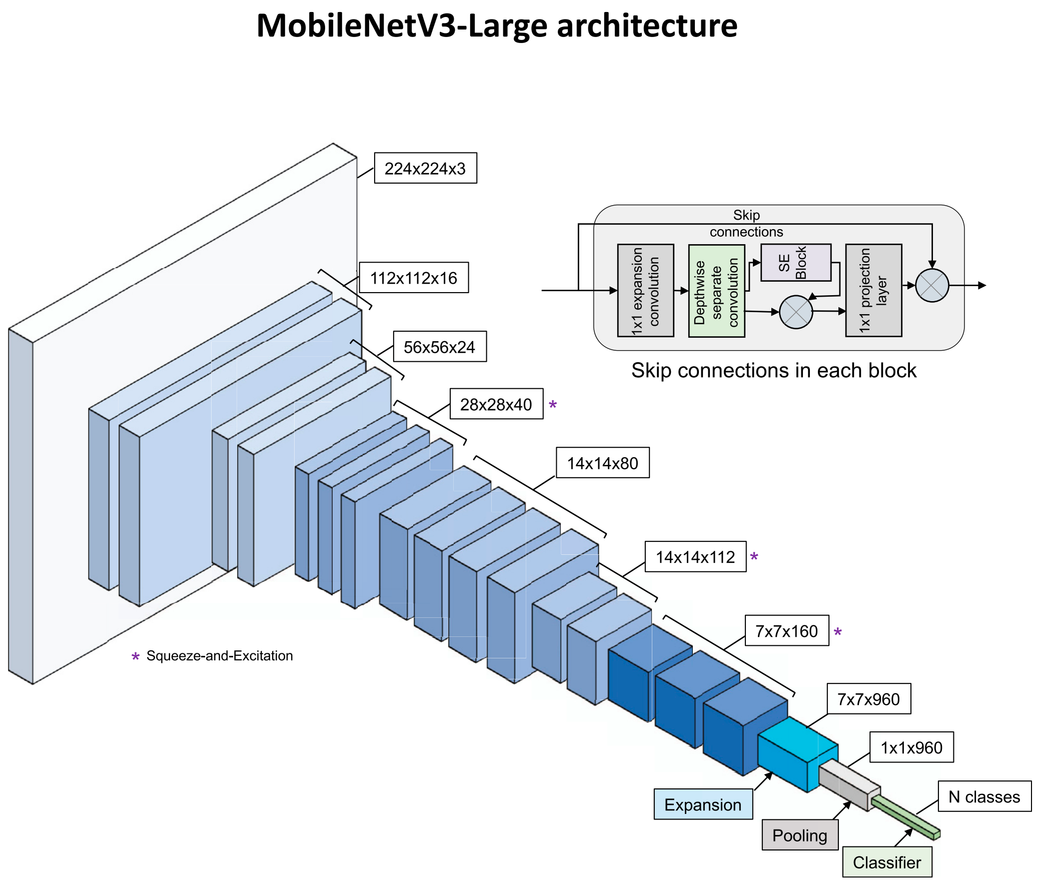


FIGURE 2
MobileNetV3-Large architecture. Main diagram shows feature map progression from 224×224×3 input through inverted residual blocks to N-class output. Blocks marked (*) use Squeeze-and-Excitation modules. Inset shows internal structure of an inverted residual block with skip connections.

segmentations for the anterior and posterior leaflets (Labels 1 and 2). To enhance the dataset for comprehensive valve analysis, manual annulus contour segmentations were added (Label 3).

2.2.1.1 Annulus annotation enhancement

Manual annulus contours were created using 3D Slicer by placing control points along the mitral annulus in 3D space using the SlicerHeart analysis module. These control points were exported as JSON markup files containing world coordinates. To convert these sparse control points into volumetric segmentations, an automated spline-based approach was developed: 1. control points were fitted with a smooth 3D B-spline using scipy's splprep function with zero smoothing factor, 2. the spline was evaluated at 100 equally spaced parameter values to create a dense point cloud, 3. a cylindrical tube with 1.5 mm radius was generated around the spline using VTK libraries, and 4. the tube was voxelized into the original image space using VTK's vtkPolyDataToImageStencil method.

To ensure anatomically consistent segmentations, we applied morphological post-processing, including connected component analysis, to retain only the most significant component with the highest mean z-coordinate, effectively removing spurious disconnected regions.

2.2.2 Deep learning model training

2.2.2.1 Architecture and framework

We employed MONAI's Auto3DSeg framework, which automatically generates and optimizes multiple 3D segmentation architectures for medical imaging applications. The framework was configured to use SegResNet as the primary architecture, a 3D residual U-Net variant designed explicitly for volumetric medical image segmentation.

2.2.2.2 Training configuration

The enhanced MVSEG2023 dataset was divided into 5-fold cross-validation splits with random stratification (seed = 42). Training data organization followed MONAI's standard format. The Auto3DSeg pipeline automatically handled data preprocessing, augmentation strategies, and hyperparameter optimization. The training was set up following the configuration of the MVSEG2023 challenge winner, with the specified modality being magnetic resonance imaging.

2.2.2.3 Model ensemble

The Auto3DSeg pipeline trains a model for each fold and enables ensemble prediction by averaging the outputs of all models, which improves performance at the expense of longer inference time. For prediction, models from all five folds were used to obtain the best segmentation.

2.2.3 Cardiac phase detection and temporal analysis

2.2.3.1 End-systole identification

To identify the optimal cardiac phase for valve area measurement, an automated mid-diastole detection algorithm based on temporal analysis of segmented structures was developed. For each frame in the 4D TEE sequences, we performed the following analysis pipeline:

a. Annulus Skeletonization: The segmented annulus (Label 3) was skeletonized using 3D morphological thinning to extract its centerline representation.
b. 3D Point Ordering: Skeleton points were spatially ordered using a nearest-neighbor approach with orientation constraints to prevent backtracking, ensuring anatomically consistent point sequences along the annulus perimeter.
c. Plane Fitting: Principal Component Analysis (PCA) was applied to the ordered annulus points to determine the best-fitting plane, with the plane normal defined as the eigenvector corresponding to the smallest eigenvalue.
d. Area Calculation: All segmented structures (leaflets and annulus) were projected onto this optimal plane, and areas were calculated using pixel-based methods with appropriate spatial calibration.

2.2.3.2 Temporal peak detection

Mid-diastole was identified as the frame exhibiting maximum effective valve area, corresponding to the point of maximum valve opening during the cardiac cycle.

2.2.4 Geometric analysis and area quantification

2.2.4.1 Valve plane projection

The projection process involved: I. determination of the optimal valve plane using PCA analysis of annulus centerline points, II. orthogonal projection of all segmented voxels onto this plane, III. conversion to 2D coordinates using orthonormal basis vectors derived from the plane normal, and IV. creation of high-resolution 2D images with pixel sizes calculated from the original voxel spacing and projection angle.

2.2.4.2 Effective orifice area

Functional valve opening area was determined through morphological analysis of the projected segmentation, using flood-fill algorithms to identify the central opening region.

2.2.4.3 Spatial calibration

All measurements were performed in physical units (mm^2) using voxel spacing information extracted from DICOM headers. The projection method accounted for oblique viewing angles by adjusting pixel sizes based on the angle between the valve plane and the image coordinate system.

2.2.5 Data selection process

4D TEE volumes from both M-TEER (Mitraclip) and surgical MVR patients were included. Only TEE exams with available 3D acquisition, performed at the Montreal Heart Institute starting from 1 March 2024, were used, as raw data extraction was only enabled at the end of February 2024. Each TEE examination was assigned an internal code corresponding to its specific exam type in the institutional database. Only TEE exams performed within 12 months before the M-TEER or MVR were used, provided that the physician's clinical report with mitral valve analysis and MVA measurement, as performed by a cardiologist, was available. Intraprocedural TEE exams and exams from patients with prior MV procedures were excluded.

2.2.6 Data extraction process

4D TEE volumes meeting the selection criteria were identified through a series of internal SQL scripts executed across complementary databases, including a report database and an exam type database. The identified 4D TEE DICOMs were then transferred to an internal research server using pydicom-batch (https://github.com/MHI-AI-CoreLab/pydicom-batch).

2.2.7 Data cleaning process

An expert cardiologist performed a manual curation process to identify TEE exams in which the mitral valve was acquired and deemed suitable for analysis.

2.3 2D-TEE-based automatic MV segmentation: scallop-level analysis

For this part, we chose a U-Net architecture (Ronneberger et al., 2015), which is a fully convolutional network consisting of an encoder and a decoder. The model accepts 3-channel ultrasound images $x \in R3 \times 256 \times 256$ as input and outputs four results: a final segmentation map $\phi(x) \in [0,1]11 \times 256 \times 256$ and three deep supervision outputs $\psi1(x) \in [0,1]11 \times 128 \times 128$, $\psi2(x) \in [0,1]11 \times 64 \times 64$, and $\psi3(x) \in [0,1]11 \times 32 \times 32$. Each of the 11 output channels corresponds to one of the following anatomical structures: the six scallops of the mitral valve (A1, A2, A3 for the anterior leaflet, matching P1, P2, P3 respectively for the posterior leaflet), the anterior and posterior papillary muscles, the chordae, the annulus, and the background. We modified the original U-Net architecture to suit our task better, as shown in Figure 3.

2.3.1 Model architecture

2.3.1.1 Encoder

The original encoder has been replaced with a ResNet34-based backbone (Kaiming et al., 2015). Three types of blocks were used: the first is a wide convolution (7 × 7) followed by batch normalization and a ReLU activation function; the second is a residual downsampling block that reduces spatial resolution by a factor of 2 using a 3 × 3 convolution with stride 2; the third is a standard residual block with 3 × 3 convolutions. Both residual block types use skip connections to improve gradient flow during training.

2.3.1.2 Decoder

In the decoder, bilinear up-sampling was used instead of transposed convolutions to reduce checkerboard artifacts (Odena et al., 2016). The rest of the decoder followed the original U-Net structure, with skip connections passed through attention gates,

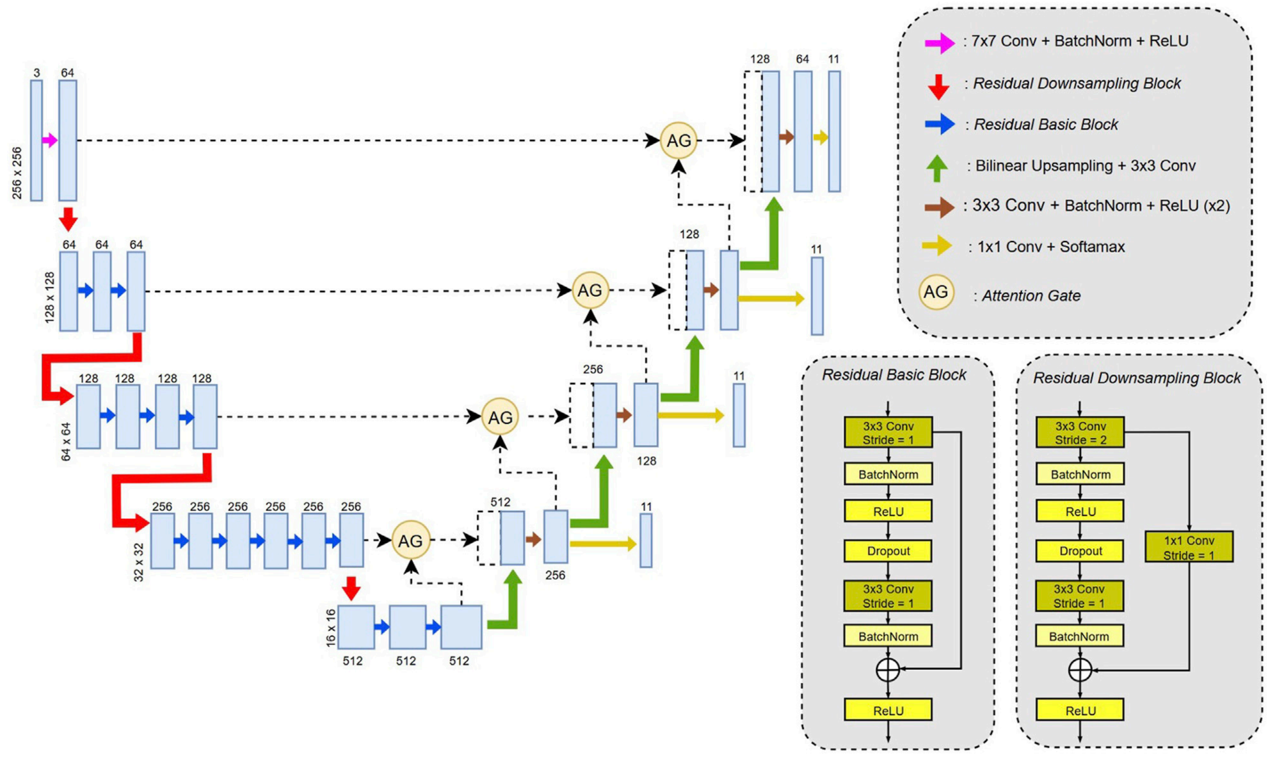


FIGURE 3
U-Net modified architecture model used for the segmentation of mitral valve leaflets scallops.

concatenated, and followed by two 3 × 3 convolutions, batch normalization, and ReLU. Each decoder stage also includes a final 1 × 1 convolution and a SoftMax activation to produce intermediate outputs for deep supervision.

2.3.1.3 Attention gates

To improve the focus on the mitral valve and reduce the segmentation of non-relevant muscular structures, we integrated attention gates. These modules, introduced by Oktay et al. (Schlemper et al., 2019), highlight specific regions of interest during training. Each attention gate takes as input a skip connection x from the encoder and a gating signal g from the corresponding decoder stage and returns a refined feature map with the same dimensionality as x (Supplementary Figure 1).

2.3.2 Loss functions

To optimize the segmentation network, we used a combination of Dice loss and Focal loss, which are well-suited for highly imbalanced multiclass segmentation tasks.

2.3.2.1 Dice loss

The generalized Dice loss is defined as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{1}{|C'|}\sum_{c\in C'} \frac{2\sum_{i=1}^{N} p_{i,c}\, g_{i,c} + \varepsilon}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} g_{i,c} + \varepsilon}$$

where C' represents the set of classes present in the image, N is the number of pixels in the image, $p_{i,c}$ is the predicted probability for pixel i belonging to class c, $g_{i,c}$ is the corresponding ground truth (one-hot encoded), and ε is a small constant to avoid division by zero.

2.3.2.2 Focal loss

To further address class imbalance and focus training on hard-to-classify pixels, we also employed the Focal loss (Lin et al., 2018), defined as:

$$\mathcal{L}_{Focal} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in C}\left(1 - p_{i,c}\right)^{\gamma} g_{i,c}\log\left(p_{i,c}\right)$$

where C denotes the set of all classes (not just those in the image), and γ is the focusing parameter (set to 2 in this study) that decreases the relative loss contribution of well-classified pixels.

2.3.2.3 Combined loss

The final training objective for a single output is a simple combination of the two losses:

$$\mathcal{L} = \frac{\mathcal{L}_{Dice} + \mathcal{L}_{Focal}}{2}$$

2.3.3 Loss functions with deep supervision

The final loss is applied not only to the network's final output but also to intermediate outputs. This deep supervision strategy, introduced in (Lee et al., 2014), encourages lower decoder layers to focus on relevant regions early in the network.

TABLE 1 Legend of labels used for mitral valve leaflets scallops and sub-apparatus structures annotations.

Anatomical structure	Label
A1	a_1
A2	a_2
A3	a_3
P1	p_1
P2	p_2
P3	p_3
MV chordae	Chordae
MV annulus	Annulus
AL papillary muscle	Papillary_anterior
PM papillary muscle	Papillary_posterior

AL, anterolateral; MV, mitral valve; PM, posteromedial.

Let $\phi(x) \in [0,1]^{C\times H\times W}$ be the final output, and $\{\psi_k(x)\}^3_{k=1}$ be the three intermediate deep supervision outputs. The total loss is then computed as:

$$\mathcal{L}_{total} = \mathcal{L}\left(\phi(x), G\right) + \sum_{k=1}^{3} \mathcal{L}\left(\psi_k(x), G^{(k)}\right)$$

Since the intermediate outputs from the decoder have lower spatial resolution than the input image, the corresponding ground truth masks need to be downsampled to match each output size before calculating the loss.

Let $G \in \{0,1\}^{C\times H\times W}$ be the original one-hot encoded ground truth mask. For each deep supervision output $\psi_k(x) \in [0,1]^{C\times Hk\times Wk}$, the ground truth is downsampled using nearest-neighbor interpolation:

$$G^{(k)} = \mathrm{Downsample}\left(G, H_k, W_k\right)$$

where (H_k, W_k) are the height and width of the k-th intermediate output. Nearest-neighbor interpolation preserves the discrete class labels, ensuring accurate loss computation for each class. The total loss is then computed by comparing each output $\psi_k(x)$ to its corresponding downsampled ground truth $G^{(k)}$.

2.3.4 Optimization and training

The network was trained with the Adam optimizer, starting with a learning rate of 1×10^{-4} and a batch size of 24. To prevent overfitting, a weight decay of 1×10^{-6} and dropout with a rate of 0.2 in the encoder layers were used. Data augmentation was extensively employed to boost the diversity of the training set, including random rotations (up to 45°), translations, and scaling (between 0.75 and 1.25). These augmentations were applied during training in real-time to improve the model's ability to generalize.

2.3.5 Dataset and preprocessing

The dataset included 992 TEE images from 77 different patients who underwent a Mitraclip procedure, focusing on the mitral valve, with 11 segmentation classes representing various anatomical structures, with corresponding labels (Table 1). Only 2D images were analyzed. A total of 2,200 ground truth annotations were made by a physician on the Labelbox platform. The data were split into 80% for training (N = 821) and 20% for validation (N = 171), ensuring that all images from the same patient remained in the same subset to prevent data leakage.

Before training, all images were normalized to have zero mean and unit variance. Both images and their corresponding masks were resized to 256 × 256 pixels when needed.

2.3.6 Evaluation

The segmentation performance was assessed using multiple metrics, including Dice coefficient, precision, recall, and false positive rate (Supplementary Figure 2).

The Dice score is a widely used metric to assess segmentation performance by quantifying the spatial overlap between the predicted segmentation and the ground truth. It is defined as:

$$\mathrm{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

where A represents the predicted segmentation and B the reference segmentation.

A Dice score of 1 indicates perfect agreement, whereas a score of 0 indicates no overlap.

This metric is particularly well-suited for medical image analysis because it remains robust to class imbalance (e.g., small anatomical structures occupying only a fraction of the image) and has become a standard benchmark for evaluating segmentation algorithms.

All experiments were conducted in PyTorch 2.6 and trained on an NVIDIA RTX A600 GPU.

3 Results

3.1 Automatic classification of 2D- TTE and TEE images: quality views assessment

ECHO-PREP first algorithm was trained to identify key diagnostic views in TTE and diagnostic and procedural views in TEE examinations (algorithm 1, Figure 1) to determine the level of image quality needed for an M-TEER, based on a dataset of 530 TTE and 800 TEE pre-M-TEER acquisitions, respectively. The total number of TTE and TEE analyzed frames was 58.749 and 52.058, respectively. The dataset distribution of TTE and TEE diagnostic views is shown in Supplementary Tables 1, 2, respectively. The algorithm successfully determined whether the TTE was of good quality with a frame-level accuracy of 95.7% (Figure 4A) and performed well in view classification (Figure 4B). For TEE, the algorithm produced similar results, accurately identifying whether TEE views were of sufficient quality for patient eligibility and procedural guidance of M-TEER in 91% of cases, with a high overall accuracy for TEE view classifications, as demonstrated by AUC values (Figure 5).

3.2 4D-TEE-based automatic MV segmentation and MV area measurements

ECHO-PREP second algorithm was trained on a total of 135 TEE 4D volumes from the MVSEG2023 dataset, with a

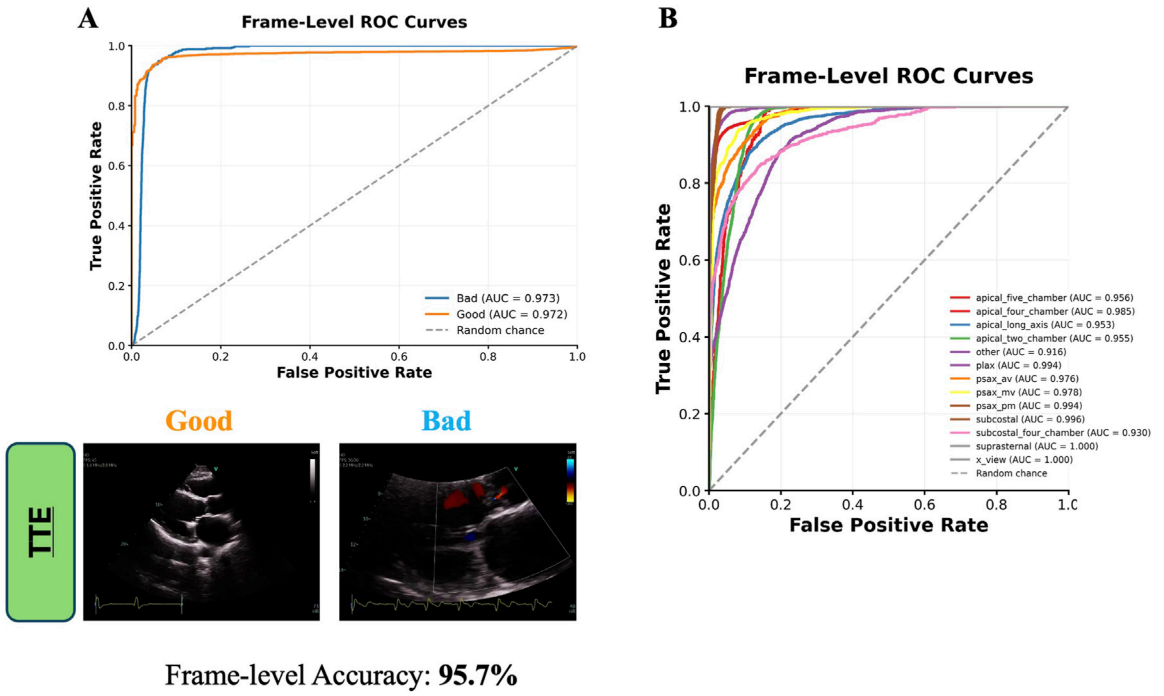


FIGURE 4
Frame-Level ROC curves for TTE Quality Image assessment **(A)** and Views classification **(B)**. TTE images are direct outputs from the algorithm to illustrate the classification of "good" versus "bad" views. AUC = area under curve; av = aortic valve; mv = mitral valve; pm = papillary muscle; TTE = transthoracic echocardiography; plax = parasternal long axis; psax = parasternal short axis; ROC = receiver operating characteristic curve.

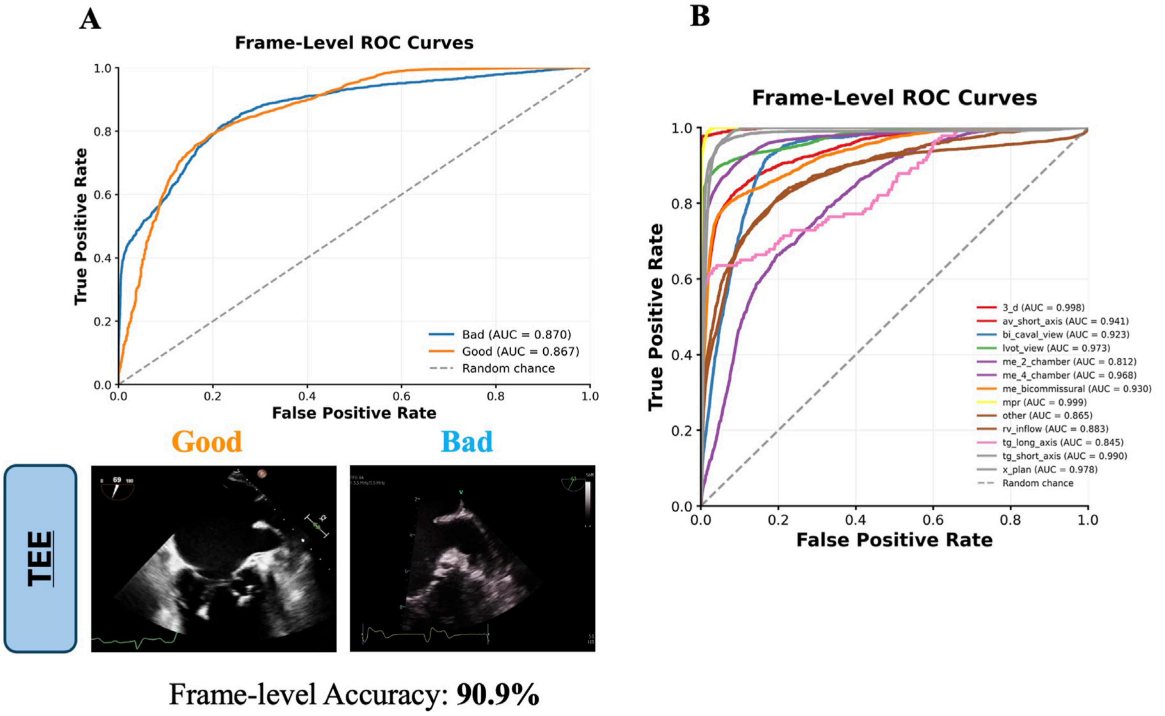


FIGURE 5
Frame-Level ROC curves for TEE Quality Image assessment **(A)** and Views classification **(B)**. TEE images are direct outputs from the algorithm to illustrate the classification of "good" versus "bad" views. AUC = area under curve; av = aortic valve; me = mid-esophageal; mpr = multiplanar reconstruction; rv = right ventricle; tg = transgastric; lvot = left ventricular outflow tract; 3_d = three-dimensional; TEE = transesophageal echocardiography; ROC = receiver operating characteristic curve.

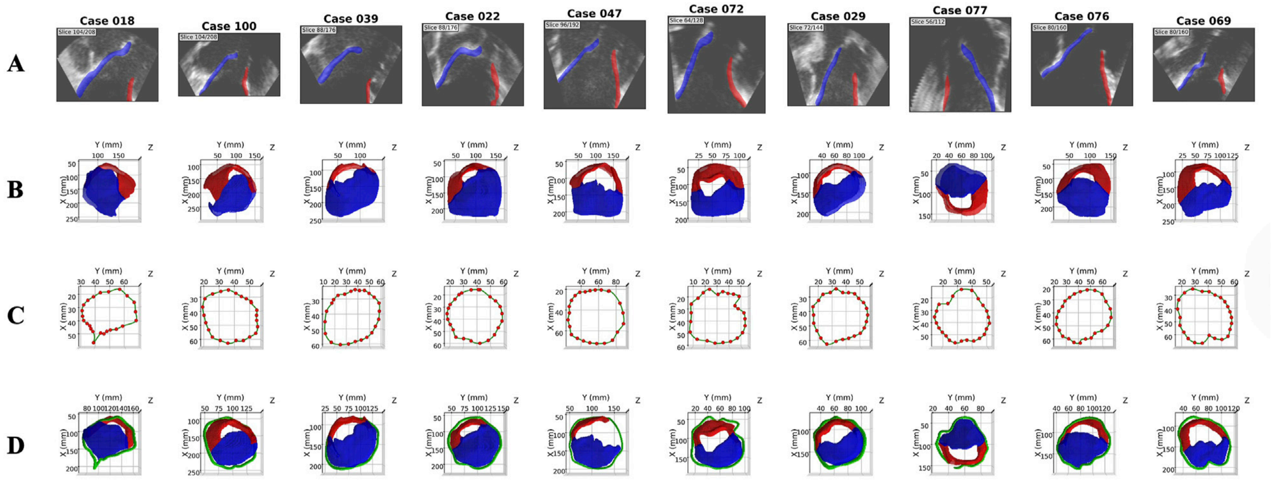


FIGURE 6
Enhanced Annotation from MVSEG 2023: 10 representative cases. Rows **(A)** 2D sagittal slice showing original echocardiography with leaflet segmentation overlay. **(B)** 3D superior view of MVSEG 2023 baseline dataset (leaflets only). **(C)** Manual annulus annotation with control points and B-spline fitting. **(D)** Final enhanced dataset with complete mitral valve (leaflets + annulus). 2D = two-dimensional; 3D = three-dimensional.

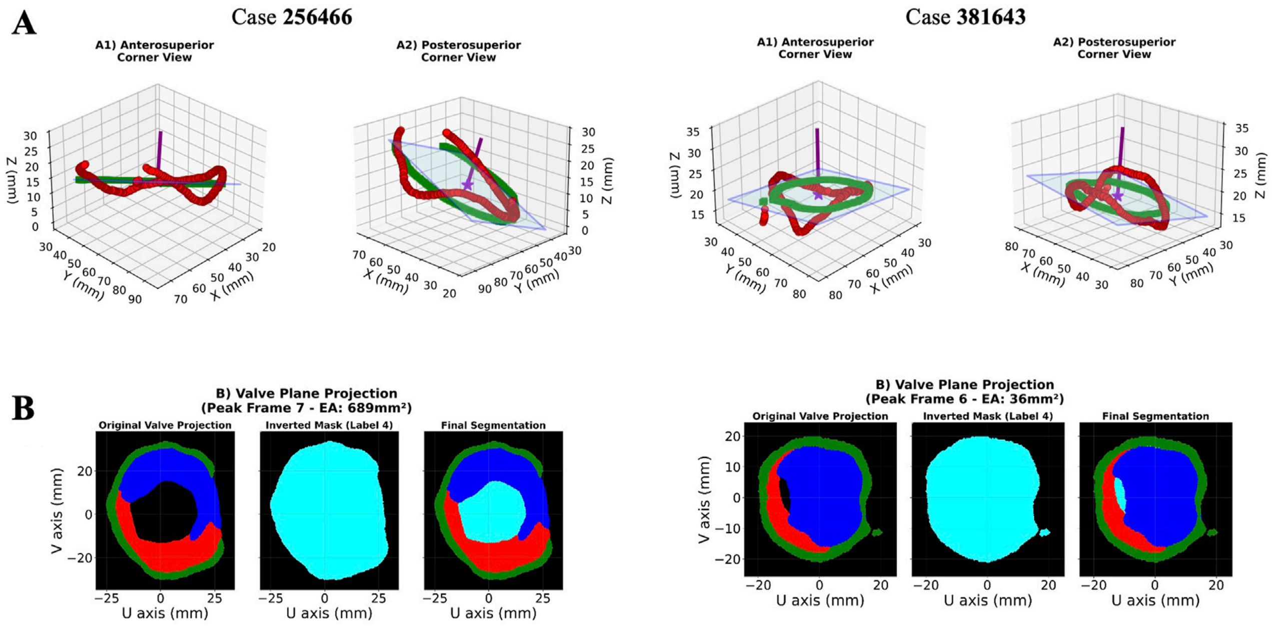

A
Case 256466
A1) Anterosuperior Corner View
A2) Posterosuperior Corner View
Case 381643
A1) Anterosuperior Corner View
A2) Posterosuperior Corner View
Z (mm)
Y (mm)
X (mm)
B
B) Valve Plane Projection (Peak Frame 7 - EA: 689mm²)
Original Valve Projection
Inverted Mask (Label 4)
Final Segmentation
B) Valve Plane Projection (Peak Frame 6 - EA: 36mm²)
Original Valve Projection
Inverted Mask (Label 4)
Final Segmentation
V axis (mm)
U axis (mm)

FIGURE 7
PCA-Based Optimal Plane Projection for 3D Mitral Valve Quantification: 2 representative cases. Rows **(A)** 3D visualization of annulus skeleton points (red) and their projection onto the PCA-derived optimal plane (green squares) from anterosuperior (A1) and posterosuperior (A2) viewpoints. The semi-transparent blue plane represents the best-fitting 2D projection surface with a normal vector (purple). **(B)** Valve plane projection showing effective area measurement from the peak cardiac frame with color-coded anatomical structures (posterior leaflet = red, anterior leaflet = blue, annulus = green, functional area = light blue). 2D = two-dimensional; 3D = three-dimensional; PCA = Principal component analysis.

70%–30% split for training and validation, respectively. Segmentation of relevant anatomical features, including the mitral anterior and posterior leaflets and annulus, was performed using the MONAI Auto3DSeg software after identifying the mid-diastole frame (Figure 6). A logical stepwise understanding from anatomy to segmentation can be derived from Figure 7, which effectively illustrates the use of a PCA-based optimal plane and segmentation pipeline for valve analysis. The figure clearly contrasts two cases (256466 vs. 381643) using 3D visualizations (on top) and 2D valve plane projections (below) at the peak frame. In Case 256466 (left panel), the effective area (EA) is much larger (689 mm^2), and the valve appears more symmetric and complete

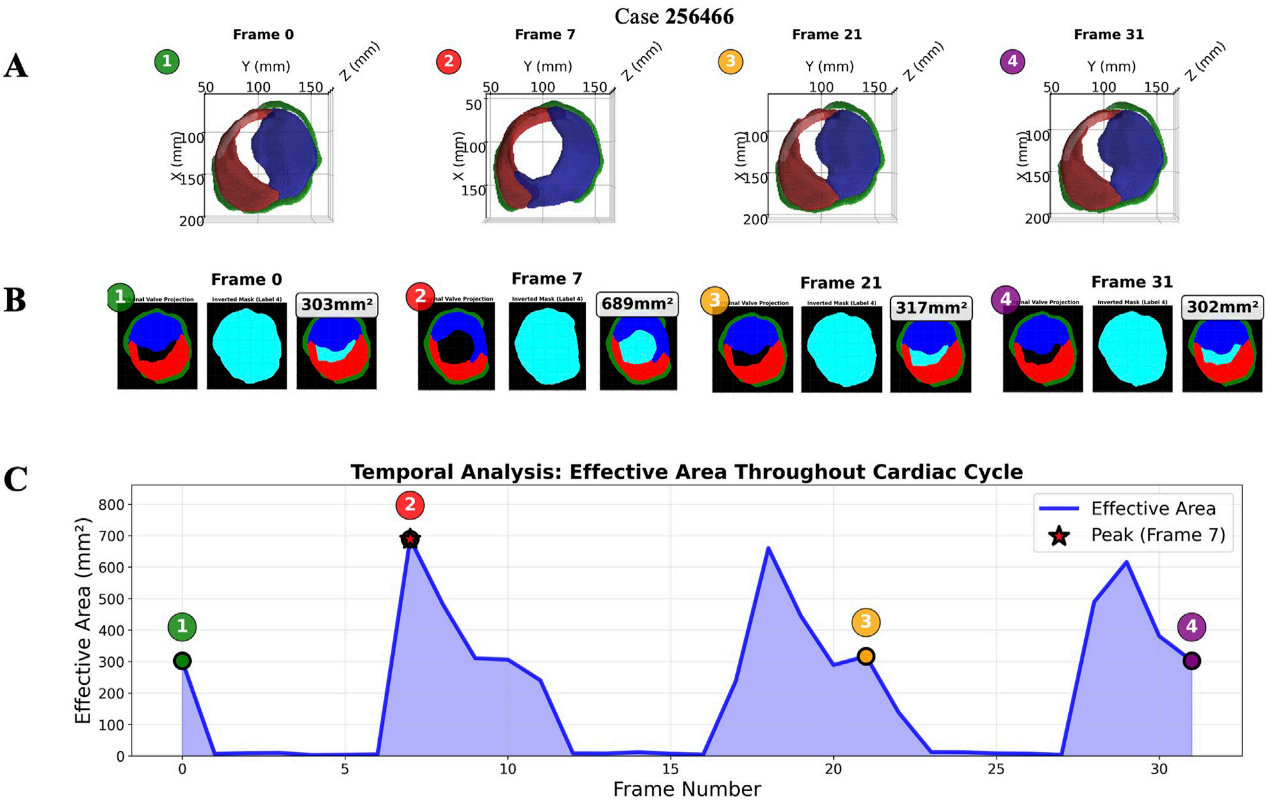


FIGURE 8
Temporal analysis showing valve geometry evolution across the cardiac cycle: representative case. Rows **(A)** 3D-views and **(B)** 2D-projections at four time points (posterior leaflet = red, anterior leaflet = blue, annulus = green, functional area = light blue). **(C)** Effective area curve with peak detection (red star) and frame markers (numbered circles). 2D = two-dimensional; 3D = three-dimensional.

in both 3D and projection views. Segmentation appears clean, with well-demarcated leaflets. In case 381643 (right panel), the EA is significantly smaller (36 mm^2), indicating severe restriction. The 3D views show distorted, irregular geometry, and the projection and segmentation reveal significant leaflet malcoaptation or incomplete opening. Combining 3D multi-angle views with 2D projection and segmentation provides complementary perspectives: the segmentation masks (original, inverted, final) offer transparency into the algorithm's steps and show good alignment between the quantitative data (EA values) and visual impression. The contrast with case 256466 (EA 689 mm^2) highlights the method's robustness in capturing extreme phenotypes. The added value of a comprehensive temporal analysis of mitral valve dynamics is demonstrated in Figure 8. Instead of static geometry, it captures the valve's physiological motion and functional variability. The top row shows 3D superior views at four timepoints across the cardiac cycle (start-peak-mid-end), with valve structures clearly delineated. It demonstrates valve opening dynamics, from partial opening at Frame 0 to maximal separation at Frame 7. The middle row shows 2D valve projections at the same key frames. Effective orifice area (EA) values are: start: 303 mm^2, peak: 689 mm^2, mid: 317 mm^2, end: 302 mm^2. This visualization complements the 3D view by quantifying leaflet separation. The bottom panel displays the temporal analysis graph with EA plotted across all 32 frames: peak EA occurs at Frame 7 (689 mm^2). The cycle demonstrates typical dynamic variation, with large fluctuations between systolic closure and diastolic opening (mean EA: 198 mm^2; range: 3–689 mm^2). This patient (case 256466) shows normal dynamic opening and closure patterns, with a large peak EA, consistent with preserved mitral valve function. The data highlights the algorithm's ability to continuously track valve dynamics throughout the cardiac cycle, not just at isolated frames. The temporal profile offers a clear functional fingerprint that could distinguish healthy from pathological valves.

The validation of the algorithm for quantifying the mitral EA involved analyzing a total of 221 TEE 4D volumes performed at our center as part of a pre-procedural assessment of mitral regurgitation. Images were divided into two groups: those from patients who later underwent M-TEER with Mitraclip (121 4D volumes from 30 patients) and those from patients who had surgical mitral valve replacement (100 4D volumes from 18 patients). A physician reviewed the available images from the center's database and preliminarily excluded videos with unsuitable views for calculating the MVA, such as poor image quality, artifacts, or the presence of a previous surgical prosthesis or valve ring. The validation of the AI-predicted MVA quantification was performed by comparing it to the gold standard of manual measurements from physician clinical reports. In Figure 9A, the scatter plot shows a strong positive correlation (Pearson's R = 0.84) between the MVA measurements from clinical reports and those predicted by our algorithm. The correlation is statistically significant ($p < 0.001$), demonstrating excellent agreement between the AI and human expert measurements.

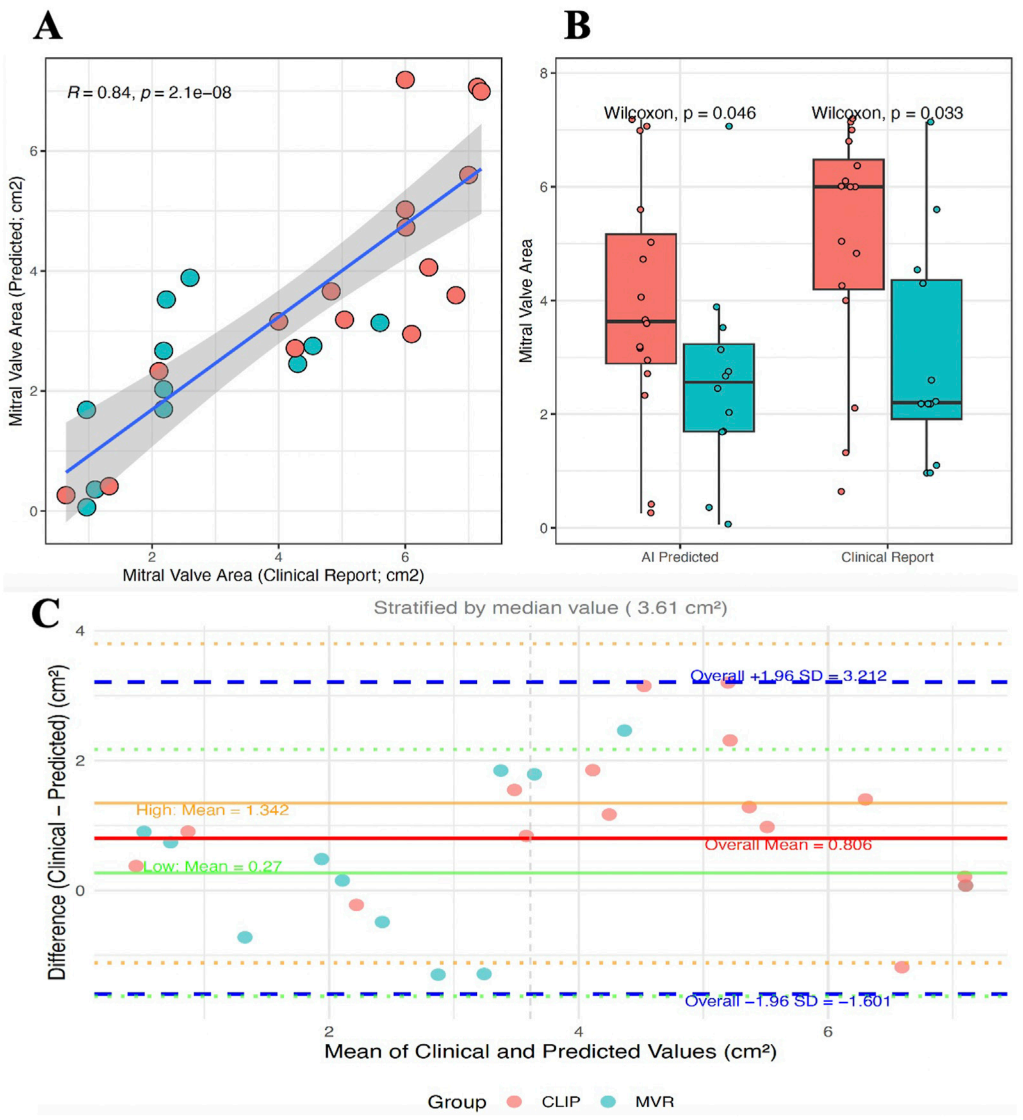


FIGURE 9
Validation of AI-Based Mitral Valve Area Quantification Against Clinical Reports. **(A)** Correlation between clinical reports and AI predictions showing strong agreement (R = 0.84). **(B)** Distribution comparison between measurement methods for Mitraclip (red) and surgical MVR (blue) patient groups. Wilcoxon tests show statistical significance of group differences. **(C)** Bland–Altman plot assessing agreement between clinical and AI-predicted mitral valve area, stratified by median value (3.61 cm^2). The overall mean difference was 0.806 cm^2 (solid red line). Agreement was better in the low-value group (mean difference 0.27 cm^2, solid green line) compared to the high-value group (mean difference 1.34 cm^2, solid orange line). Dashed blue lines indicate the 95% limits of agreement (mean ± 1.96 SD). MVR = mitral valve replacement; SD = standard deviation.

Further validation was conducted through group discrimination. Since our center handles high volumes and specializes in mitral repair with a high success rate across various mitral regurgitation scenarios, it was assumed that patients who ultimately underwent surgical MVR were more likely to have a non-repairable valve due to factors like a stenotic or restrictive valve with a smaller MVA, after excluding patients with endocarditis or prior valve procedures. In Figure 9B, the box plots compare the distribution of MVA measurements between two patient groups: Mitraclip (red) and surgical MVR (blue). As expected, the MVR group exhibits a significantly smaller mitral valve area. The difference between the two groups is statistically significant for both measurement methods (clinical reports: Wilcoxon p = 0.033; AI predictions: Wilcoxon p = 0.046). A Bland-Altman analysis was performed to complement the correlation and illustrate agreement between AI-derived and physician-reported mitral valve area (Figure 9C).

Even with some limitations, this remains an important validation step, showing that the algorithm not only aligns with clinical reports on individual measurements but also keeps the clinically relevant physiological differences between different patient groups. The higher significance (p = 0.033) in the clinical reports is expected, as they are the reference standard.

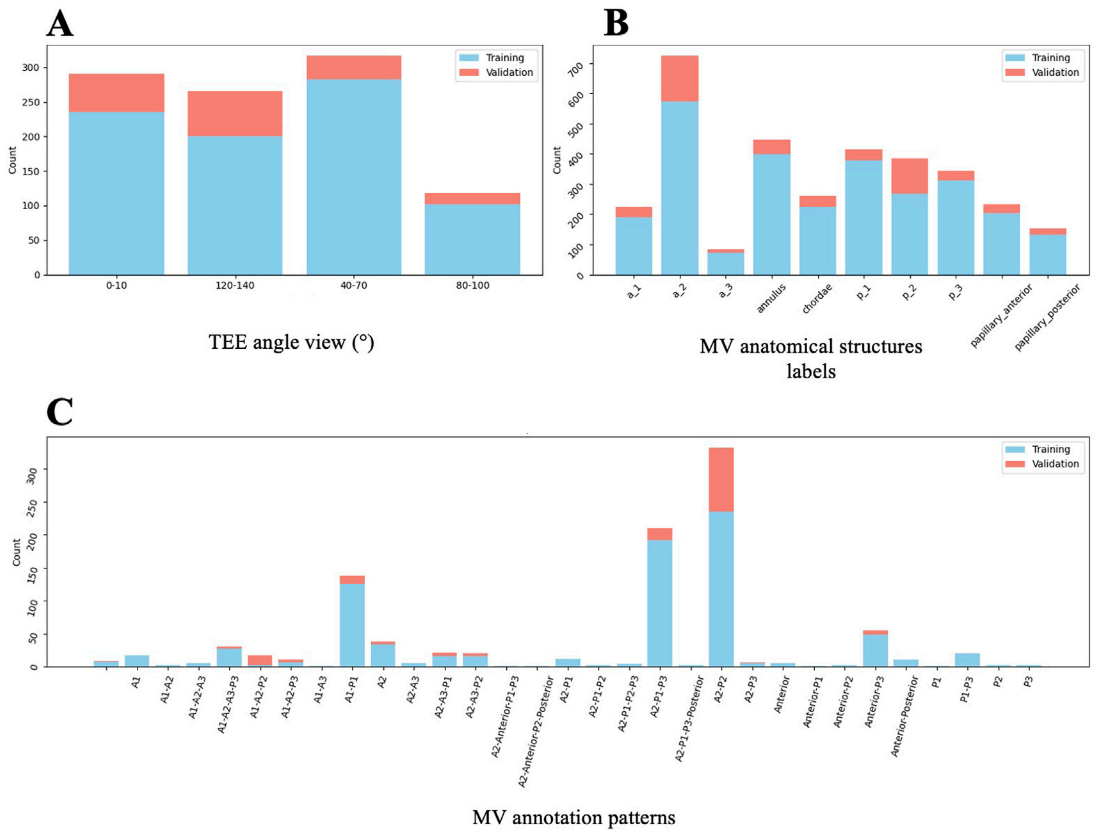


FIGURE 10
Dataset overview for 2D TEE Segmentation Model (Algorithm 3). Distribution of TEE angle views **(A)**, anatomical MV structures **(B)**, and MV annotation patterns **(C)** from the analysed image dataset, divided into training and validation subsets. 2D = two-dimensional; MV = mitral valve; TEE = transoesophageal echocardiography.

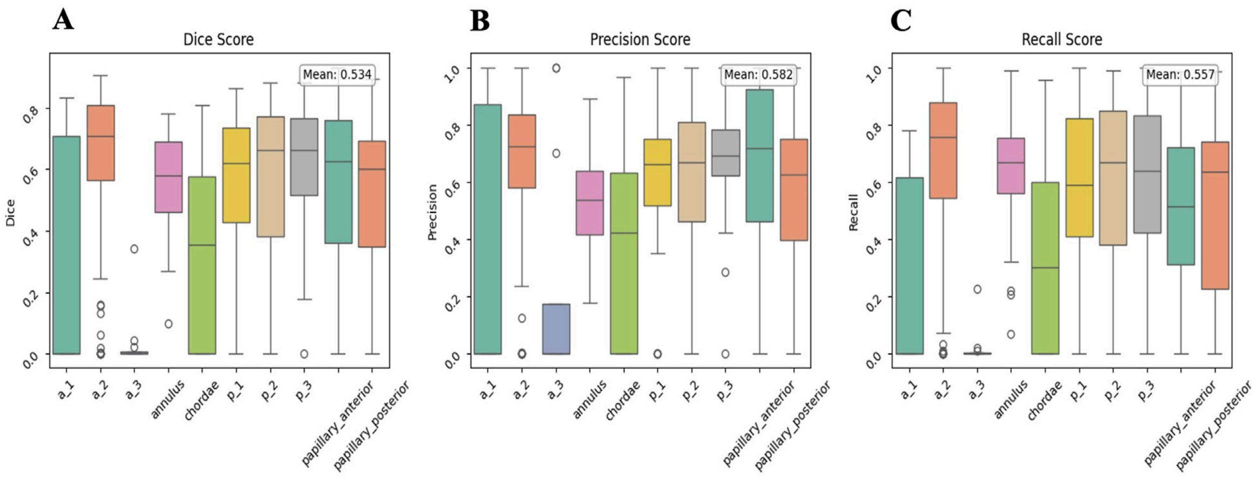


FIGURE 11
Global validation results for the 2D TEE Segmentation Model (Algorithm n.3). Mean Dice **(A)**, Precision **(B)**, and Recall **(C)** scores are displayed for all MV labels. Annotation labels correspond to the respective MV structures as previously defined. 2D = two-dimensional; MV = mitral valve; TEE = transoesophageal echocardiography.

TABLE 2 Segmentation performance metrics for mitral valve structures annotations, as identified with their labels.

Label	Dice	Precision	Recall	FPR
a_1	0.289 ± 0.352	0.369 ± 0.432	0.253 ± 0.318	0.000 ± 0.001
a_2	0.640 ± 0.236	0.673 ± 0.233	0.671 ± 0.274	0.001 ± 0.001
a_3	0.034 ± 0.098	0.225 ± 0.414	0.021 ± 0.065	0.000 ± 0.000
Annulus	0.557 ± 0.162	0.520 ± 0.168	0.639 ± 0.193	0.002 ± 0.001
Chordae	0.342 ± 0.292	0.393 ± 0.313	0.344 ± 0.319	0.001 ± 0.001
p_1	0.548 ± 0.266	0.599 ± 0.272	0.566 ± 0.310	0.001 ± 0.001
p_2	0.550 ± 0.281	0.605 ± 0.287	0.583 ± 0.324	0.001 ± 0.001
p_3	0.607 ± 0.205	0.682 ± 0.186	0.609 ± 0.272	0.000 ± 0.001
Papillary anterior	0.546 ± 0.247	0.647 ± 0.295	0.526 ± 0.269	0.002 ± 0.002
Papillary posterior	0.492 ± 0.297	0.537 ± 0.329	0.514 ± 0.333	0.002 ± 0.003

Values are Mean ± Standard Deviation.
FPR, false positive risk.

3.3 2D-TEE-based automatic MV segmentation: scallop-level analysis

A dataset overview, showing the distribution of TEE angle views and annotations of the MV structures during segmentation, is presented in Figure 10. The most common anatomical patterns were A1-P1, A2-P2, and A2-P1-P3 (Figure 10C). These were mainly mid- esophageal (ME) five- (5C) and four-chamber (4C) views, ME long axis, and ME commissural views, primarily used to evaluate the MV, especially in pre-procedural M-TEER assessment. The overall validation results are shown in Figure 11, with a mean Dice score of 0.534 across the entire validation dataset. Individual Dice scores for each MV structure annotation are listed in Table 2. As expected, performance is slightly lower in commissural regions such as A3 and A1. When analyzing these results, it is essential to note that if the segmentation involves a small anatomical structure, such as a short or retracted posterior mitral leaflet, which is common in functional mitral regurgitation, the metric results, particularly the Dice score, may be misleading.

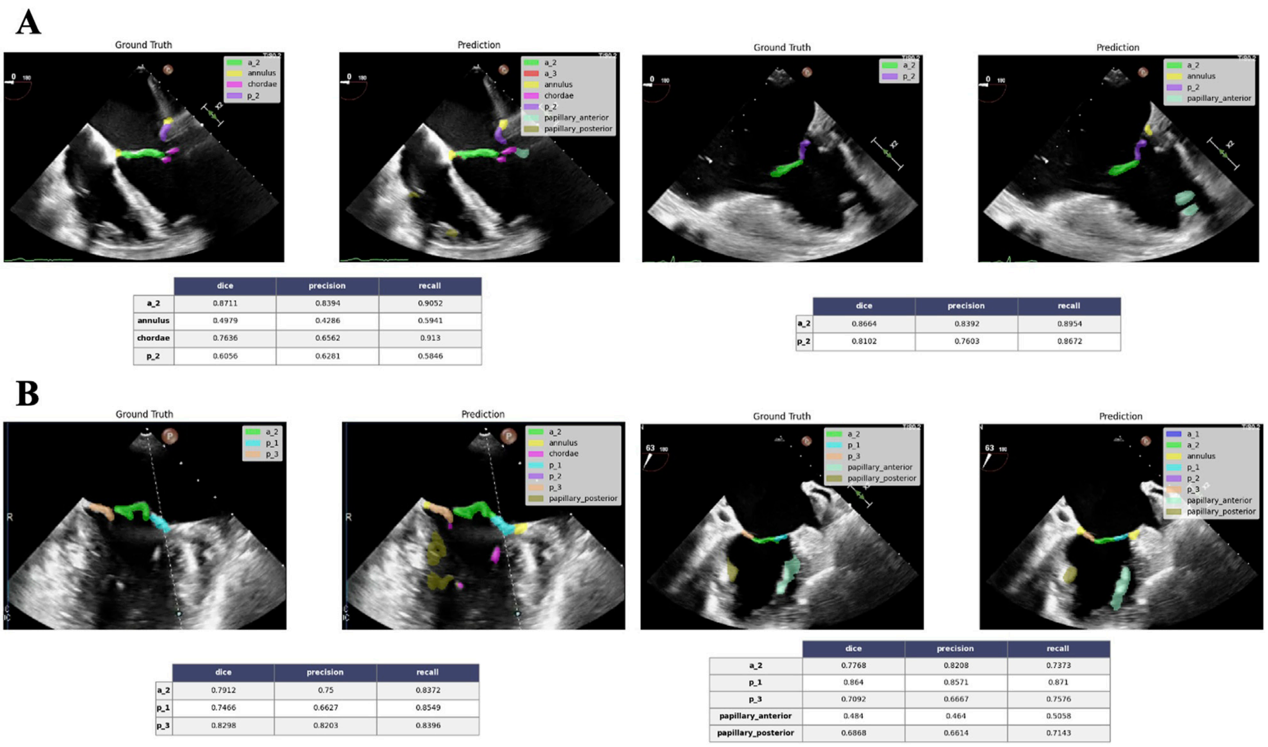
Generally, a Dice score above 0.7 is considered a good (visually) result. Regarding leaflet scallop segmentation, the ME Long Axis views and ME Commissural views are typically segmented very well by the neural network, with the P3-A2-P1 sequences being highly represented in the dataset. An example is illustrated in Figure 12. However, the results are less accurate for other views. The poorer outcomes mainly stem from the ME 4 C and ME 5 C views, which often confuse the A2- P2 and A1- P1 sequences. The segmentation of the annulus produces quite good results. It is important to note that in 2D images, annulus annotation is very small and can be biased when calcifications are absent, making the Dice score very sensitive. Even in images without annulus annotations, the neural network seems capable of detecting the annulus correctly. Since the papillary muscles and mitral chordae are located within the ventricle (mostly represented by dark pixels), the network is highly sensitive to noise and bright areas within the ventricle, which can lead to confusion with these structures. As a result, the outcomes for the papillary muscles- and even more so for the chordae- are not optimal.

4 Discussion

In this study, we present ECHO-PREP, an integrated multi-stage deep learning framework for pre-procedural mitral valve assessment in candidates for M-TEER procedure. Our approach encompasses three complementary modules: automated quality assessment of echocardiographic views, 4D segmentation with functional valve area quantification, and 2D scallop-level analysis of valve anatomy. Together, these components aim to address the current challenges of variability, subjectivity, and inefficiency in echocardiographic interpretation for M-TEER planning.

The first significant finding was the high performance of the quality assessment algorithm for both TTE and TEE images, achieving frame-level accuracies above 90%. This step, although often overlooked, is clinically critical: poor-quality imaging is a common reason for inconclusive evaluations and may delay intervention. By introducing automation at this stage, our framework could improve workflow efficiency and ensure that downstream analyses are only performed on diagnostically valid inputs. The second major result was the successful implementation of 4D TEE-based segmentation with automated mitral valve area (MVA) quantification. The algorithm showed strong correlation with physicians' clinical reports ($R = 0.84$, $p < 0.001$), confirming its reliability for valve sizing and functional assessment. Notably, the system not only reproduced static area measurements but also captured temporal variations of valve opening and closure, offering a dynamic fingerprint of valve physiology. This longitudinal perspective may become a powerful discriminator between repairable and non-repairable valves, as suggested by the observed differences between patients undergoing M-TEER and those treated with surgical valve replacement.

However, the validation of MVA quantification across patient groups relied on the assumption that all surgically replaced valves (MVR) were non-repairable. In our high-volume center, which has a strong track record of surgical valve repair, it is reasonable to infer that patients selected for MVR likely presented with severely remodeled or rheumatic valves, resulting in significantly smaller valve areas in this group. Nevertheless, even after applying strict



	dice	precision	recall
a_2	0.8711	0.8394	0.9052
annulus	0.4979	0.4286	0.5941
chordae	0.7636	0.6562	0.913
p_2	0.6056	0.6281	0.5846

	dice	precision	recall
a_2	0.8664	0.8392	0.8954
p_2	0.8102	0.7603	0.8672

	dice	precision	recall
a_2	0.7912	0.75	0.8372
p_1	0.7466	0.6627	0.8549
p_3	0.8298	0.8203	0.8396

	dice	precision	recall
a_2	0.7768	0.8208	0.7373
p_1	0.864	0.8571	0.871
p_3	0.7092	0.6667	0.7576
papillary_anterior	0.484	0.464	0.5058
papillary_posterior	0.6868	0.6614	0.7143

FIGURE 12
Good performance segmentation example. ME 4C views at 0° **(A)** and ME commissural views at 63° **(B)** show good correlation between the ground truth and the neural network predicted annotation for A2-P2 and P3-A2-P1 combinations, respectively. Tables (below) show the corresponding performance metrics score for each label annotation. ME = mid-esophageal.

exclusion criteria, additional factors may have influenced the decision to replace rather than repair, thereby weakening the correlation between smaller MVA and surgical replacement. This limitation reduces the ability of the AI model to correctly classify MVR patients based solely on pre-procedural imaging. Moreover, the retrospective design and the single-center context limit the generalizability of our findings, particularly in centers with lower surgical expertise in mitral valve repair.

Finally, the scallop-level analysis represents a novel and ambitious contribution toward standardized, automated scallop identification, a task that today relies heavily on operator expertise. While segmentation of large structures (annulus, anterior and posterior leaflets, central scallops) reached acceptable accuracy, finer anatomical elements such as commissural leaflet scallops or chordae tendinae were more difficult to identify consistently. The mean Dice score of 0.53 on the overall structures dataset reflects these challenges. Nonetheless, the network correctly reproduced frequent anatomical configurations (e.g., A2-P2, P1-A2-P3), especially in mid-esophageal long-axis and commissural views, which are crucial for procedural planning. This constitutes a meaningful step. One important consideration in interpreting scallop-level results is the potential role of overfitting. Our dataset, while curated and enriched with physician annotations, remains limited in size compared to the complexity of the task. Neural networks trained on relatively small, homogeneous datasets are prone to overfitting, i.e., capturing dataset-specific patterns rather than generalizable features. This phenomenon may explain why performance was higher in anatomical regions and views more frequently represented in the training set (e.g., A2-P2 in long-axis views), while less common configurations showed reduced accuracy. Overfitting risk is further heightened by the high class imbalance inherent in scallop annotation: commissural scallops, papillary muscles, and chordae are both smaller in size and underrepresented, leading to disproportionate errors in Dice score evaluation. Another factor to consider is that Dice scores, while informative, may not fully reflect clinical usability. For small structures, a low Dice value may correspond to visually acceptable segmentation. Conversely, a higher Dice in a large structure might still fail to capture clinically relevant details such as leaflet clefts or tethering. This highlights the need for evaluation metrics that combine geometric accuracy with clinical relevance, possibly integrating expert qualitative scoring.

Compared to earlier approaches, which focused on annulus-only segmentation or static 3D models (Costa et al., 2019; Carnahan et al., 2021; Aly et al., 2022; Chen et al., 2023; Munafò et al., 2024; Andreassen et al., 2019; Andreassen et al., 2022), our pipeline integrates quality control, 4D functional analysis, and scallop-level anatomy into a unified framework. Recent semi-supervised methods (Munafò et al., 2025) demonstrated reliable 4D segmentation, but they did not extend to scallop analysis or clinical validation against surgical and percutaneous cohorts.

By validating our algorithm against both manual measurements and group-level clinical outcomes, we provide an important translational step toward clinical applicability.

4.1 Limitations and core challenges

Several limitations should be taken into account. Our datasets are relatively modest and partially monocentric, raising concerns about generalizability. Although validation against clinical reports is encouraging, clinical measurements and annotations themselves are subject to intra- and inter-operator variability, especially for complex structures like scallops and pathology zones, which could influence correlations. This is the concept of the noisy ground truth: an AI model trained on one expert's labels may perform poorly when judged by another expert's standards, highlighting the inherent ambiguity in the task. Scallop-level segmentation performance is limited and susceptible to overfitting and to the "rare event" challenge: pathologies like commissural lesions, complex Barlow's disease with multiple prolapses, or specific calcification patterns are less frequent than standard A2/P2 pathologies. A deep learning model trained on an imbalanced dataset will inherently be biased towards the more common cases and will struggle with these rare but clinically crucial edge cases. External validation on larger, more diverse datasets will be essential to confirm robustness.

Finally, while our system successfully quantifies pre-procedural imaging, its real-time intra-procedural utility remains untested. A model can perfectly segment a valve and measure lengths, but determining the feasibility and the clip strategy requires synthesizing all that information into a clinical decision. This involves tacit knowledge that cardiologists and cardiac surgeons accumulate over years and that is rarely explicitly stated in the annotations (e.g., "leaflet is too fragile," "coaptation gap is too wide for a single clip," "the jet is too commissural for a safe grasp").

4.2 Future perspectives

Moving forward, expanding annotated datasets, ideally through multi-center collaborations and semi-automated labeling strategies, will be crucial to mitigate overfitting and improve generalizability. Incorporating advanced architectures (e.g., vision transformers or hybrid CNN–transformer models) and uncertainty quantification methods may further enhance reliability in challenging cases. Using the STAPLE algorithm (Simultaneous Truth and Performance Level Estimation) or similar statistical methods to generate a probabilistic "consensus truth" from multiple annotations could also help in building a more consolidated dataset for training. The use of generative AI techniques like Generative Adversarial Networks (GANs) or diffusion models could help to create realistic synthetic examples of rare and challenging cases to balance the training set. Furthermore, the segmentation of the valve informs the pathology classification, which tells the feasibility prediction. Design a single model that simultaneously learns to segment, classify views, classify pathology, and detect calcifications makes the model more robust and generalizable than a set of separate models.

Dabiri et al. (2022) conducted a simulation study to assess how the number and location of MitraClips influence residual MR and valve hemodynamics. This study emphasizes that procedural success depends not only on patient selection but also on real-time strategic decisions regarding clip quantity and positioning.

In fact, beyond pre-procedural planning, a promising future direction involves integrating DL into intra-procedural guidance. Real-time segmentation and scallop identification could assist operators during clip placement by continuously updating valve anatomy and coaptation maps as the device interacts with the leaflets. Automated tracking of leaflet grasping zones and prediction of residual regurgitation jets could help reduce procedure time, cut down on unnecessary clip deployments, and improve procedural safety. Such integration would need further optimization of inference speed, user-friendly visualization tools, and compatibility with procedural echocardiography systems. Ultimately, combining imaging-derived AI quantification with biomechanical simulation could create a comprehensive decision-support system, predicting both procedural feasibility and the hemodynamic trade-offs of various clip strategies.

The most transformative future direction, however, involves a core shift from a reconstructive to a predictive and simulative model. Current models, including our own, analyze the pre-procedural anatomy in a static way. One future use of our ECHO-PREP workflow will be to train the model on paired data: pre-procedural 3D TEE volumes and their corresponding post-procedural 3D TEE volumes with the clip deployed and a good result. Instead of just identifying what exists, the AI will learn what a successful outcome looks like and apply that knowledge to guide the pre-procedural plan, by understanding the mechanical changes caused by clip implantation on the valve and the optimal morphological features of a pre-procedural valve that lead to a successful post-procedural result. This "backward-forward" AI approach has the highest potential to truly standardize and democratize M-TEER planning worldwide, allowing less experienced centers to leverage the collective expertise embedded in the AI from high-volume centers, all while using the standard imaging equipment they already have. These analyses, together with the integration of fluid-dynamics simulations, are envisioned as central elements of a comprehensive multi-imaging simulation platform for transcatheter procedures, which we are currently advancing through our ongoing multicenter study, ENVISAGE (NCT07213531).

If validated prospectively, this capability could transform ECHO-PREP from a pre-procedural planning tool into a real-time decision-support system integrated into the cath lab workflow. We envision a future where the interventional cardiologist is empowered not just with tools, but with foresight. This is the true promise of AI: not to replace physicians, but to enhance their capabilities, making their expertise more powerful, precise, and accessible to every patient in need.

Data availability statement

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

Author contributions

SC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing. TG: Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing. OT: Data curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing. DC: Data curation, Formal Analysis, Methodology, Software, Writing – original draft, Writing – review and editing. TM: Supervision, Writing – review and editing. SvB: Supervision, Validation, Writing – review and editing. FL: Conceptualization, Methodology, Software, Supervision, Validation, Writing – review and editing. WBA: Conceptualization, Investigation, Project administration, Validation, Writing – original draft, Writing – review and editing.

Funding

The authors declare that no financial support was received for the research and/or publication of this article.

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.

Generative AI statement

The authors declare that Generative AI was used in the creation of this manuscript. For correcting the grammar and aiding in generating descriptions for figure legends.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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/fnetp.2025.1701758/full#supplementary-material

References

Aly, A., Khandelwal, P., Aly, A., Kawashima, T., Mori, K., Saito, Y., et al. (2022). Fully automated 3d segmentation and diffeomorphic medial modeling of the left ventricle mitral valve complex in ischemic mitral regurgitation. *Med. Image Anal.* 80, 102513. doi:10.1016/j.media.2022.102513

Andreassen, B., Veronesi, F., Gerard, O., Solberg, A., and Samset, E. (2019). Mitral annulus segmentation using deep learning in 3-D transesophageal echocardiography. *IEEE J. Biomed. Health Infor* 24, 994–1003. doi:10.1109/JBHI.2019.2959430

Andreassen, B., Völgyes, D., Samset, E., and Solberg, A. (2022). Mitral annulus segmentation and anatomical orientation detection in TEE images using periodic 3D CNN. *IEEE Access* 10, 51472–51486. doi:10.1109/access.2022.3174059

Carnahan, P., Moore, J., Bainbridge, D., Eskandari, M., Chen, E., and Peters, T. (2021). "DeepMitral: fully automatic 3D echocardiography segmentation for patient specific mitral valve modelling," in *Medical image computing and computer assisted Intervention-MICCAI 2021: 24th international conference, strasbourg, France, September 27-October 1, 2021, proceedings, part V 24*, 459–468.

Chen, J., Li, H., He, G., Yao, F., Lai, L., Yao, J., et al. (2023). Automatic 3D mitral valve leaflet segmentation and validation of quantitative measurement. *Biomed. Sig Process Control* 79, 104166. doi:10.1016/j.bspc.2022.104166

Costa, E., Martins, N., Sultan, M., Veiga, D., Ferreira, M., Mattos, S., et al. (2019). Mitral valve leaflets segmentation in echocardiography using convolutional neural networks. *2019 IEEE 6th Portuguese Meet. Bioeng. (ENBENG)*, 1–4. doi:10.1109/enbeng.2019.8692573

Dabiri, Y., Mahadevan, V. S., Guccione, J. M., and Kassab, G. S. (2022). A simulation study of the effects of number and location of MitraClips on mitral regurgitation. *JACC Adv.* 1 (1), 100015. doi:10.1016/j.jacadv.2022.100015

Elaziz, E. A., Al-qaness, M., Dahou, M., Alsamhi, S. H., Abualigah, L., Ibrahim, R. A., et al. (2023). Evolution toward intelligent communications: impact of deep learning applications on the future of 6G technology. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 14, e1521. doi:10.1002/widm.1521

Hien, M., Großgasteiger, M., Weymann, A., Rauch, H., and Rosendal, C. (2014). Reproducibility in echocardiographic two-and three-dimensional mitral valve assessment. *Echocardiography* 31, 311–317. doi:10.1111/echo.12365

Kaiming, H., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv:1512.03385*. Available online at: https://arxiv.org/abs/1512.03385.

Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., et al. (2019). Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans. Med. Imaging* 38 (9), 2198–2210. doi:10.1109/TMI.2019.2900516

Lee, C., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2014). Deeply-supervised nets. *arXiv:1409.5185.* doi:10.48550/arXiv.1409.5185

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection. *arXiv: 1708.02002*. Available online at: https://arxiv.org/abs/1708.02002.

Maisano, F., Franzen, O., Baldus, S., Schäfer, U., Hausleiter, J., Butter, C., et al. (2013). Percutaneous mitral valve interventions in the real world: early and 1-year results from the ACCESS-EU, a prospective, multicenter, nonrandomized post-approval study of the MitraClip therapy in Europe. *J. Am. Coll. Cardiol.* 62 (12), 1052–1061. doi:10.1016/j.jacc.2013.02.094

Munafò, R., Saitta, S., Ingallina, G., Denti, P., Maisano, F., Agricola, E., et al. (2024). A Deep Learning-Based Fully Automated Pipeline for Regurgitant Mitral Valve Anatomy Analysis From 3D Echocardiography. *IEEE Access*, 12, 5295–5308. doi:10.1109/ACCESS.2024.3349698

Munafò, R., Saitta, S., Tondi, D., Ingallina, G., Denti, P., Maisano, F., et al. (2025). Automatic 4D mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach. *Med. Biol. Eng. Comput.* doi:10.1007/s11517-024-03275-w

Nkomo, V. T., Gardin, J. M., Skelton, T. N., Gottdiener, J. S., Scott, C. G., and Enriquez-Sarano, M. (2006). Burden of valvular heart diseases: a population-based study. *Lancet* 368 (9540), 1005–1011. doi:10.1016/S0140-6736(06)69208-8

Odena, A., Dumoulin, V., and Olah, C. (2016). "Deconvolution and checkerboard artifacts," in *Distill*. doi:10.23915/distill.00003

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., et al. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 580, 252–256. doi:10.1038/s41586-020-2145-8

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *arXiv 1505.04597*, 234–241. doi:10.1007/978-3-319-24574-4_28

Schlemper, J., Oktay, O., Shaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: learning to leverage salient regions in medical images. *arXiv: 1808.08114* 53, 197–207. doi:10.1016/j.media.2019.01.012

Stone, G. W., Lindenfeld, J., Abraham, W. T., Kar, S., Lim, D. S., Mishell, J. M., et al. (2018). Transcatheter mitral-valve repair in patients with heart failure. *N. Engl. J. Med.* 379 (24), 2307–2318. doi:10.1056/NEJMoa1806640

Synapse (2025). Synapse.org. Available online at: https://www.synapse.org/Synapse:syn51186045/wiki/621356.

Taskén, A., Berg, E., Grenne, B., Holte, E., Dalen, H., Stølen, S., et al. (2023). Automated estimation of mitral annular plane systolic excursion by artificial intelligence from 3D ultrasound recordings. *Artif. Intell. Med.* 144, 102646. doi:10.1016/j.artmed.2023.102646

Thomas, N., Unsworth, B., Ferenczi, E., Davies, J. E., Mayet, J., and Francis, D. P. (2008). Intraobserver variability in grading severity of repeated identical cases of mitral regurgitation. *Am. Heart J.* 156, 1089–1094. doi:10.1016/j.ahj.2008.07.017

Votta, E., Caiani, E., Veronesi, F., Soncini, M., Montevecchi, F. M., and Redaelli, A. (2008). Mitral valve finite-element modelling from ultrasound data: a pilot study for a new approach to understand mitral function and clinical scenarios. *Philo Trans. R. Soc. A Math. Phys. Eng. Sci.* 366, 3411–3434. doi:10.1098/rsta.2008.0095

Wifstad, S., Kildahl, H., Grenne, B., Holte, E., Hauge, S. W., Sæbø, S., et al. (2024). Mitral valve segmentation and tracking from transthoracic echocardiography using deep learning. *Ultrasound Med. Biol.* 50, 661–670. doi:10.1016/j.ultrasmedbio.2023.12.023

Zamorano, J., Badano, L., Bruce, C., Chan, K. L., Gonçalves, A., Hahn, R. T., et al. (2011). EAE/ASE recommendations for the use of echocardiography in new transcatheter interventions for valvular heart disease. *Eur. Heart J.* 32, 2189–2214. doi:10.1093/eurheartj/ehr259

TYPE Original Research
PUBLISHED 28 November 2025
DOI 10.3389/fcvm.2025.1614979

OPEN ACCESS

EDITED BY
Ulrich Parlitz,
Max Planck Institute for Dynamics and Self-Organization, Germany

REVIEWED BY
Naufal Zagidullin,
Bashkir State Medical University, Russia
Siming Tao,
The Affiliated Hospital of Yunnan University, China

*CORRESPONDENCE
Ju Mei
✉ ju_mei63@126.com
Zhaolei Jiang
✉ wojiangzhaolei@163.com

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

‡These authors have contributed equally to this work

RECEIVED 20 April 2025
REVISED 16 November 2025
ACCEPTED 17 November 2025
PUBLISHED 28 November 2025

CITATION
Wang C, Muradil M, Huang J, Cai J, Ding F, Zhang L, Li M, Fu C, Mei J and Jiang Z (2025) Identification of key genes associated with atrial fibrillation and hypoxia using WGCNA and machine learning technology.
Front. Cardiovasc. Med. 12:1614979.
doi: 10.3389/fcvm.2025.1614979

Identification of key genes associated with atrial fibrillation and hypoxia using WGCNA and machine learning technology

Chao Wang[1†], Mardan Muradil[2†], Jianbin Huang[1], Jie Cai[1], Fangbao Ding[1], Li Zhang[1], Mengda Li[3], Chenglai Fu[1,4], Ju Mei[1*‡] and Zhaolei Jiang[1*‡]

[1]Department of Cardiothoracic Surgery, Xinhua Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China, [2]Spine Center, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China, [3]Neurological Surgery, Henan Provincial People's Hospital, Henan, Zhengzhou, China, [4]Institute for Developmental and Regenerative Cardiovascular Medicine, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Background: Atrial fibrillation (AF) is among the most prevalent cardiac arrhythmias worldwide, and its incidence is steadily rising due to global aging. Hypoxia, a well-recognized trigger of AF, plays a pivotal role in the onset and progression of AF. However, the molecular mechanisms underlying the interplay between AF and hypoxia remain unclear, and specific biomarkers for this condition are lacking. This study aimed to identify key hypoxia-related genes associated with AF through an integrated bioinformatics approach that combines weighted gene co-expression network analysis (WGCNA) with machine learning (ML) algorithms, and to assess their potential diagnostic significance.

Methods: This study employed an integrative approach combining weighted gene co-expression network analysis (WGCNA) and machine learning (ML) to identify key genes associated with AF under hypoxic conditions. AF-related gene expression data were sourced from the Gene Expression Omnibus (GEO) database, and hypoxia-related gene sets from the Molecular Signatures Database (MSigDB) database. WGCNA was employed to identify gene modules associated with AF, which were then intersected with hypoxia-related genes. Candidate hub genes were identified using random forest and least absolute shrinkage and selection operator regression. Their diagnostic performance was evaluated using receiver operating characteristic (ROC) curve analysis. A predictive nomogram was developed, and immune infiltration analysis and gene set enrichment analysis (GSEA) were performed to explore associated biological pathways and alterations in the immune landscape.

Results: WGCNA identified 34 gene modules, with the most AF-relevant module comprising 624 genes. Intersection analysis and ML algorithms identified SLC6A6, BGN, and PFKP as key genes. ROC analysis demonstrated strong diagnostic potential. Immune cell profiling showed increased infiltration of M2 macrophages and dendritic cells in AF samples, with significant correlations to the expression of these hub genes.

Conclusion: This study identified SLC6A6, BGN, and PFKP as key genes associated with AF under hypoxic conditions and successfully developed a diagnostic model with promising clinical applicability. These genes likely play important roles in hypoxia-mediated AF pathogenesis and are closely associated with immune cell infiltration, providing potential biomarkers for early diagnosis and precision treatment of AF. This study provides novel insights into the molecular mechanisms underlying the interplay between hypoxia and AF.

KEYWORDS

atrial fibrillation, hypoxia, weighted gene co-expression network analysis (WGCNA), machine learning, hub genes

1 Introduction

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia worldwide, with an increasing incidence due to the progressive aging of the population. An epidemiological study in the United States estimated that the number of individuals with AF ranges from 3 to 6 million, and this figure is projected to rise to approximately 6 to 16 million by 2050 (1). Hypoxia is one of the common triggering factors of AF and serves as a critical driver of its sustained progression. The most frequently encountered conditions leading to hypoxia include coronary artery disease (2, 3) and obstructive sleep apnea syndrome (4). Multiple mechanisms underlie the induction of AF under hypoxic conditions, primarily involving atrial electrophysiological and structural remodeling, inflammatory responses, and oxidative stress (2, 5). Although extensive clinical and fundamental research has demonstrated a close association between AF and hypoxia, a thorough investigation into the molecular mechanisms concerning AF onset in a hypoxic state remains insufficient, particularly regarding the identification of definitive biomarkers. In recent years, weighted gene co-expression network analysis (WGCNA) has become an effective bioinformatics approach for finding key gene modules related to specific diseases based on gene expression data (6). Additionally, machine learning (ML) techniques like random forest (RF) and least absolute shrinkage and selection operator (LASSO) regression, have been widely applied in gene selection and disease prediction (7, 8). By integrating WGCNA with ML approaches, it is possible to identify critical genes related to AF and hypoxia with greater precision. Therefore, this study aimed to elucidate the molecular mechanisms linking hypoxia signaling with AF, identify key hub genes, and construct a diagnostic model based on these genes. Our findings will offer new perspectives and lay the theoretical groundwork for unveiling the pathogenesis of AF in hypoxic states and developing innovative diagnostic tools.

2 Methods

2.1 Data collection and preprocessing

The gene expression profiling data for AF were sourced from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), with details presented in Table 1. The GSE115574 dataset comprises 59 samples, including 31 left atrial appendage (LAA) tissues from AF patients and 28 control tissues from people with sinus rhythm (SR). The GSE14975 dataset encompasses 10 samples, including 5 AF LAA tissues and 5 SR control samples. The GSE41177 dataset includes 38 samples, comprising 32 AF LAA tissues and 6 SR control samples. The gene-related information for all AF datasets was derived from the GPL570 platform. The GSE115574 and GSE14975 datasets were combined to create a training set, while GSE41177 was designated as an external validation set. After the raw dataset files were downloaded, genes with zero or negative expression values were excluded before $\log_2$ transformation. Low-expression probes (mean $\log_2$ intensity < 1 or detected in $<50\%$ of samples) and low-variance genes [median absolute deviation (MAD) < 0.15] were removed. A total of 21,653 genes meeting these criteria were retained for subsequent analyses. Expression values for each gene were then normalized to ensure their independence, facilitating subsequent computational processing. Raw expression data were retrieved using R 4.4.1 and the GEOquery package. Subsequently, an expression matrix was constructed and probes were mapped to their corresponding gene symbols. Duplicate genes and missing values were eliminated. If a probe corresponded to multiple genes, that particular gene was excluded to ensure data integrity. The final gene matrix was integrated, and the batch effect was corrected utilizing the ComBat function from the SVA package 3.52.0. Information on hypoxia-upregulated genes was sourced from the Molecular Signatures Database (MSigDB) (http://www.gsea-msigdb.org). Specifically, 200 hypoxia-related genes were obtained by querying the database using the "HALLMARK_HYPOXIA" keyword (Figure 1).

2.2 WGCNA

To ensure the accuracy of WGCNA, the integrated gene matrix was further filtered, and genes expressed in over 50% of the samples were retained. The WGCNA package 1.73 was employed to assess the quality of samples and genes, ensuring that the data matrix was suitable for WGCNA. Sample hierarchical clustering was performed on the integrated gene-sample data to identify potential outliers (Figure 3A).

TABLE 1 Brief description of hypoxia and AF source dataset.

Disease/Pathological state	Data chip	Sample size		Data source	Year
		Normal control	Disease		
AF	GSE115574	28	31	Gene expression data from human left and right atrial tissues in patients with degenerative MR in SR and AF	2021
AF	GSE14975	5	5	Transcriptional profiling of left atrial myocardium from AF and SR patients	2019
AF	GSE41177	6	32	Region-specific gene expression profiles in left atria of patients with valvular atrial fibrillation	2019
Hypoxia	M5891	200		Genes up-regulated in response to low oxygen levels (hypoxia).	2015

AF, atrial fibrillation; SR, sinus rhythm.

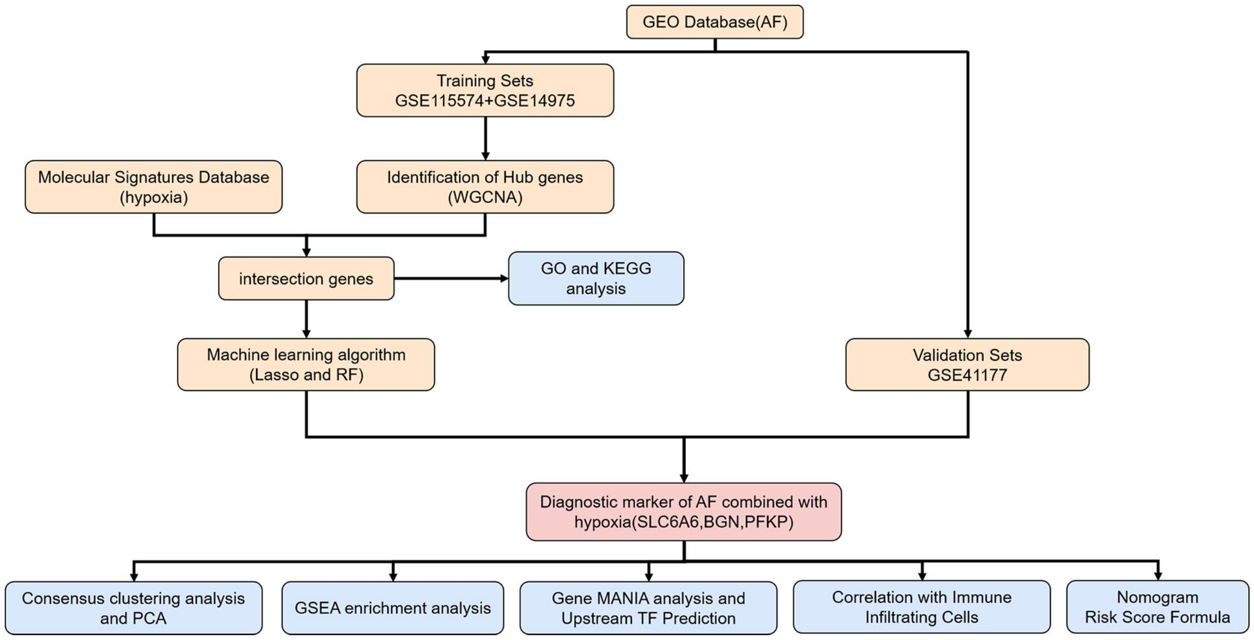


FIGURE 1
Flowchart of the study. AF, atrial fibrillation; WGCNA, weighted gene Co-expression network analysis; GO, gene ontology; KEGG, Kyoto encyclopedia of genes and genomes; LASSO, least absolute shrinkage and selection operator; RF, random forest; PCA, principal component analysis; GSEA, gene set enrichment analysis gene; MANIA, gene multiple association network integration algorithm.

A biologically significant scale-free network was developed utilizing the soft-thresholding parameter (β) as per the scale-free topology requirement. The "pickSoftThreshold" function was adopted for computing and selecting an appropriate β value, ensuring that the scale-free topology model fitting index (sftr squared) was approximately 0.8, thereby guaranteeing network robustness and biological relevance. Based on network topology analysis, a CutHeight value (height threshold) greater than 0.8 was selected for gene module construction, with at least 50 module genes. The gene module most strongly linked to AF was identified. Gene modules were defined via a topological overlap matrix (TOM) in combination with the dynamic tree-cut method. After module delineation, the eigengene was calculated for every module (module eigengene, ME). Finally, Pearson correlation coefficients were employed to evaluate the link of modules to clinical traits. The module most strongly related to clinical characteristics was identified as the key module and visualized within the trait-gene network.

2.3 Functional enrichment analysis of hypoxia- and AF-associated Hub genes

Our study identified the intersection between hypoxia-related genes and the most AF-relevant module genes from WGCNA. Their involvement in biological processes (BP), molecular functions (MF), and cellular components (CC) was explored through Gene Ontology (GO) analysis on the intersecting genes. Additionally, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted to characterize and describe gene functions. To ascertain the statistical enrichment of genes within KEGG and GO pathways, the ClusterProfiler package 4.12.6 was utilized. Pathways containing three or more significantly enriched genes with $p < 0.05$ suggesting significance.

2.4 ML-based key genes identification

Two commonly used ML algorithms: RF and LASSO, were utilized following the identification of hypoxia-AF intersecting genes. LASSO regression analysis was enabled by the "glmnet" package 4.1–8, while the RF model was generated via the "randomForest" package 4.7–1.2. Genes overlapping between the RF and LASSO analyses were considered potential hypoxia-associated AF biomarkers. Furthermore, to evaluate the diagnostic value, receiver operating characteristic (ROC) curves comprehensively present the model's classification performance at different thresholds.

2.5 Construction and evaluation of the predictive model

A nomogram model was built on the identified key gene set. Every risk factor was assigned a corresponding score, with the total score mapped to AF occurrence probability. Calibration curves reflected the concordance between observed and predicted results. The net benefit of the model in forecasting AF occurrence was examined through decision curve analysis (DCA).

2.6 Relationship between key genes and infiltrating immune cells

To investigate immune cell infiltration within the samples, this study utilized the CIBERSORT package 0.1.0, which employs the LM22 immune cell-specific gene matrix (LM22 signature) for correlation analysis. CIBERSORT was used to estimate the proportion of immune cells in each sample, and the results were integrated with sample classification data. Bar plots presented overall immune cell infiltration, and box plots illustrated differences across AF and normal samples. To explore the potential relationships of key genes with immune cells, the links of selected genes to each immune cell type were unraveled via Spearman correlation analysis. The obtained correlation coefficients and significance values were visualized in a heatmap. Data visualization was performed using the R packages "reshape" 0.8.9, "tidyverse" 2.0.0, "ggplot2" 3.5.1, and "pheatmap" 1.0.12. $p < 0.05$ denoted statistical significance.

2.7 Gene set enrichment analysis (GSEA) and prediction of upstream transcription factors

GSEA was employed to interpret the biological significance of specific genes in biological processes or diseases via the GSVA package 2.0.5 in R, with visualization implemented via the "ggplot2" 3.5.1 and "enrichplot" 1.24.4 packages. Gene sets meeting the adjusted $p < 0.05$ threshold were significant. A co-expression gene network was formed via GeneMANIA (http://www.genemania.org). NetworkAnalyst (https://www.networkanalyst.ca) helped to analyze the link of key genes to their associated transcription factors, facilitating the prediction of upstream transcriptional regulators.

2.8 Consensus clustering analysis and principal component analysis (PCA)

Clustering analysis algorithms were leveraged for validating and confirming the biological significance and effectiveness of identified key genes. Based on hypoxia- and AF-associated key genes, unsupervised consensus clustering was performed through the "ConsensusClusterPlus" package 1.70.0. AF patient subtypes were defined across all AF samples, with 50 iterations conducted to assess result stability. Key operational parameters were an 80% item resampling rate and a maximum k-value of 10. PCA was utilized to examine the differentiation among clustering groupings and to corroborate the clustering outcomes.

3 Results

3.1 Construction and processing of the AF dataset

The batch effect was corrected on the GSE115574 and GSE14975 datasets (Figures 2A,B), comprising 69 samples, including 33 AF samples and 36 SR samples. 21,653 genes were identified, and PCA was undertaken to detect differences before and after correction (Figures 2C,D).

3.2 Identification of AF-associated genes using WGCNA

The optimal soft-threshold (β) was determined as per the scale-free topology criterion. When the soft-threshold was 5, R^2 exceeded 0.8 (Figure 3B). Gene clustering resulted in 34 modules, with the module dendrogram illustrated in Figures 3C–E. A heatmap illustrates the correlation of gene modules with AF-related clinical characteristics (Figure 4). Notably, the green module showed the highest relation to AF ($r = 0.4$, $P = 7\text{e-}04$), encompassing 624 genes. Intersecting the genes from the key AF-related module with the hypoxia gene set yielded 16 overlapping genes (Figure 5). An overlap of 16 genes was observed between the AF-associated "green module" ($n = 624$) and the HALLMARK_HYPOXIA gene set ($K = 200$), which was significantly greater than expected by chance (expected 5.76; hypergeometric $p = 0.00056$; OR = 2.56, 95% CI: 2–11), indicating a non-random enrichment of hypoxia-responsive genes within the AF-associated module.

3.3 Enrichment analysis

To clarify the shared molecular biological processes of disease-linked genes, GO and KEGG enrichment analyses were carried out

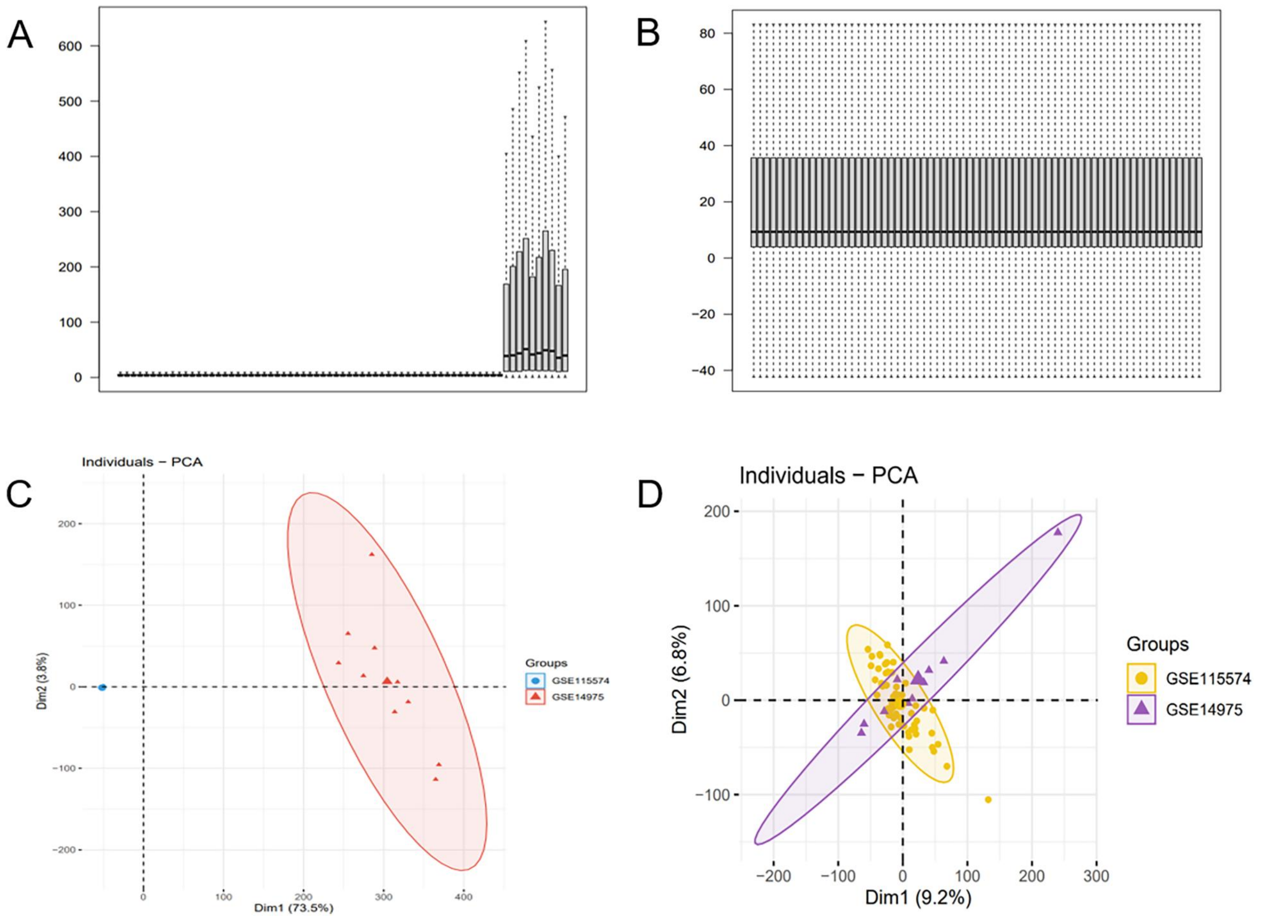


FIGURE 2
Batch effect correction results. (A) Boxplot of the AF dataset before batch effect correction. (B) Boxplot of the AF dataset after batch effect correction. (C) PCA of the AF dataset before batch effect correction. (D) PCA of the AF dataset after batch effect correction.

on the overlapping genes (Figure 6). The GO enrichment analysis highlighted the top 10 pathways linked to biological processes (BP) and molecular functions (MF), while the KEGG enrichment analysis identified the top 9 pathways. In the BP category, GO terms were predominantly linked to carbohydrate metabolism, ADP catabolic processes, and nucleotide metabolism. In the MF category, functional enrichment was observed in pathways related to monosaccharide binding and carbohydrate kinase activity. 9 pathways were identified after KEGG pathway enrichment analysis, revealing the clustering of hub genes in many pathways, including the HIF-1 signaling pathway, glycolysis/gluconeogenesis, glycosaminoglycan biosynthesis, as well as carbon metabolism.

3.4 Key genes selection via ML and evaluation of model diagnostic performance

To further refine key gene selection, a 10-fold cross-validation analysis on the 16 overlapping genes was carried out via the LASSO algorithm (Figure 7B), and the optimal Lambda value (Lambda.min) was 0.04 (Figure 7A). Eight key genes were identified: HK1, PFKP, BGN, CDKN1A, CAV1, SLC6A6, P4HA1, and CHST. In the RF algorithm, the optimal tree number was 56, corresponding to the lowest error rate of 0.18 (Figure 7C). Genes with Mean Decrease Gini (MDG) scores > 2 were retained, corresponding to approximately the top 3% of features above the inflection point of the MDG distribution, thereby representing variables of high relative importance in the Random Forest model. Genes scoring over 2 were selected (Figure 7D), including SLC6A6, MYH9, CHST2, B3GALT6, GAPDH, PFKP, and BGN. Notably, SLC6A6, BGN, and PFKP were shared between LASSO and RF algorithms (Figure 8). To evaluate the predicting accuracy of critical genes for AF under hypoxic conditions, ROC curves were constructed, and the effectiveness was analyzed using metrics such as area under the curve (AUC), sensitivity, and specificity. The performance of the AF forecasting models built on RF and LASSO algorithms was examined (Figure 9). When candidate hub genes were used, the AUC values were as follows: SLC6A6, 0.736; BGN, 0.705; PFKP, 0.726. External validation using the GSE41177 dataset further corroborated the model's performance, yielding AUC values of SLC6A6, 0.891; BGN, 0.943; and PFKP, 0.953. Similarly, in the GSE79768 dataset, AUC values were 0.887 for SLC6A6, 0.613 for BGN, and 0.548 for PFKP.

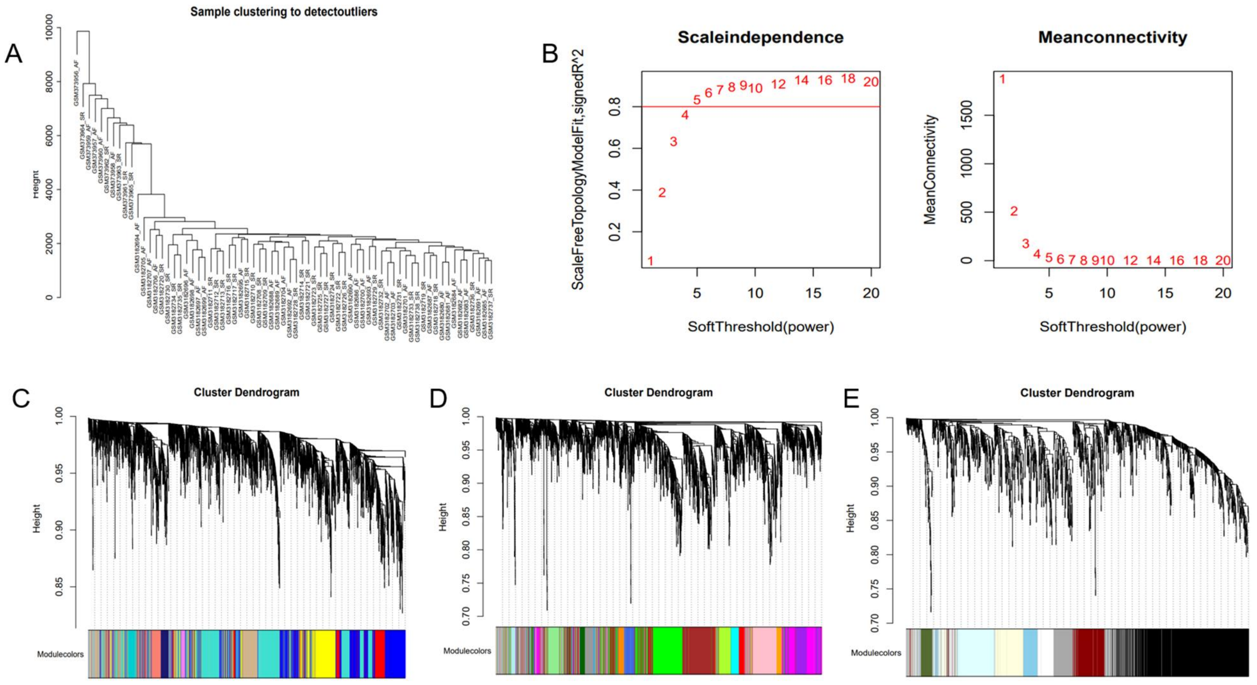


FIGURE 3
WGCNA. **(A)** Sample clustering dendrogram of AF. **(B)** Relationship of the fitting index with soft threshold (left) and the relationship of mean connectivity with soft threshold (right). **(C–E)** Module clustering dendrogram of the AF co-expression network with different colors representing different modules.

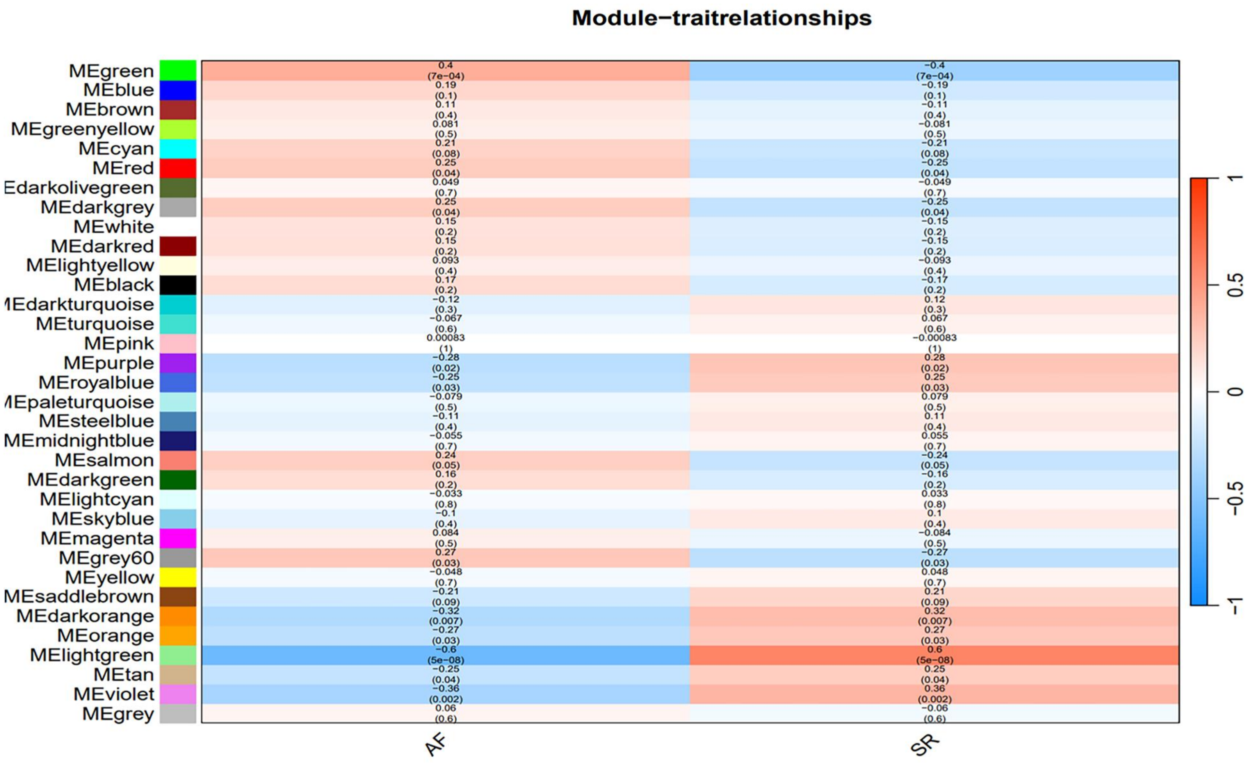


FIGURE 4
Heatmap of the correlation between gene modules and clinical characteristics of AF. Red indicates a positive correlation, blue indicates a negative correlation.

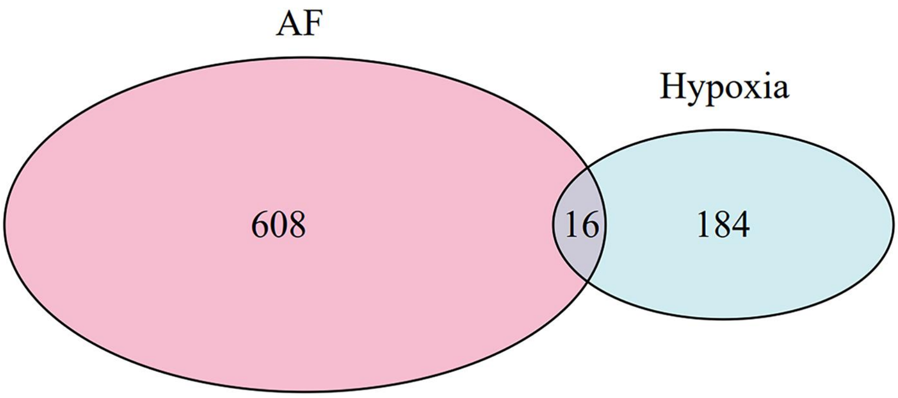


FIGURE 5
Venn diagram of AF-related genes identified by WGCNA and hypoxia-related genes.

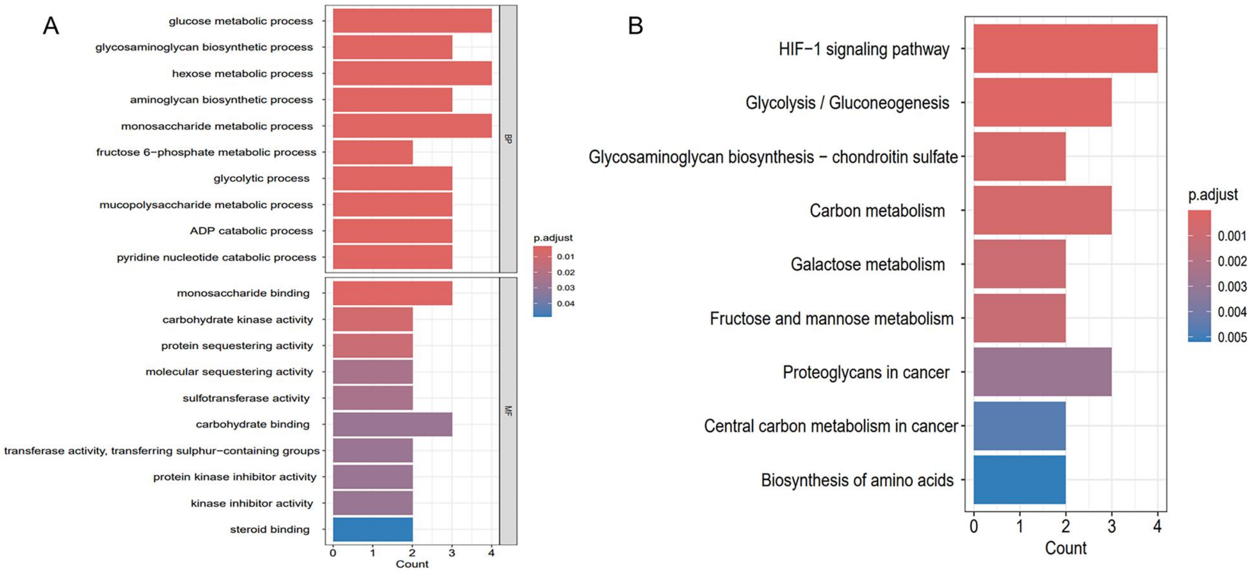


FIGURE 6
Enrichment analysis results. **(A)** GO enrichment analysis of key genes shared between hypoxia and AF. **(B)** KEGG enrichment analysis of key genes shared between hypoxia and AF.

These findings collectively demonstrate robust predictive performance across datasets.

3.5 Differential expression of key genes and prognostic model development and evaluation

The differential expression of key genes linked to AF and hypoxia was analyzed. Box plots (Figure 10A) showed that key genes exhibited markedly elevated expression levels in AF samples relative to SR samples ($P < 0.05$). Subsequently, a prognostic nomogram was constructed based on the expression levels of SLC6A6, BGN, and PFKP (Figure 10B) for risk assessment. The calculated score for each gene predicted the probability of AF occurrence. Calibration curves demonstrated minimal deviation between the observed and bias-corrected curves relative to the ideal curve (Figure 10C), indicating favorable predictive accuracy. Additionally, DCA demonstrated a significant net benefit (Figure 10D), highlighting the substantial clinical utility of the model in predicting AF during follow-up.

3.6 Immune cell infiltration analysis

Significant differences existed in immune cell infiltration patterns across AF and SR samples (Figure 11A). Further comparative analysis of immune cell proportions (Figure 11C)

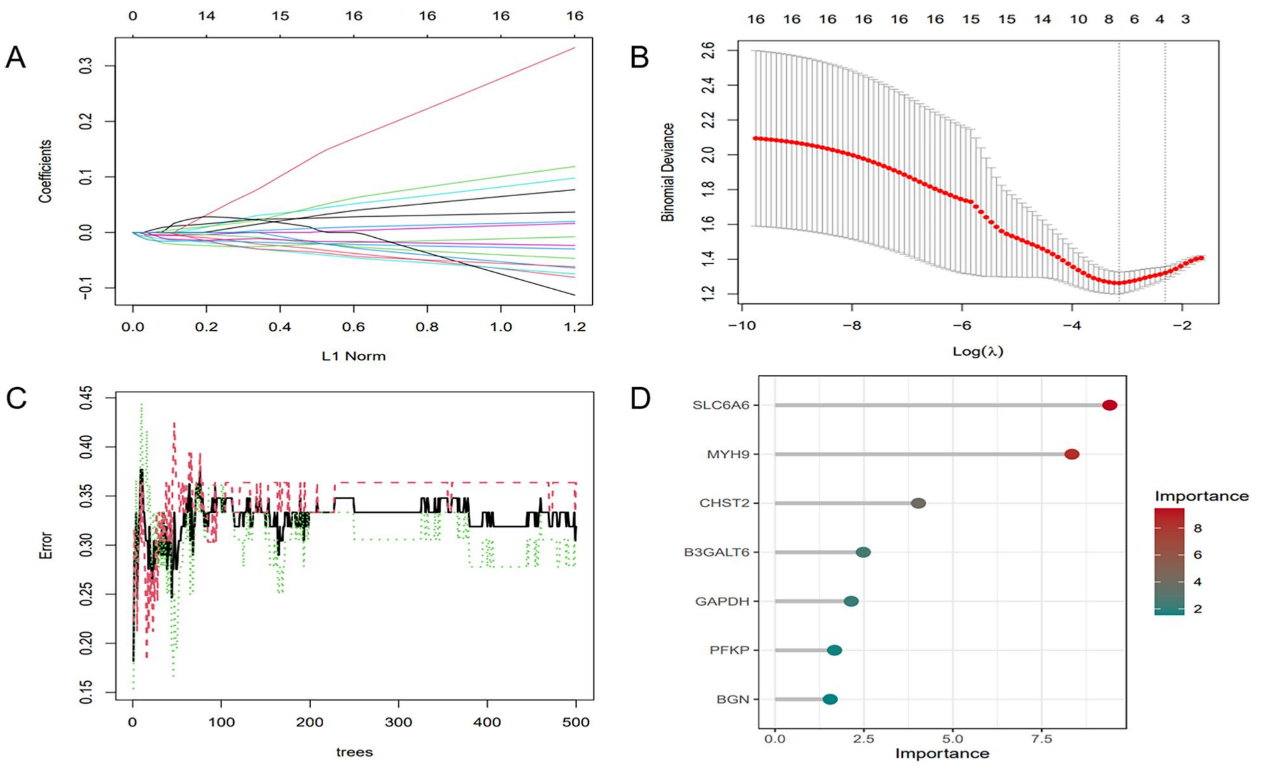


FIGURE 7
ML. (A) Regularization path of LASSO regression. (B) Cross-validation curve of LASSO regression. (C) RF model: the trend of error variation with the number of decision trees. (D) Feature selection results of the RF model.

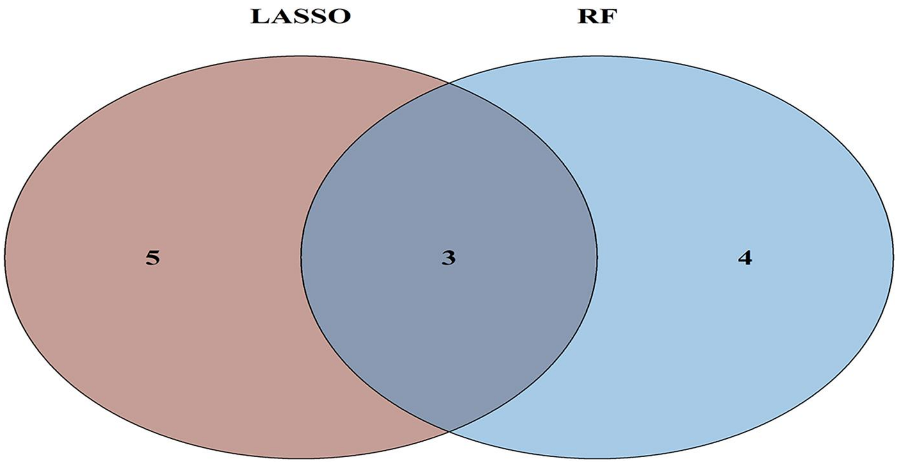


FIGURE 8
Venn diagram of ML results from RF and LASSO algorithms.

demonstrated that the proportions of M2 macrophages ($P < 0.05$) and resting dendritic cells(DCs) ($P < 0.01$) were notably higher in AF samples, whereas the proportion of regulatory T cells (Tregs) ($P < 0.05$) was evidently higher in SR samples. Correlation analysis among different immune cell populations (Figure 11B) revealed strong negative or positive correlations, such as an inverse relation of M2 to M0 macrophages, and positive correlations of M2 with M1 macrophages as well as resting mast cells. Moreover, immune cell infiltration varied significantly among different key genes (Figure 11D). For instance, SLC6A6 expression was positively correlated with resting DCs and CD4+ T cells, PFKP was negatively correlated with activated DCs, and BGN was positively correlated with M2 macrophages but negatively correlated with activated DCs.

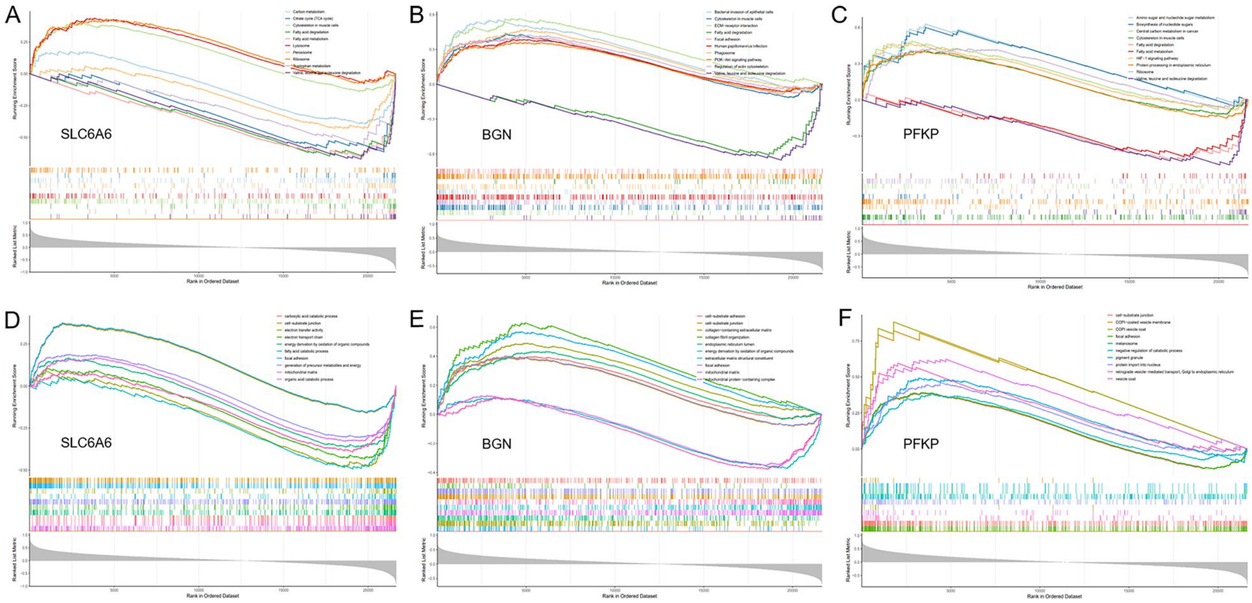


FIGURE 9
GSEA. **(A–C)** GO enrichment analysis of the key gene set (SLC6A6, PFKP, BGN) using GSEA. **(D–F)** KEGG enrichment analysis of the key gene set (SLC6A6, PFKP, BGN) using GSEA.

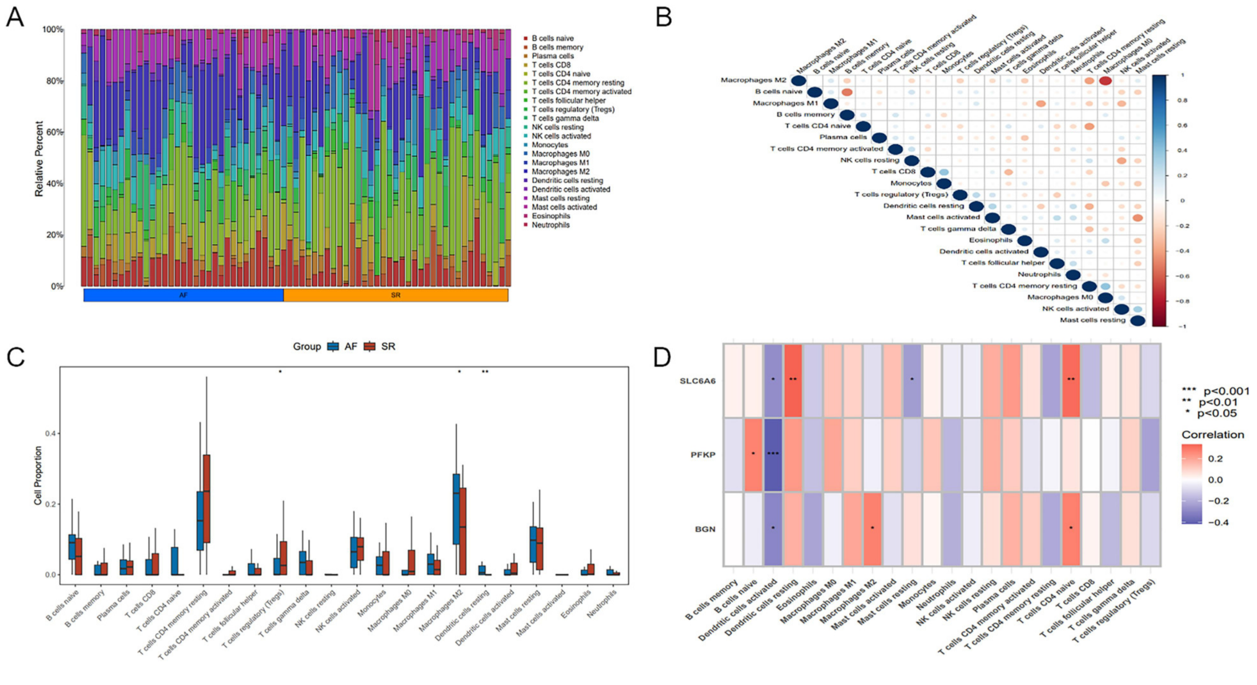


FIGURE 10
Immune infiltration analysis of key genes. **(A)** Heatmap illustrating differences in immune cell proportions between AF and SR samples. **(B)** Relationship among different immune cell types. **(C)** Boxplot of differences in immune cell proportions between AF and SR samples, blue represents AF patients, and red represents SR patients. **(D)** Correlation between the key gene set and immune cell populations. Statistical significance: **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

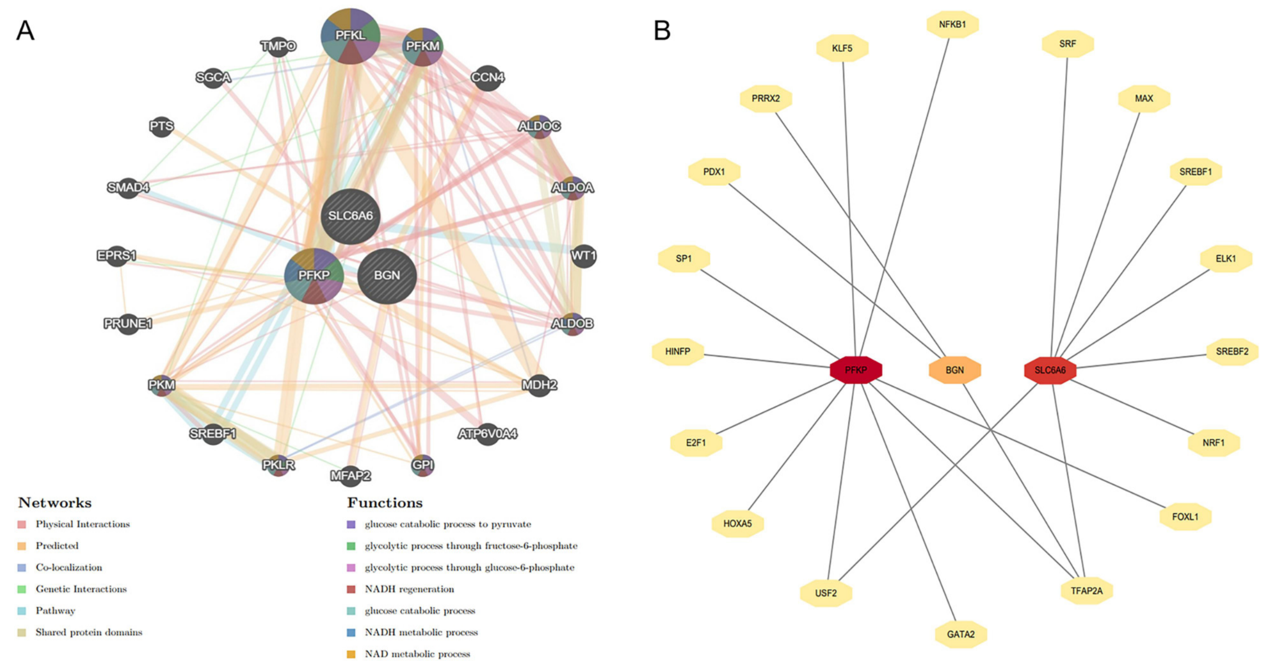


FIGURE 11
Construction of gene interaction network and prediction of upstream transcription factors. (A) The gene network analysis of the key gene set based on the GeneMANIA database. (B) Prediction of upstream transcription factors for the key gene set.

3.7 GSEA enrichment analysis and consensus clustering

To unveil the biological roles of key gene sets in AF, GSEA on three genes was carried out (Figure 12). The results revealed that the key genes (SLC6A6, BGN, and PFKP) are involved in distinct biological pathways, primarily including fatty acid and amino acid metabolism, interactions between cells and the extracellular matrix, and extracellular matrix biosynthesis. Furthermore, AF sample subtypes were classified through consensus clustering based on three genes. According to the cumulative distribution function (CDF) plot (Figure 13A) and Delta area plot (Figure 13B), heatmap analysis indicated that the optimal clustering of AF samples was into two groups (Figure 13C). PCA plot further illustrated the distribution of the two clusters (Figure 13D).

3.8 Construction of the gene interaction network and prediction of upstream transcription factors

The gene interaction network of key genes was formed using GeneMANIA (Figure 14A), providing insights into their potential roles in cellular functional regulation, transcriptional control, metabolic processes, and disease progression. The results demonstrated that this gene network is primarily involved in biological processes such as glycolysis and glucose catabolism. Moreover, upstream transcription factors of the key genes were forecast using the JASPAR transcription factor database (Figure 14B). Subsequently, Cytoscape 3.9.0 was employed to generate a related network diagram, illustrating the upstream transcription factors of key genes. The color intensity reflects the density of linked transcription factors.

4 Discussion

Hypoxia induces electrophysiological changes in atrial cells, enhancing atrial excitability and susceptibility, thereby promoting AF. Oxidative stress and inflammatory responses further contribute to AF development by affecting cardiomyocyte function and electrical activity. Moreover, sympathetic activation, vagal responses, and atrial structural remodeling due to prolonged hypoxia [for instance, fibrosis (9)] play critical roles in AF pathogenesis. Hypoxia-inducible factor-1α (HIF-1α), a core molecular marker in hypoxia signaling pathways, has been implicated in AF onset and progression (10, 11), whereas studies on its downstream regulatory molecules remain limited. Therefore, identifying biomarkers related to hypoxia-induced AF is critical for diagnosing and treating this AF subtype. Our study leveraged WGCNA and ML approaches to identify three hypoxia-related key genes (SLC6A6, BGN, and PFKP). Based on these findings, a nomogram model was constructed to assess the diagnostic value of these key genes in predicting hypoxia-associated AF. Additionally, GSEA was conducted to elucidate their biological functions and specific involvement in biological pathways. Based on the current transcriptomic results and previous mechanistic evidence, it is

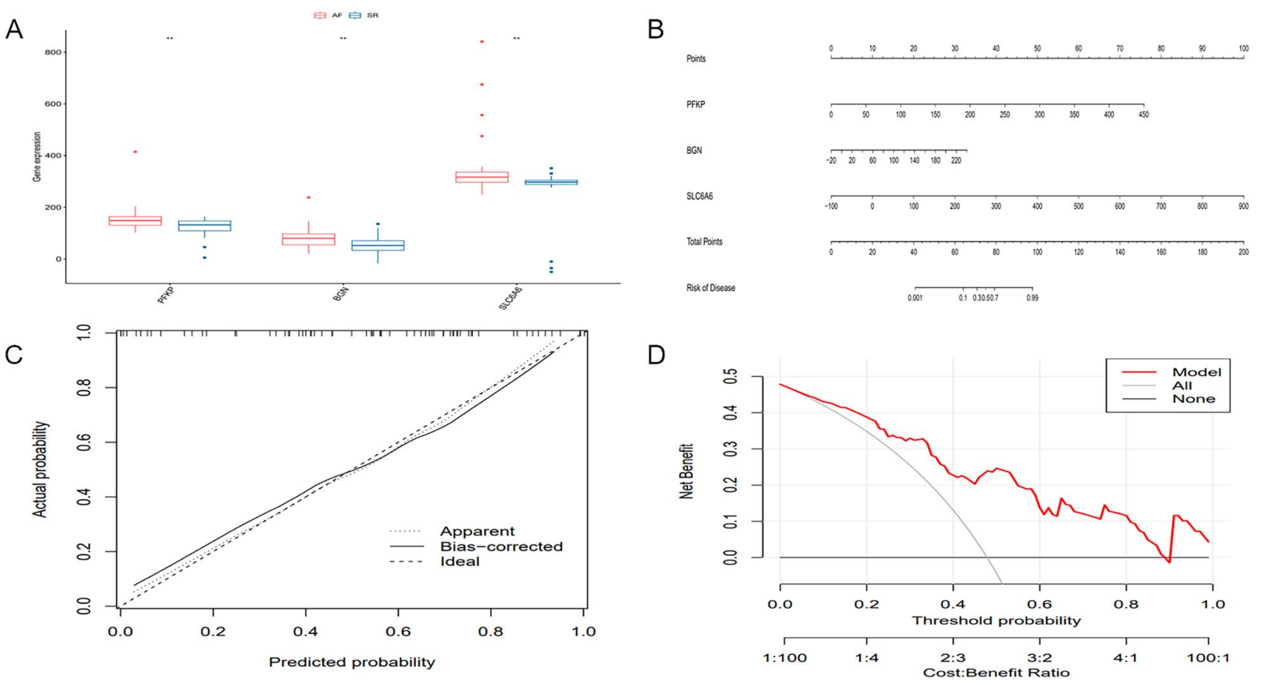


FIGURE 12
Differential expression of key genes and development and evaluation of the predictive model. **(A)** Boxplot of differential gene expression of the key gene set across different sample groups. **(B)** Nomogram model for predicting the probability of AF occurrence. **(C)** Calibration curve of the nomogram model. **(D)** DCA of the nomogram model.

proposed that hypoxia may promote AF pathogenesis through HIF-1α–mediated metabolic and extracellular remodeling pathways. As illustrated in Figure 15, SLC6A6, BGN, and PFKP occupy distinct yet convergent nodes within this regulatory network, potentially linking hypoxia-responsive signaling to atrial structural and electrophysiological alterations. Further experimental investigations are warranted to validate this hypothetical framework.

Enrichment analysis results indicate that the key genes all correlate with the HIF-1α signaling pathway. Studies have demonstrated that AF patients secondary to myocardial hypoxia exhibit elevated HIF-1α levels (12). Furthermore, HIF-1α may contribute to fibrotic remodeling, forming the pathological basis for AF induction (10). SLC6A6 primarily encodes a sodium-ion-dependent taurine transporter, which regulates cellular proliferation, differentiation, and apoptosis (13). Under hypoxic conditions, SLC6A6 is predominantly involved in energy metabolism-related activities. Existing research has shown that SLC6A6 is highly expressed in vascular smooth muscle cells (VSMCs), where its overexpression reduces reactive oxygen species (ROS) production and inhibits the Wnt/β-catenin pathway, thereby suppressing VSMC proliferation, migration, and dedifferentiation (14). Moreover, SLC6A6 overexpression further prevents vascular stenosis and atherosclerosis formation by inhibiting VSMC proliferation, dedifferentiation, and migration (15). Through its regulatory effects on cardiac energy metabolism and myocardial cell stability, SLC6A6 may indirectly participate in AF formation under hypoxic conditions. The BGN gene (Biglycan) encodes a glycosaminoglycan (GAG)-binding protein that primarily interacts with the extracellular matrix (ECM). BGN is expressed in multiple tissues, playing a crucial role in ECM structure and function. In this study, enrichment analysis revealed that BGN is mainly involved in ECM remodeling and energy metabolism-related biological processes, thereby promoting atrial fibrosis, electrical conduction heterogeneity, and oxidative stress, all of which contribute to AF development. The PREDICT-AF study, conducted by Nicoline et al., identified an association between BGN and AF, with elevated BGN expression observed in AF patients. The underlying mechanism is believed to involve fibroblast activation and interaction with collagen. During tissue remodeling in AF patients, increased BGN expression may serve as an early indicator of ECM remodeling in the atria (16). The PFKP gene encodes phosphofructokinase (PFK), a key enzyme in glycolysis that directly influences energy metabolism across various organs (17). In normal cardiac tissue, approximately 70% of energy supply is from fatty acid oxidation (FAO), while the remaining 30% originates primarily from glycolysis and the oxidation of lactate-derived pyruvate, which enters the mitochondria for oxidative phosphorylation (18). However, in the terminal stages of heart failure or under hypoxic conditions, the capacity for FAO is significantly damaged, leading to a marked shift toward increased glucose uptake and utilization (19). Consequently, PFKP is pivotal in the regulation of cardiac energy metabolism. Based on the enrichment analysis results in this study, PFKP is primarily involved in glycolysis-mediated

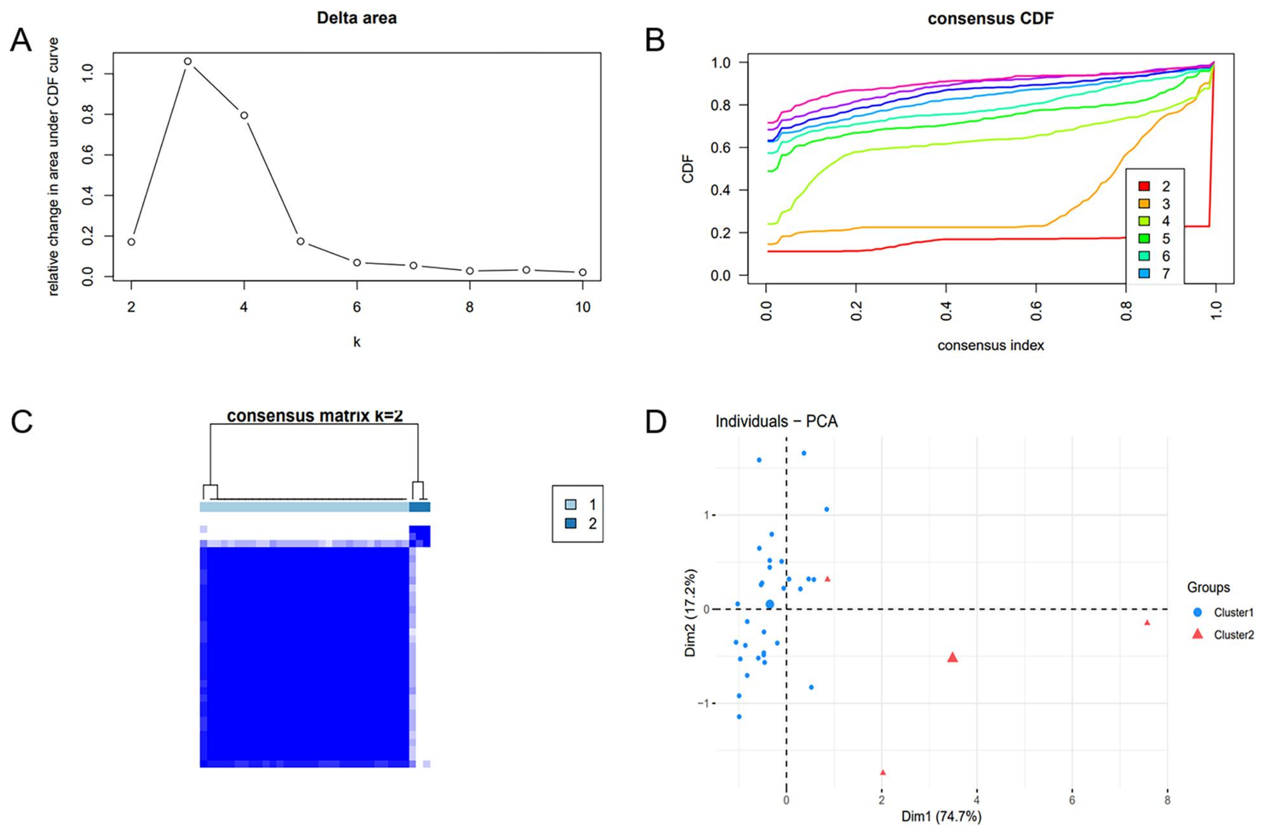


FIGURE 13
Consensus clustering analysis and PCA. (A) Delta area plot. (B) Cumulative distribution function (CDF) plot at $K = 2$. (C) Consensus matrix plot at $K = 2$. (D) PCA of AF samples.

energy supply and myocardial energy metabolism regulation. Aberrant PFKP expression or activity may result in insufficient ATP synthesis, affecting ion channel function, thereby altering myocardial electrophysiology, shortening the effective refractory period, and ultimately promoting AF development (20–22). A study by Marta et al. identified PFKP as a key factor in pathological cardiac hypertrophy (23). Additionally, PFKP is involved in the dynamic balance of focal adhesions, mediating cell-matrix adhesion via integrin regulation and potentially promoting collagen deposition and fibrosis through the TGF-β signaling pathway (24). Research has demonstrated that PFKP overexpression in proximal renal tubular epithelial cells exacerbates glycolysis and renal fibrosis triggered by TGF-β (25). Furthermore, Laurent et al. found that TGF-β1 induces PFKP expression, with a stronger induction observed in the pulmonary arteries of pulmonary arterial hypertension (PAH) individuals and cultured pulmonary arterial endothelial cells. This TGF-β1-induced PFKP expression can be inhibited by pioglitazone (26).

Furthermore, the key genes (SLC6A6, BGN, and PFKP) were positively linked to CD4+ T and B cells, and M2 macrophages and inversely related to DCs. Among the 22 immune cell types analyzed, M2 macrophages and resting DCs were notably elevated in AF samples, whereas Tregs were markedly reduced. Tregs are critical in preserving immunological tolerance and avoiding excessive immune responses. There was an evident reduction in the proportion of Tregs in patients with AF (27), possibly owing to impaired immune regulation and chronic inflammation, which may suppress cell proliferation during the pathogenesis of AF. The role of M2 macrophages in AF development has been well documented (28–30). Our immune infiltration analysis revealed a predisposition of M2 macrophages to infiltrate atrial tissue in AF patients. Upon activation by associated immune-inflammatory responses, this infiltration was accompanied by increased fibrotic area in cardiac tissue, enhanced collagen deposition, and upregulated fibroblast-to-myofibroblast transition, with concurrent activation of the TGF-β/Smad downstream signaling pathway, thereby further promoting fibrosis progression (30). DCs function as antigen-presenting cells and are essential for immune responses. However, how DCs contribute to AF pathogenesis remains unclear. Previous studies have indicated a marked rise in the proportion of immune cells in AF samples, with a significantly higher number of DCs in comparison to samples from individuals with SR (31, 32). In contrast, this study demonstrated a notable increase in the proportion of resting DCs in AF patients, whereas significant changes were not noted in activated DCs across AF and SR groups. Despite the high diagnostic accuracy of the hub genes identified through WGCNA and ML methodologies, which have been

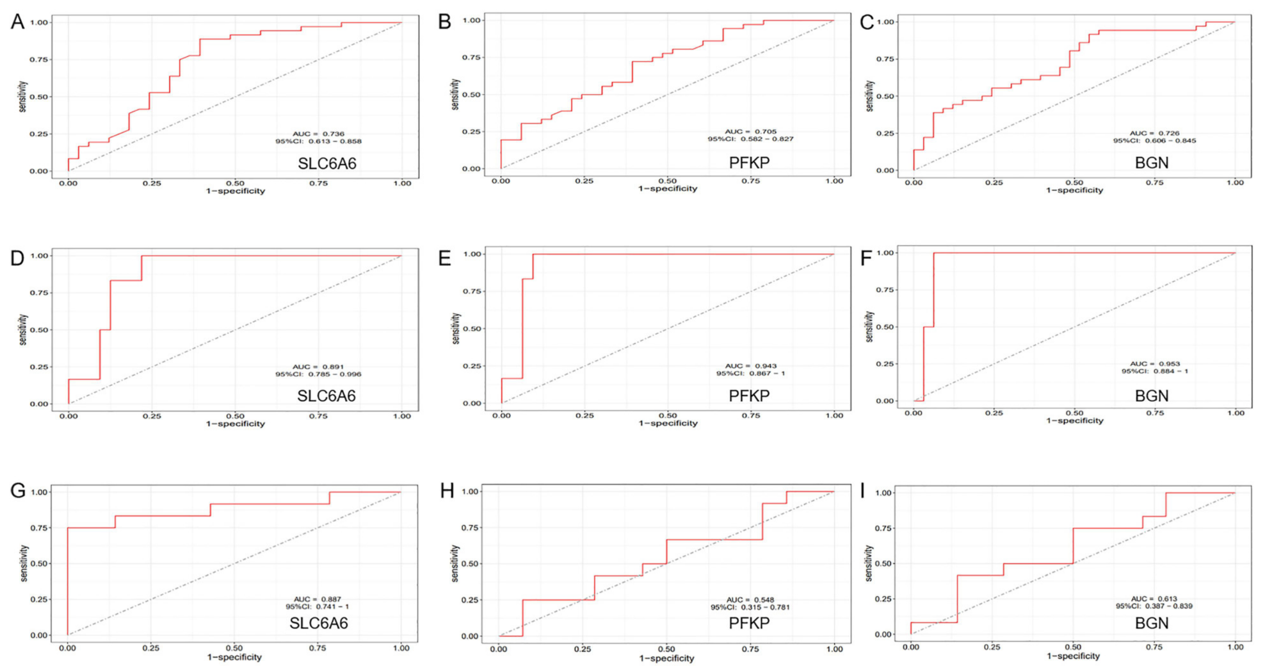


FIGURE 14
ROC Curve Analysis. **(A–C)** ROC curve analysis of the diagnostic efficiency of the key gene set (SLC6A6, PFKP, BGN) in the training dataset (GSE14975 combined with GSE115574). **(D–F)** ROC curve analysis of the diagnostic efficiency of the key gene set (SLC6A6, PFKP, BGN) in the external validation dataset (GSE41177). **(G–I)** ROC curve analysis of the diagnostic efficiency of the key gene set (SLC6A6, PFKP, BGN) in the external validation dataset (GSE79768).

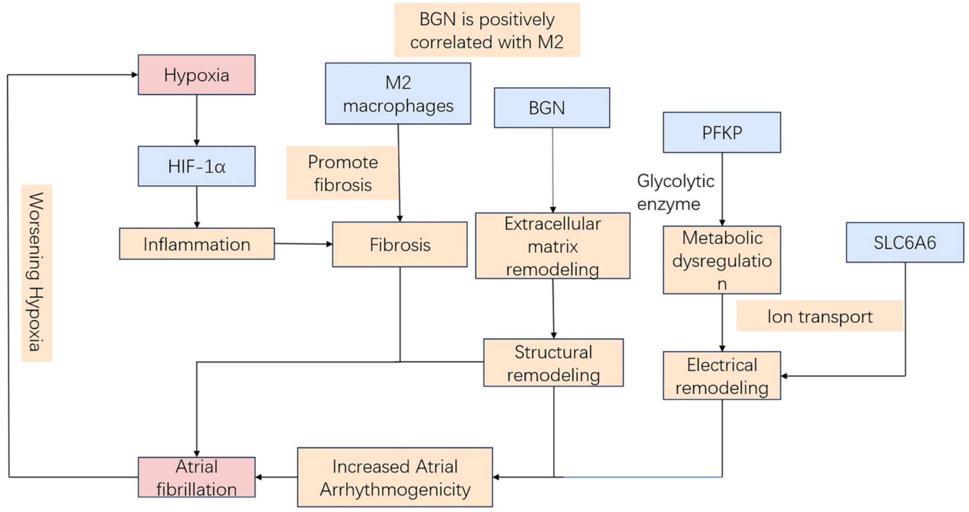


FIGURE 15
Schematic model of hypoxia-related mechanisms in AF.

validated using external datasets, our study has limitations. First, the foregoing findings primarily rely on bioinformatics analyses of hub genes and *in vivo* and *in vitro* experimental validation is lacking. Therefore, conclusions regarding gene expression implicated in the molecular mechanisms of hypoxia-related AF should be interpreted with caution, warranting further experimental confirmation. Second, the possible influence of external clinical characteristics on the data were not accounted for. Additionally, limited genetic data were used in our immune cell infiltration analysis, and *in vivo* and *in vitro* studies were necessitated for unveiling the specific regulatory mechanisms.

4.1 Limitations

This study has limitations. First, Although the expression patterns and biological annotations of SLC6A6, BGN, and PFKP suggest that they may serve as molecular nodes integrating AF-related structural remodeling with hypoxia-responsive transcriptional networks, comprehensive experimental studies are necessary to confirm their causal roles in hypoxia-induced AF pathogenesis. Second, the external validation cohort (GSE41177) exhibited higher AUC values than the training cohort, likely due to the small control sample size and biological heterogeneity between datasets. Hence, these results should be interpreted as supportive rather than definitive evidence of model generalizability. Third, the immune cell infiltration analysis was performed using limited genetic data, and further *in vitro* and *in vivo* experiments are required to elucidate the underlying regulatory mechanisms. Lastly, the transcriptomic data were derived from left atrial tissue without direct measurement of oxygen tension, the identified hypoxia-related genes reflect molecular signatures of hypoxia responsiveness rather than confirmed evidence of actual tissue hypoxia.

5 Conclusion

In conclusion, our findings suggest that SLC6A6, BGN, and PFKP serve as potential hypoxia-related biomarkers and therapeutic targets in AF. Further investigations into immune responses may elucidate the molecular mechanisms underlying this condition and provide novel insights into the management of its comorbidities.

Data availability statement

The gene expression profiling datasets used in our study are publicly available in the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) under the following accession numbers: GSE115574, GSE14975, GSE41177, and GSE79768.

Author contributions

CW: Writing – review & editing. MM: Writing – review & editing. JH: Writing – original draft, Data curation. JC: Visualization, Investigation, Writing – original draft. FD: Supervision, Writing – original draft. LZ: Software, Writing – original draft. ML: Writing – original draft, Investigation. CF: Methodology, Formal analysis, Writing – review & editing. JM: Conceptualization, Methodology, Writing – review & editing, Software. ZJ: Writing – review & editing, Conceptualization, Methodology.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Key Clinical Specialty; Shanghai Science and Technology Grant (Grant No: 22QC1401100); and the Medical industry cross research fund of “Jiaotong University star” program of Shanghai Jiaotong University (Grant No: YG2022ZD008).

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and risk factors in atrial fibrillation (ATRIA) study. *JAMA*. (2001) 285(18):2370–5. doi: 10.1001/jama.285.18.2370

2. Liang F, Wang Y. Coronary heart disease and atrial fibrillation: a vicious cycle. *Am J Physiol Heart Circ Physiol*. (2021) 320(1):H1–12. doi: 10.1152/ajpheart.00702.2020

3. Thomas VB, Almassi GH, Shroyer ALW. Editorial review: guidance for future atrial fibrillation research. *Vessel Plus*. (2024) 8:23. doi: 10.20517/2574-1209.2023.147

4. Goudis CA, Ketikoglou DG. Obstructive sleep and atrial fibrillation: pathophysiological mechanisms and therapeutic implications. *Int J Cardiol*. (2017) 230:293–300. doi: 10.1016/j.ijcard.2016.12.120

5. Saleeb-Mousa J, Nathanael D, Coney AM, Kalla M, Brain KL, Holmes AP. Mechanisms of atrial fibrillation in obstructive sleep apnoea. *Cells*. (2023) 12(12):1661. doi: 10.3390/cells12121661

6. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. *J Biopharm Stat*. (2010) 20(2):281–300. doi: 10.1080/10543400903572753

7. Subramanian M, Wojtusciszyn A, Favre L, Boughorbel S, Shan J, Letaief KB, et al. Precision medicine in the era of artificial intelligence: implications in chronic disease management. *J Transl Med*. (2020) 18(1):472. doi: 10.1186/s12967-020-02658-5

8. Antonio S, Isabella L, Jolanda S, Margarita B, Chiara S, Nicole C, et al. Is artificial intelligence the new kid on the block? Sustainable applications in cardiology. *Vessel Plus*. (2024) 8:12. doi: 10.20517/2574-1209.2023.123

9. Shinagawa K, Shi YF, Tardif JC, Leung TK, Nattel S. Dynamic nature of atrial fibrillation substrate during development and reversal of heart failure in dogs. *Circulation*. (2002) 105(22):2672–8. doi: 10.1161/01.CIR.0000016826.62813.F5

10. Abe I, Teshima Y, Kondo H, Kaku H, Kira S, Ikebe Y, et al. Association of fibrotic remodeling and cytokines/chemokines content in epicardial adipose tissue with atrial myocardial fibrosis in patients with atrial fibrillation. *Heart Rhythm*. (2018) 15(11):1717–27. doi: 10.1016/j.hrthm.2018.06.025

11. Xu Y, Sharma D, Du F, Liu Y. The role of toll-like receptor 2 and hypoxia-induced transcription factor-1α in the atrial structural remodeling of non-valvular atrial fibrillation. *Int J Cardiol*. (2013) 168(3):2940–1. doi: 10.1016/j.ijcard.2013.03.174

12. Ogi H, Nakano Y, Niida S, Dote K, Hirai Y, Suenari K, et al. Is structural remodeling of fibrillated atria the consequence of tissue hypoxia? *Circ J*. (2010) 74(9):1815–21. doi: 10.1253/circj.CJ-09-0969

13. Desforges M, Parsons L, Westwood M, Sibley CP, Greenwood SL. Taurine transport in human placental trophoblast is important for regulation of cell differentiation and survival. *Cell Death Dis*. (2013) 4(3):e559. doi: 10.1038/cddis.2013.81

14. An W, Luong LA, Bowden NP, Yang M, Wu W, Zhou X, et al. Cezanne is a critical regulator of pathological arterial remodelling by targeting β-catenin signalling. *Cardiovasc Res*. (2022) 118(2):638–53. doi: 10.1093/cvr/cvab056

15. Rong Z, Li F, Zhang R, Niu S, Di X, Ni L, et al. Ant-Neointimal formation effects of SLC6A6 in preventing vascular smooth muscle cell proliferation and migration via wnt/β-catenin signaling. *Int J Mol Sci*. (2023) 24(3):3018. doi: 10.3390/ijms24033018

16. van den Berg NWE, Neefs J, Kawasaki M, Nariswari FA, Wesselink R, Fabrizi B, et al. Extracellular matrix remodeling precedes atrial fibrillation: results of the PREDICT-AF trial. *Heart Rhythm*. (2021) 18(12):2115–25. doi: 10.1016/j.hrthm.2021.07.059

17. Wang H, Penaloza T, Manea AJ, Gao X. PFKP: more than phosphofructokinase. *Adv Cancer Res*. (2023) 160:1–15. doi: 10.1016/bs.acr.2023.03.001

18. Stanley WC, Recchia FA, Lopaschuk GD. Myocardial substrate metabolism in the normal and failing heart. *Physiol Rev*. (2005) 85(3):1093–129. doi: 10.1152/physrev.00006.2004

19. Akki A, Smith K, Seymour AM. Compensated cardiac hypertrophy is characterised by a decline in palmitate oxidation. *Mol Cell Biochem*. (2008) 311(1-2):215–24. doi: 10.1007/s11010-008-9711-y

20. Specterman MJ, Aziz Q, Li Y, Anderson NA, Ojake L, Ng KE, et al. Hypoxia promotes atrial tachyarrhythmias via opening of ATP-sensitive potassium channels. *Circ Arrhythm Electrophysiol*. (2023) 16(9):e011870. doi: 10.1161/CIRCEP.123.011870

21. Ma S, Yang S, Xu P, Li W, Wang Y, Wang C, et al. Regulation of ankyrin-G on Nav1.5 channel in hypoxic HL-1 cardiac muscle cells. *Discov Med*. (2024) 36(190):2191–201. doi: 10.24976/Discov.Med.202436190.201

22. Zhang K, Ma Z, Song C, Duan X, Yang Y, Li G. Role of ion channels in chronic intermittent hypoxia-induced atrial remodeling in rats. *Life Sci*. (2020) 254:117797. doi: 10.1016/j.lfs.2020.117797

23. Vigil-Garcia M, Demkes CJ, Eding JEC, Versteeg D, de Ruiter H, Perini I, et al. Gene expression profiling of hypertrophic cardiomyocytes identifies new players in pathological remodelling. *Cardiovasc Res*. (2021) 117(6):1532–45. doi: 10.1093/cvr/cvaa233

24. Akhurst RJ, Hata A. Targeting the TGFβ signalling pathway in disease. *Nat Rev Drug Discov*. (2012) 11(10):790–811. doi: 10.1038/nrd3810

25. Yang S, Wu H, Li Y, Li L, Xiang J, Kang L, et al. Inhibition of PFKP in renal tubular epithelial cell restrains TGF-β induced glycolysis and renal fibrosis. *Cell Death Dis*. (2023) 14(12):816. doi: 10.1038/s41419-023-06347-1

26. Calvier L, Chouvarine P, Legchenko E, Hoffmann N, Geldner J, Borchert P, et al. PPARγ links BMP2 and TGFβ1 pathways in vascular smooth muscle cells, regulating cell proliferation and glucose metabolism. *Cell Metab*. (2017) 25(5):1118–34.e7. doi: 10.1016/j.cmet.2017.03.011

27. Sulzgruber P, Koller L, Winter MP, Richter B, Blum S, Korpak M, et al. The impact of CD4(+)CD28(null) T-lymphocytes on atrial fibrillation and mortality in patients with chronic heart failure. *Thromb Haemost*. (2017) 117(2):349–56. doi: 10.1160/TH16-07-0531

28. Yang M, Xu X, Zhao XA, Ge YN, Qin J, Wang XY, et al. Comprehensive analysis of immune cell infiltration and M2-like macrophage biomarker expression patterns in atrial fibrillation. *Int J Gen Med*. (2024) 17:3147–69. doi: 10.2147/IJGM.S462895

29. Sheng Y, Wang YY, Chang Y, Ye D, Wu L, Kang H, et al. Deciphering mechanisms of cardiomyocytes and non-cardiomyocyte transformation in myocardial remodeling of permanent atrial fibrillation. *J Adv Res*. (2024) 61:101–17. doi: 10.1016/j.jare.2023.09.012

30. Wu Y, Zhan S, Chen L, Sun M, Li M, Mou X, et al. TNFSF14/LIGHT Promotes cardiac fibrosis and atrial fibrillation vulnerability via PI3Kγ/SGK1 pathway-dependent M2 macrophage polarisation. *J Transl Med*. (2023) 21(1):544. doi: 10.1186/s12967-023-04381-3

31. Shiba M, Sugano Y, Ikeda Y, Okada H, Nagai T, Ishibashi-Ueda H, et al. Presence of increased inflammatory infiltrates accompanied by activated dendritic cells in the left atrium in rheumatic heart disease. *PLoS One*. (2018) 13(9):e0203756. doi: 10.1371/journal.pone.0203756

32. Smorodinova N, Bláha M, Melenovský V, Rozsívalová K, Přidal J, Ďurišová M, et al. Analysis of immune cell populations in atrial myocardium of patients with atrial fibrillation or sinus rhythm. *PLoS One*. (2017) 12(2):e0172691. doi: 10.1371/journal.pone.0172691

TYPE Original Research
PUBLISHED 11 December 2025
DOI 10.3389/fnetp.2025.1658470

OPEN ACCESS

EDITED BY
Ulrich Parlitz,
Max Planck Institute for Dynamics and Self-Organization, Germany

REVIEWED BY
Madhan K.,
St. Joseph's College of Engineering, India
Jorge Gomez,
University of Córdoba, Colombia

*CORRESPONDENCE
Silvia Gaftandzhieva,
✉ sissiy88@uni-plovdiv.bg

RECEIVED 02 July 2025
REVISED 19 November 2025
ACCEPTED 19 November 2025
PUBLISHED 11 December 2025

CITATION
Baratpur AZ, Vahdat-Nejad H, Arslan E, Hassannataj Joloudari J and Gaftandzhieva S (2025) Coronary artery disease prediction using Bayesian-optimized support vector machine with feature selection.
Front. Netw. Physiol. 5:1658470.
doi: 10.3389/fnetp.2025.1658470

Coronary artery disease prediction using Bayesian-optimized support vector machine with feature selection

Abdul Zahir Baratpur[1], Hamed Vahdat-Nejad[1], Emrah Arslan[2], Javad Hassannataj Joloudari[1,3,4] and Silvia Gaftandzhieva[5]*

[1]Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran, [2]Departmant of Computer Engineering, Faculty of Engineering, KTO Karatay University, Konya, Türkiye, [3]Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran, [4]Department of Computer Engineering, Bab.C, Islamic Azad University, Babol, Iran, [5]Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria

Introduction: Cardiovascular diseases, particularly Coronary Artery Disease (CAD), remain a leading cause of mortality worldwide. Invasive angiography, while accurate, is costly and risky. This study proposes a non-invasive, interpretable CAD prediction framework using the Z-Alizadeh Sani dataset.

Methods: A hybrid decision tree–AdaBoost method is employed to select 30 clinically relevant features. To prevent data leakage, SMOTE oversampling is applied exclusively within each training fold of a 10-fold cross-validation pipeline. The Support Vector Machine (SVM) model is optimized using Bayesian hyperparameter tuning and compared against Sea Lion Optimization Algorithm (SLOA) and grid search. SHapley Additive exPlanations (SHAP) analysis is utilized to interpret the feature contributions.

Results: The SVM_Bayesian model achieves 97.67% accuracy, 95.45% precision, 100.00% sensitivity, 97.67% F1-score, and 99.00% AUC, outperforming logistic regression (93.02% accuracy, 92.68% F1-score), random forest (95.45% accuracy, 93.33% F1-score), standard SVM (77.00% accuracy), and SLOA-optimized SVM (93.02% accuracy). Ablation studies and Wilcoxon signed-rank tests confirm the statistical superiority of the proposed model.

Discussion: SHAP analysis reveals clinically meaningful feature contributions (e.g., Typical Chest Pain, Age, EFTTE). 95% bootstrap confidence intervals and temporal generalization on an independent test set ensure robustness and prevent overfitting. Future work includes validation on external real-world datasets. This framework provides a transparent, generalizable, and clinically actionable tool for CAD risk stratification, aligned with the principles of network physiology by focusing on interconnected cardiovascular features in predicting systemic disease.

KEYWORDS

coronary artery disease prediction, support vector machine, Bayesian optimization, sealion optimization, feature selection, network physiology

1 Introduction

Cardiovascular diseases (CVDs) have become one of the most prevalent and deadly health challenges in developing countries in recent decades. According to the National Health and Nutrition Examination Survey, between 2013 and 2016, approximately 48% of adults over the age of 20 were affected by some form of CVD, with incidence rates rising progressively with age (Belgiu and Drăguţ, 2016). Despite extensive efforts by medical professionals to prevent, diagnose, and treat these conditions, CVD-related mortality continues to grow. In 2019 alone, an estimated 18.6 million deaths were attributed to heart diseases. According to the World Health Organization (WHO), CAD accounted for approximately 32% of global deaths in 2020, and projections estimate this number will reach 23.6 million annually by 2030.

CAD is primarily caused by the narrowing or blockage of coronary arteries due to plaque buildup, leading to reduced oxygen supply to heart muscles (El-Ibrahimi et al., 2025; Han et al., 2025; Hefti et al., 2025). Risk factors for CAD include hypertension, diabetes, smoking, high cholesterol, poor diet, sedentary lifestyle, psychological stress, and genetic predispositions (Velusamy and Ramasamy, 2021; Mohammedqasim et al., 2022). One of the standard diagnostic tools for CAD is coronary angiography, which offers high precision and spatial clarity for examining coronary vessel structure. However, this method is invasive, costly, and requires highly skilled operators, making it impractical for widespread use as a screening tool.

Non-invasive alternatives such as electrocardiography (ECG) and echocardiography are commonly used in clinical evaluations, though they lack the sensitivity and accuracy offered by invasive coronary angiography (Alizadehsani et al., 2022). In response to these limitations, researchers have increasingly turned to artificial intelligence-based machine learning (ML) methods to improve the diagnostic capabilities of non-invasive approaches. ML algorithms have proven effective in diverse fields, including big data analytics, cybersecurity, IoT, and particularly in medical image analysis and disease prediction (Nasarian et al., 2020).

Several studies have investigated CAD Prediction using ML algorithms. For instance, Alizadeh Sani et al. applied C4.5 decision tree and Bagging classifiers to a dataset of 303 numerical samples for CAD detection (Alizadehsani et al., 2013). Similarly, Hassan Nataj et al. utilized random forest-based feature ranking to identify important predictive features (Joloudari et al., 2020; Arabasadi et al., 2017) experimented with artificial neural networks and genetic algorithms both independently and in combination on the same dataset.

The success of any ML-based disease detection system largely depends on the algorithm used and the number of predictive features selected (Fajri et al., 2022). Feature selection, which involves identifying the most relevant input variables, significantly enhances model accuracy and generalizability (Velusamy and Ramasamy, 2021). Feature selection techniques are broadly categorized into three types: filter, wrapper, and embedded methods. These methods aim to reduce the dimensionality of datasets while preserving essential information (Zebari et al., 2020).

High-dimensional datasets—those with numerous input variables—pose serious challenges for ML models. As feature dimensionality increases, models become more complex, making it harder to optimize and increasing the risk of overfitting. Overfitting occurs when a model learns training data too closely and performs poorly on unseen data. Dimensionality reduction helps alleviate these issues by simplifying models and enhancing their generalization capabilities.

Numerous studies have focused on effective feature selection to reduce dataset dimensions. For example (Hassannataj Joloudari et al., 2022), applied a genetic algorithm for optimization, while (Velusamy and Ramasamy, 2021) used the Boruta wrapper method (Jin and Li, 2022). implemented recursive feature elimination using random forests, and (Mohammedqasim et al., 2022) employed whale optimization in combination with k-nearest neighbor algorithms.

Another common challenge in CAD-related datasets is class imbalance (Nasarian et al., 2020), where samples of one class significantly outnumber those of others. This imbalance, often seen in scenarios like fraud detection or rare disease Prediction, can lead ML models to favor the majority class, reducing overall accuracy. Most ML algorithms are designed to minimize overall error without explicitly considering class distribution, thereby degrading performance on the minority class.

To address this, several studies have incorporated resampling techniques. For example (Nasarian et al., 2020), employed Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) to balance class distributions. Similarly (Gupta et al., 2022; Mohammedqasim et al., 2022), and Velusamy and Ramasamy (Velusamy and Ramasamy, 2021) utilized SMOTE to enhance prediction accuracy on imbalanced datasets.

In current research, the Z-Alizadeh Sani CAD dataset is statistically analyzed, and missing data are examined. During preprocessing, the dataset is normalized, and feature selection is performed using decision tree and AdaBoost algorithms. The dataset is then split into training and testing subsets using 10-fold cross-validation. To address class imbalance, the SMOTE algorithm is employed to synthetically balance the target class. For classification, a Support Vector Machine (SVM) model is used, with its hyperparameters optimized via Bayesian optimization (Frazier, 2018) and compared against the performance of the Sea Lion Optimization Algorithm (SLOA) (Masadeh et al., 2019; Kumaraswamy and Poonacha, 2021). The final models are evaluated based on accuracy, sensitivity, and F1-score.

1.1 Contribution of the study

While previous studies have explored CAD detection using ML methods, this study introduces several innovations that enhance accuracy, interpretability, statistical rigor, and clinical applicability.

1. Combining decision tree and AdaBoost algorithms for effective and interpretable feature selection, reducing dimensionality to 30 clinically relevant features without sacrificing predictive power.
2. Employing Bayesian optimization to fine-tune SVM hyperparameters, achieving 97.67% accuracy, 100.00% sensitivity, and 99.00% AUC, outperforming standard SVM (77.00% accuracy), SLOA-optimized SVM (93.02%), logistic

regression (93.02%), and random forest (95.45%), while enabling a systematic and efficient search compared to grid or random methods.
3. Including a direct comparison with the SLOA, a recent metaheuristic, demonstrating Bayesian optimization's superior efficiency and performance.
4. Addressing class imbalance using SMOTE within a pipeline-based 10-fold cross-validation framework, preventing data leakage and ensuring robust detection of minority cases.
5. Evaluating the model using a comprehensive metric suite (Accuracy, Precision, Sensitivity, F1-score, AUC), reporting mean ± std across folds, 95% bootstrap confidence intervals, and temporal generalization on an independent held-out set.
6. Providing clinical interpretability via SHapley Additive exPlanations (SHAP) analysis, highlighting Typical Chest Pain, Age, and EF-TTE as top predictors, fully aligned with ESC/AHA guidelines, and including calibration assessment (Brier score = 0.088) and cost-sensitive threshold optimization.
7. Delivering collectively a transparent, generalizable, and deployment-ready framework for non-invasive CAD risk stratification.

1.2 The workflow of the study

The study follows a structured six-phase workflow: related works, methodology, results, interpretability analysis, clinical validation, and conclusion with future directions. Recent CAD prediction studies were reviewed to identify challenges and gaps. Methodology covers data preprocessing, hybrid decision tree–AdaBoost feature selection (30 features), pipeline-based SMOTE for imbalance, 10-fold cross-validation, and training of logistic regression, random forest, and SVM with Bayesian hyperparameter optimization compared to SLOA and grid search. Results include performance metrics, ablation studies, Wilcoxon tests, and temporal generalization, confirming Bayesian-optimized SVM superiority (97.67% accuracy, 100.00% sensitivity). Interpretability uses SHAP for feature explanations. Clinical validation assesses calibration, Brier score, and thresholds. Conclusion summarizes findings, implications, and future directions like external validation and trials.

2 Related works

Today, one of the most critical challenges facing human societies is the prevalence of widespread diseases, many of which lead to high mortality rates. According to the World Health Organization, cardiovascular diseases, particularly CAD, are among the leading causes of death globally, especially in middle-aged and older populations. Currently, various clinical techniques such as exercise stress testing, chest X-rays, Computed Tomography (CT) scans, cardiac magnetic resonance imaging (MRI), coronary angiography, and electrocardiography (ECG) are employed to assess the severity of heart conditions.

In recent years, numerous studies have focused on the application of artificial intelligence techniques for CAD detection using clinical datasets (Liu et al., 2025). Jain and Lee proposed a CAD detection model based on the Whale Optimization Algorithm (WOA) integrated with k-nearest neighbors (k-NN) for feature selection and a stacked model for Prediction (Jin and Li, 2022). The WOA was applied to perform continuous-to-binary transformation and identify optimal feature subsets for each primary coronary artery. Subsequently, a two-layer stacked model was developed to diagnose the left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA). Their method selected 17 features for each Prediction task and achieved classification accuracies of 89.68%, 88.71%, and 85.81% for LAD, LCX, and RCA respectively.

Nasarian et al. introduced a novel hybrid feature selection algorithm named HFS2, which was applied to the Nasarian CAD dataset (Nasarian et al., 2020). This dataset included not only clinical variables but also workplace and environmental features. To address data imbalance, SMOTE and ADASYN aproaches were used. Various classifiers, including Decision Tree, Gaussian Naive Bayes, Random Forest, and XGBoost were employed. Their proposed method, when combined with SMOTE and XGBoost, achieved a classification accuracy of 81.23%. Moreover, the approach was validated on other well-known CAD datasets, yielding classification accuracies of 83.94%, 81.58%, and 92.58% on Hungarian, Long-Beach-VA, and Z-Alizadeh Sani datasets, respectively.

(Alizadehsani et al., 2013) applied Decision Tree C4.5 and Bagging classifiers to a dataset of 303 numerical samples for CAD detection. Feature selection was conducted using information gain and Gini index. The Bagging classifier, when combined with these feature selection methods, outperformed C4.5, achieving classification accuracies of 79.54%, 65.09%, and 66.31% for detecting stenosis in three major coronary arteries. In comparison, the C4.5 algorithm achieved respective accuracies of 76.56%, 63.10%, and 63.38%.

(Arabasadi et al., 2017) used neural networks and genetic algorithms both individually and in combination for CAD Prediction on a dataset of 303 samples. Feature selection employed several techniques including SVM weighting, Gini index, information gain, and principal component analysis (PCA). Results showed that the neural network alone achieved an accuracy of 84.62%, while the hybrid neural-genetic algorithm reached 93.85% using 10-fold cross-validation.

(Khozeimeh et al., 2023) proposed an active learning method combined with an ensemble of classifiers for CAD detection. Their framework incorporated four classifiers: three focused on diagnosing stenosis in the three main coronary arteries, and one to predict the overall presence of CAD. Among 19 active learning algorithms, their ensemble method paired with an SVM classifier achieved the best performance with an accuracy of 97.01%.

(Hassannataj Joloudari et al., 2022) introduced a novel hybrid machine learning model combining Genetic Algorithm and Analysis of Variance (ANOVA) as the kernel function for SVM. This model was evaluated on the Z-Alizadeh Sani dataset, with feature selection handled by a genetic optimizer. Additionally, multiple SVM variants such as ANOVA-SVM, linear SVM, and RBF-kernel SVM—were applied. Using 10-fold cross-validation and 31 selected features, an accuracy of 89.45% was achieved.

In another study, a two-level genetic algorithm was integrated with NuSVM to create a hybrid model named N2Genetic-NuSVM, tested on 303 samples (Abdar et al., 2019). The dual-level genetic algorithm simultaneously optimized the SVM parameters and selected relevant features. The model achieved a CAD detection accuracy of 93.08% using 10-fold cross-validation.

(Eyupoglu and Karakuş, 2024), focusing on the challenge of feature redundancy in CAD Prediction, demonstrated that reducing features while maintaining accuracy can facilitate early detection. Their study combined eight search techniques with PCA and AdaBoostM1, achieving 91.8% accuracy on the Z-Alizadeh Sani dataset using only five features: age, blood pressure, typical chest pain, inverted T wave, and wall motion abnormality.

(Hashemi et al., 2024) also utilized the Z-Alizadeh Sani dataset for CAD detection. By applying genetic algorithms for feature selection in neural networks, they achieved an accuracy of 94.71%, sensitivity of 96.29%, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 93.5%, demonstrating strong diagnostic performance using machine learning techniques.

(Koloi et al., 2024) introduced a machine learning framework for early-stage CAD prediction using clinical and laboratory test data from 19,826 patients. They demonstrated that their approach could accurately identify early CAD cases, achieving a classification accuracy of 79% and an AUC-ROC of 0.79. The study highlighted the potential of machine learning in enhancing early diagnostic capabilities for CAD.

(Brendel et al., 2025) applied deep learning to detect CAD using photon-counting coronary CT angiography (PC-CCTA). Their deep learning model achieved an AUC-ROC of 0.90 at the patient level and 0.92 at the vessel level, indicating high diagnostic performance. The study underscored the effectiveness of combining PC-CCTA imaging with advanced learning algorithms for accurate CAD diagnosis.

(Wang et al., 2024) developed an explainable CAD prediction model using Automated Machine Learning (AutoML). The AutoGluon-based ensemble model achieved an accuracy of 91.67% and an AUC of 0.9562 in 4-fold cross-bagging. The integration of SHAP values provided transparency in feature importance, enhancing the interpretability and trustworthiness of the model in clinical applications.

(Akella and Akella, 2021) evaluated six open-source machine learning algorithms for CAD prediction using the Cleveland dataset. Among the tested models, the neural network achieved the highest accuracy of 93% and a recall of 93.8%. The study underscored the potential of accessible machine learning-based CAD prediction tools for enhancing diagnostic capabilities in clinical settings.

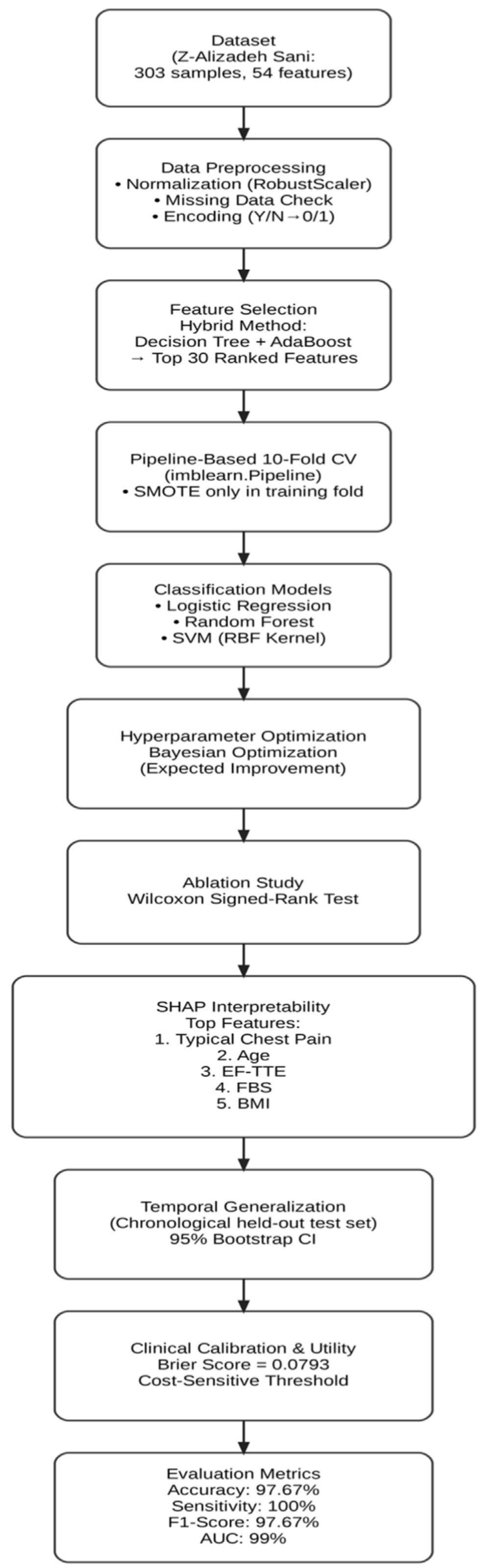


FIGURE 1
The proposed methodology.

Despite the growing body of research on CAD Prediction using artificial intelligence, several open challenges remain in improving diagnostic accuracy, reducing feature dimensionality, and effectively handling uncertainty in medical predictions. Prior studies have primarily focused on integrating optimization algorithms such as genetic algorithms and whale optimization with classifiers like SVM, decision tree, and XGBoost. Although these approaches have reported respectable accuracies ranging from 85% to 93%, several methodological limitations persist. For instance, studies such as (Hassannataj Joloudari et al., 2022; Abdar et al., 2019) have employed evolutionary algorithms to tune the hyperparameters of SVMs, but probabilistic modeling using Bayesian theory has largely been overlooked. Incorporating prior distributions and Bayesian inference could offer a more principled alternative to stochastic parameter search, potentially reducing overfitting and improving model robustness. In addition, feature selection techniques used in earlier research such as principal component analysis and the Gini index are typically deterministic and fail to account for the inherent uncertainty in medical data. A Bayesian-driven evaluation of feature

TABLE 1 Quantitative features of the Z-Alizadehsani dataset.

Feature	Mean	Std. Dev	Min	25%	Median (50%)	75%	Max
Age	58.90	10.39	30.00	51.00	58.00	66.00	86.00
Weight	73.83	11.99	48.00	65.00	74.00	81.00	120.00
Length	164.72	9.33	140.00	158.00	165.00	171.00	188.00
BMI	27.25	4.10	18.12	24.51	26.78	29.41	40.90
BP	129.55	18.94	90.00	120.00	130.00	140.00	190.00
PR	75.14	8.91	50.00	70.00	70.00	80.00	110.00
FBS	119.18	52.08	62.00	88.50	98.00	130.00	400.00
CR	1.06	0.26	0.50	0.90	1.00	1.20	2.20
TG	150.34	97.96	37.00	90.00	122.00	177.00	1,050.00
LDL	104.64	35.40	18.00	80.00	100.00	122.00	232.00
HDL	40.23	10.56	15.90	33.50	39.00	45.50	111.00
BUN	17.50	6.96	6.00	13.00	16.00	20.00	52.00
ESR	19.46	15.94	1.00	9.00	15.00	26.00	90.00
HB	13.15	1.61	8.90	12.20	13.20	14.20	17.60
K	4.23	0.46	3.00	3.90	4.20	4.50	6.60
Na	141.00	3.81	128.00	139.00	141.00	143.00	156.00
WBC	7562.05	2413.74	3700.00	5800.00	7100.00	8800.00	18000.00
Lymph	32.40	9.97	7.00	26.00	32.00	39.00	60.00
Neut	60.15	10.18	32.00	52.50	60.00	67.00	89.00
PLT	221.49	60.80	25.00	183.50	210.00	250.00	742.00
EF-TTE	47.23	8.93	15.00	45.00	50.00	55.00	60.00

relevance, based on posterior probabilities, could yield more reliable and interpretable feature subsets. Moreover, while some studies, such as (Nasarian et al., 2020), have used oversampling techniques like SMOTE and ADASYN to manage data imbalance, relatively little attention has been paid to combining these sampling methods with probabilistic weighting schemes that could more accurately reflect the uncertainty associated with minority class samples during model training. Another significant issue is that most existing approaches handle feature selection and model optimization as separate processes.

In the current study, a unified Bayesian framework that jointly selects features and tunes model parameters was proposed to improve classification accuracy, enhance model interpretability, and better handle the uncertainty inherent in complex medical datasets such as CAD. This approach was compared with the SLOA, and results demonstrated that the Bayesian-optimized SVM outperformed SLOA, confirming its superior performance and reliability.

3 Proposed methodology

In this study, a Bayesian-optimized SVM was employed for the prediction of heart disease, using a feature selection approach to identify the most significant attributes within the Z-Alizadeh Sani dataset. The proposed methodology consists of four main phases: data preparation, data preprocessing, classification modeling, and hyperparameter optimization. After completing these phases, a final decision is made regarding the presence or absence of the CAD. The entire process is illustrated in Figure 1, and each phase is described in detail in the following sections.

In addition to the workflow shown in Figure 1, a redesigned system architecture diagram has been included to provide a more comprehensive illustration of the methodology. This architecture explicitly depicts the flow of data, the preprocessing steps, the feature selection stage, and the role of each algorithm used in classification and optimization.

3.1 Phase 1: data preparation

The dataset used in this research is the Z-Alizadehsani dataset, which is publicly available through the UCI Machine Learning Repository[1]. This dataset consists of 303 samples, including

[1] https://archive.ics.uci.edu/dataset/412/z+alizadeh+sani

TABLE 2 The categorical variables.

Feature name	Number of unique values	Unique values
Sex	2	Male, Female
DM	2	0, 1
HTN	2	1, 0
Current_Smoker	2	1, 0
EX-Smoker	2	1, 0
Obesity	2	Y, N
CR	2	1, 0
CVA	2	N, Y
Airway_disease	2	N, Y
Thyroid_Disease	2	N, Y
CHF	2	N, Y
DLP	2	Y, N
Weak_Peripheral_Pulses	2	N, Y
Lung_rales	2	N, Y
Systolic_Murmur	2	N, Y
Diastolic_Murmur	2	N, Y
Typical_Chest_Pain	2	N, Y
Dyspnea	2	N, Y
Atypical	2	N, Y
Nonanginal	2	N, Y
Exertional_CP	2	N, Y
Q_Wave	2	0, 1
ST_Elevation	2	0, 1
ST_Depression	2	1, 0
Inversion_T	2	0, 1
LVH	2	N, Y
Poor_R_Progression	2	N, Y
Target variable (Cath)	2	CAD, Normal

216 patients diagnosed with CAD and 87 healthy individuals, described by 54 features. The dataset encompasses clinical characteristics, signs and symptoms, echocardiographic data, and laboratory test results.

The study variables are divided into dependent and independent variables. The target variable, labeled as cath, is the dependent variable and represents the presence or absence of disease. The independent variables refer to the input features extracted from the dataset, categorized as follows:

- Clinical Features: Age, weight, gender, Body Mass Index (BMI), diabetes mellitus, hypertension, current smoker, ex-smoker, family history, obesity, chronic renal failure, cerebrovascular accident, airway disease, thyroid disease, congestive heart failure, dyslipidemia.
- Signs and Symptoms: Systolic and diastolic blood pressure, heart rate (beats per minute), edema, weak peripheral pulse, pulmonary rales, systolic murmur, diastolic murmur, typical chest pain, dyspnea, functional class, atypical symptoms, non-anginal chest pain, exertional chest pain.
- Echocardiography: Rhythm, Q wave, ST elevation, ST depression, T wave inversion, left ventricular hypertrophy, poor R wave progression.
- Laboratory Tests and Echocardiographic Parameters: Fasting blood sugar (mg/dL), creatinine (mg/dL), triglycerides (mg/dL), low-density lipoprotein (mg/dL), high-density lipoprotein (mg/dL), blood urea nitrogen (mg/dL), erythrocyte sedimentation rate (mm/h), hemoglobin (g/dL), potassium (mEq/L), sodium (mEq/L), white blood cell count (cells/mL), lymphocyte percentage, neutrophil percentage, platelet count (×1,000/mL), ejection fraction (%), regional wall motion, abnormality score (numeric), severity of valvular heart disease.

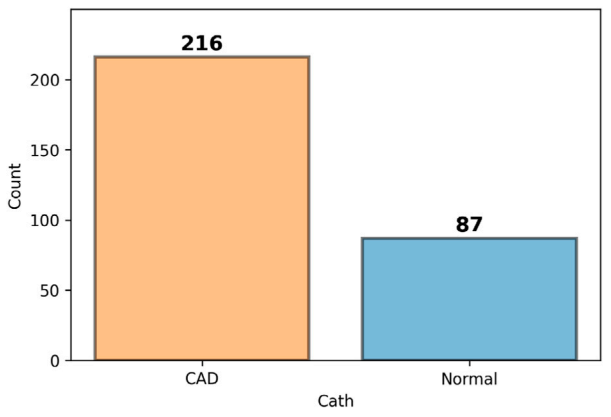


FIGURE 2
Class distribution of the target variable (Cath), indicating imbalance across categories.

Within the system architecture, this dataset serves as the input layer, from which clinical, echocardiographic, and laboratory features are extracted. The architecture highlights this stage as the foundation upon which all subsequent analysis and modeling steps are built.

3.2 Phase 2: data preprocessing

3.2.1 Statistical analysis

During the preprocessing phase, it was identified that among the 55 features in the dataset, 54 are independent variables and 1 is the dependent variable. Statistical analysis helped reveal hidden patterns and correlations among the data. Furthermore, 21 of the features were found to be quantitative in nature. These numerical features are listed in Table 1, which presents the quantitative attributes of the Z-Alizadehsani dataset. The categorical variables examined in this study include 31 items, which are presented in Table 2. In addition to

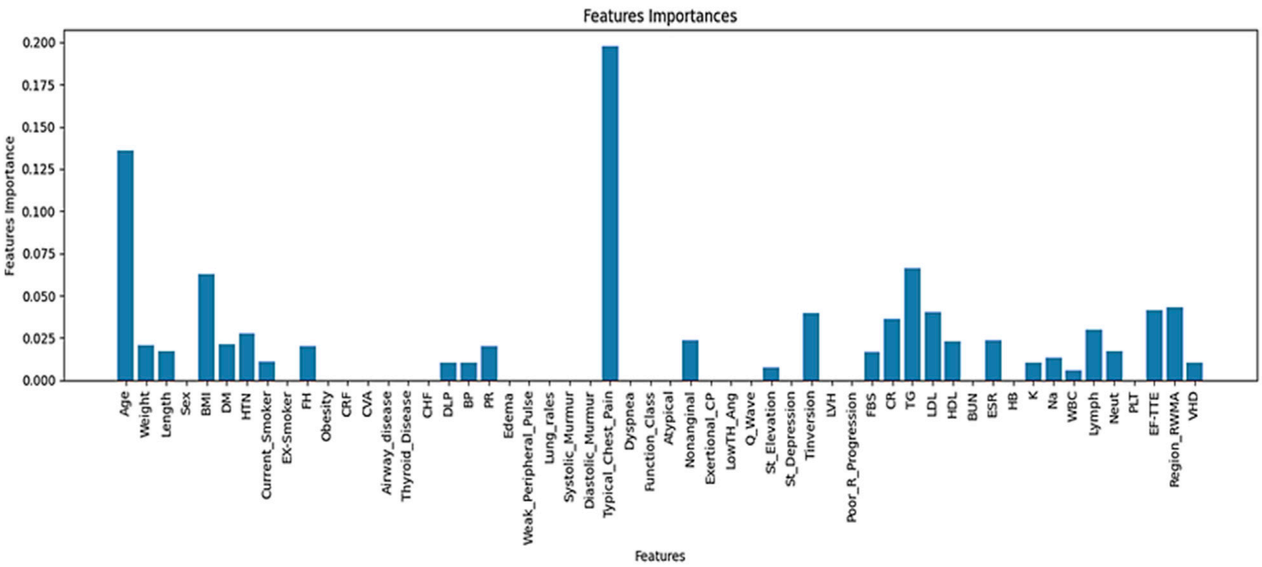


FIGURE 3
Ranked important features and their relative contributions to the prediction task.

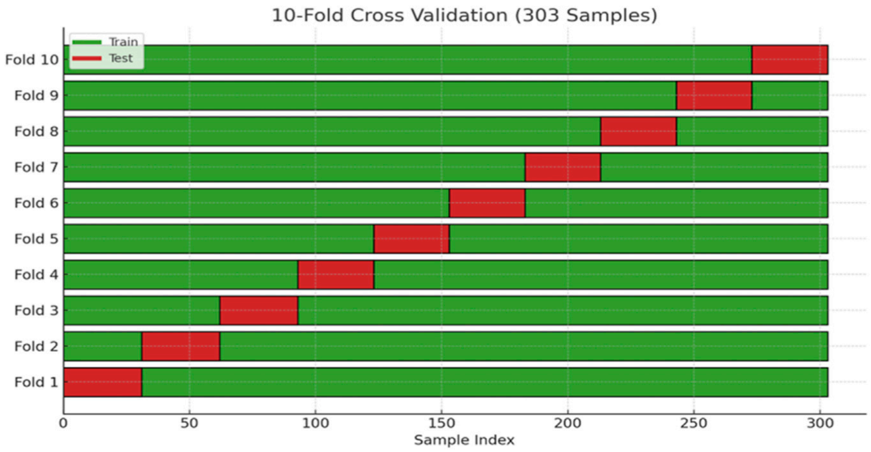


FIGURE 4
Illustration of the dataset splitting into training and testing subsets using stratified 10-fold cross-validation.

the categorical variables, three ordinal variables were also analyzed in this study. The first is Function_Class, which contains four unique values: 0, 1, 2, and 3. The second variable, Region_RWMA, includes five distinct levels: 0, 1, 2, 3, and 4. The third, VHD, consists of four qualitative levels: 'N' (normal), 'mild', 'Moderate', and 'Severe'. These variables were treated as ordinal data in subsequent analyses.

3.2.2 Missing data analysis

To assess the presence of missing values, the dataset was examined using the *pandas* library in Python. The analysis revealed that there were no missing entries in the dataset; all data points were fully recorded. As a result, no imputation or deletion procedures were required.

3.2.3 Data balance evaluation

An initial inspection of the dataset indicated that the target variable, CAD, was imbalanced. The number of instances across the different classes of this variable showed significant disparity. This imbalance is illustrated in Figure 2, highlighting the potential impact on classification performance and underscoring the need for appropriate handling strategies in the modeling phase.

3.2.4 Data normalization

The normalization process began by converting non-numeric features—such as those represented by textual values like 'y' and 'n'—into binary numerical values (1 and 0). This transformation was performed after completing the initial statistical analysis. Following

TABLE 3 Model hyperparameter settings based on Bayesian optimization.

Model name	Configuration
Random Forest	RandomForestClassifier (n_estimators = 380, max_depth = 22, min_samples_split = 2, min_samples_leaf = 13, criterion = ’gini’)
Logistic Regression	LogisticRegression (penalty = ’l2′, C = 58.48737264443094)
Support Vector Machine	SVC (C = 239.59501536334488, kernel = ’rbf’, gamma = 0.36055928693321015)

this encoding step, the dataset underwent a scaling process to bring all features, including numerical ones, into a standardized range between 0 and 1. For this purpose, the RobustScaler method was applied. Unlike the StandardScaler, which relies on mean and standard deviation and is sensitive to outliers, RobustScaler is designed based on robust statistics, specifically the median and interquartile range (IQR). This approach centers each feature by subtracting its median and then scales it by dividing by the IQR (the difference between the 75th and 25th percentiles). By doing so, it effectively minimizes the influence of outliers while preserving the underlying structure of the data distribution. This makes RobustScaler particularly suitable for datasets with skewed distributions or extreme values, ensuring that key data characteristics are maintained during normalization (Prusty et al., 2022).

The Robust Normalization formula, as shown in Equation 1, is used to scale data in a way that minimizes the influence of outliers.

$$X_{robust} = \frac{(X - \text{Median})}{\text{IQR}} \tag{1}$$

In this formula, X represents the original value from the dataset that we aim to normalize. The term Median refers to the median of all values in the corresponding column, which serves as a robust measure of central tendency. Q1 and Q3 denote the first and third quartiles, respectively. Q1 is the value below which 25% of the data fall, while Q3 is the value below which 75% of the data fall. The interquartile range, IQR = Q3 - Q1, captures the spread of the central 50% of the data and helps reduce the impact of extreme values. Finally, X_robust is the normalized value obtained after applying the robust normalization process. This method is especially useful when dealing with datasets that contain outliers, as it relies on measures (median and IQR) that are less sensitive to such anomalies compared to mean and standard deviation.

3.2.5 Feature selection

After completing the initial preprocessing and transformation phases, feature selection techniques were applied to reduce computational costs. In recent years, hybrid and ensemble methods for feature selection have shown promising results, proving effective in identifying relevant attributes within datasets (Belgiu and Drăguţ, 2016).

TABLE 4 Performance results of machine learning models.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)	AUC (%)
Logistic Regression	93.02	95.00	90.48	92.68	97.92
Random Forest	95.45	91.30	95.45	93.33	98.00
Bayesian-Optimized SVM with AdaBoost + Decision Tree feature selection	97.67	95.45	100.00	97.67	99.00

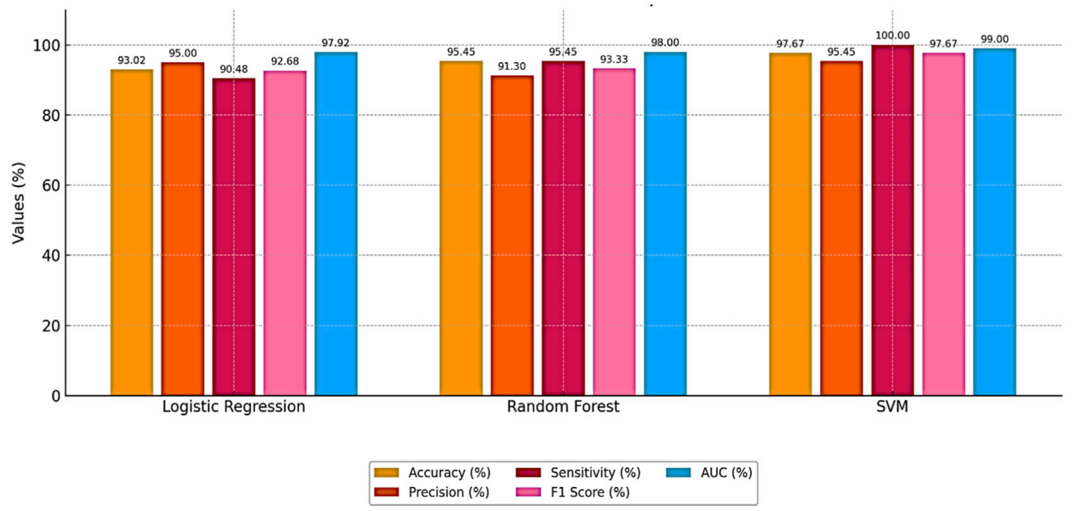


FIGURE 5
Comparative performance of the models.

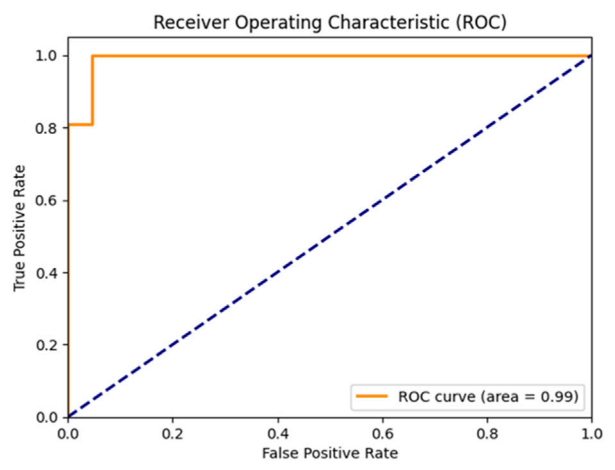


FIGURE 6
ROC curve of the Bayesian-optimized SVM.

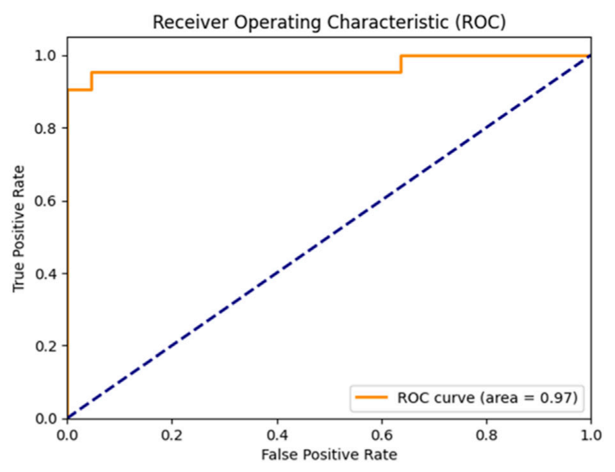


FIGURE 7
ROC curve of the sea lion optimized SVM.

TABLE 5 Accuracy comparison among different SVM variants.

Method	Accuracy (%)
Standard SVM	77.00
SLOA Optimized SVM	93.02
Bayesian Optimized SVM (Proposed)	97.67

Hybrid techniques typically combine two different feature selection strategies, such as wrapper and filter methods, or use two approaches with similar evaluation criteria. By merging the strengths of each method, these techniques enhance the effectiveness of the feature selection process. A commonly used hybrid strategy involves integrating filter and wrapper methods—where filter approaches evaluate features independently of learning models using statistical metrics, while wrapper methods assess feature subsets based on model performance. This combination enables fast elimination of irrelevant features via filtering, followed by more refined selection using wrapper-based evaluation, resulting in improved model performance and reduced computation time.

A typical hybrid pipeline might consist of the following steps (Prusty et al., 2022; Eyupoglu and Karakuş, 2024; Hashemi et al., 2024):

- Initial Filtering: A filter method such as mutual information (for nonlinear dependencies) or Pearson correlation (for linear relationships) is used to remove features with minimal relevance to the target variable. This step reduces the dimensionality by eliminating low-importance features.
- Wrapper-Based Refinement: Once the feature space is reduced, a wrapper method is used for more precise evaluation. For example, a genetic algorithm coupled with a machine learning model (such as a neural network) can be employed to explore optimal feature combinations, aiming to maximize the model's predictive power.

In contrast, ensemble methods attempt to form clusters of feature subsets and aggregate their outputs. These methods often rely on subsampling strategies, applying a given selection algorithm across multiple subsets of the data, then integrating the results to form a more robust feature set.

In general, feature selection plays a critical role in identifying important variables within a dataset, assigning higher scores to more influential features while downranking less informative ones. Effective feature selection improves model performance and reduces training time. However, different selection methods may yield different results, as a feature deemed significant by one technique may receive a lower score from another. Therefore, assigning consistent and reliable importance scores can be challenging. Despite this, hybrid methods offer flexible and effective solutions, allowing researchers to adapt the feature selection process to the specific properties of their datasets and achieve better learning outcomes.

In this study, feature selection was performed using a hybrid approach based on the combination of AdaBoost and Decision Tree algorithms. The joint importance scores provided by both models were used to identify the most relevant features, as described below:

- Model Training: Both the Decision Tree and AdaBoost models were trained on the dataset consisting of the feature matrix (X) and the target variable (Y).
- Importance Scoring: After training, each model generated importance scores for the features. Decision Trees ranked features based on their contribution to splitting nodes, while AdaBoost evaluated them according to their role in gradient boosting.
- Score Integration: The core idea of the hybrid approach lies in combining the importance scores from both models. This was achieved by averaging the individual scores, aiming to balance the biases of each model and provide a more stable and flexible assessment of feature relevance.
- Feature Ranking: Features were then ranked based on the combined scores, producing a sorted list in descending order of importance.
- Utilization of Ranked Features: This ranked list can be used for various purposes, such as reducing the feature space to

TABLE 6 Comparative evaluation with previous studies on Z-Alizadeh Sani dataset.

Authors	Methodology	Number of features	Accuracy	Sensitivity	Precision	AUC
Fajri et al. (2022)	Hybrid Feature Selection with Q-learning and Bee Algorithm	24	90.10%	N/A	94%	94.10%
Joloudari et al. (2020)	Random Forest	40	91.47%	N/A	N/A	96.70%
Kiliç and Keleş (2018)	Artificial bee colony algorithm	16	89.43%	N/A	N/A	N/A
Napi'ah et al. (2023)	Gradient Boosted Trees + Monarch Butterfly Optimization	31	90.26%	80.79%	86.82%	87.33%
Abdar et al. (2019)	N2Genetic + nuSVM	29	93.08%	N/A	N/A	N/A
Hashemi et al. (2024)	MLP with Genetic Algorithm	24	94.71%	96.29%	N/A	93.50%
Nasarian et al. (2020)	XGBoost + SMOTE	38	92.58%	92.99%	92.59%	N/A
This Study (Proposed)	Bayesian-Optimized SVM with AdaBoost + Decision Tree feature selection	29	97.67%	100%	95.45%	99%

improve computational efficiency, emphasizing the most relevant features to boost model performance, or gaining insights into the key drivers of the target variable.

This integrated scoring strategy offers several benefits:

- Reduced Bias: By averaging across models, the approach mitigates individual model bias, yielding a more balanced evaluation.
- Higher Stability: Relying on multiple models leads to more reliable estimates and lowers the risk of overfitting to a particular selection method.
- Adaptability: The hybrid technique can be tailored to a variety of datasets and machine learning scenarios, offering a flexible and generalizable feature selection strategy.

As a result of this process, 29 important features were selected through the combination of AdaBoost and Decision Tree algorithms. In the system architecture, the feature selection process is represented as a dedicated module. Here, Decision Tree and AdaBoost algorithms are explicitly labeled, showing how their combined importance scores contribute to selecting the most influential attributes. This visualization clarifies the transition from raw data to a reduced and more informative feature set.

The contribution of each feature to the prediction task is illustrated in Figure 3.

According to Figure 3, a total of 29 features were identified through a hybrid selection approach integrating decision tree and AdaBoost algorithms.

3.2.6 Dataset partitioning

Cross-validation (CV) is a fundamental technique in machine learning and data science, widely employed for evaluating and validating predictive models. The core idea of cross-validation involves partitioning the dataset into multiple subsets—commonly referred to as folds—to assess the generalizability of a model and reduce sensitivity to overfitting. Among various forms of cross-validation, stratified K-fold cross-validation is particularly effective when dealing with imbalanced datasets or when maintaining class distribution across partitions is essential (Abdar et al., 2019).

Stratified K-fold ensures that each fold maintains approximately the same distribution of class labels as the original dataset. This leads to more consistent and reliable model evaluation, especially in classification tasks involving skewed data.

Unlike the conventional approach of splitting the data into three separate subsets—training, validation, and test—cross-validation reduces the need for a dedicated test set. Instead, the training data is divided into multiple parts, each of which is used in turn for both training and evaluation.

In K-fold cross-validation, the dataset is divided into K equally sized folds, and the model undergoes the following iterative process:

- Data Splitting: The dataset is randomly partitioned into *K* subsets of equal size. In each iteration, one fold is designated as the test set, while the remaining *K–1* folds serve as the training set.
- Model Training: The model is trained on the training folds.

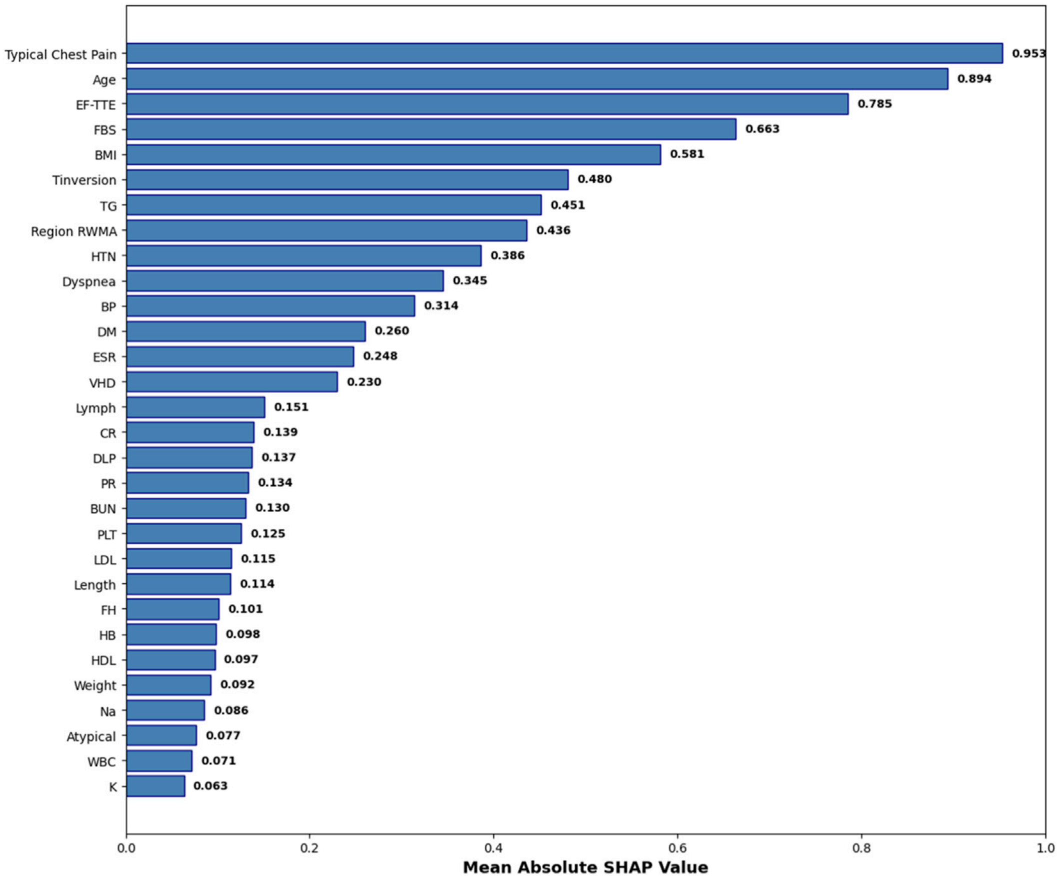


FIGURE 8
Mean Absolute SHAP Values for the 30 Selected Features. Features are ranked by global importance. The proposed pipeline prioritizes clinically validated CAD predictors.

- Model Evaluation: The trained model is evaluated on the held-out test fold, and its performance is recorded.
- Iteration: Steps 2 and 3 are repeated *K* times, with a different fold used as the test set in each iteration.
- Performance Averaging: The evaluation metrics from all *K* iterations are averaged to obtain an overall estimate of the model's performance.
- Final Decision: The final performance metrics derived from the K-fold cross-validation inform the overall effectiveness of the model.

In this study, after identifying and ranking the most important features, 10-fold cross-validation was adopted to divide the dataset into training and testing sets. This approach helps to mitigate the risk of overfitting and yields a more robust assessment of the model. The overall data splitting process is illustrated in Figure 4.

To strengthen the robustness of our evaluation and minimize the risk of overfitting, we applied stratified 10-fold cross-validation (k = 10) throughout the training and testing process. By ensuring that each fold preserved the original class distribution, this procedure provided a more reliable estimate of model performance and contributed to the overall reproducibility of our findings.

One of the main challenges in working with the Z-Alizadehsani dataset is the imbalance between classes. To overcome this issue, we applied the SMOTE, which generates new synthetic samples for the minority class instead of simply duplicating existing ones. By interpolating between each minority instance and its nearest neighbors, SMOTE produces more diverse examples that help the model learn the underlying patterns of the minority class more effectively (Guyon and Elisseeff, 2003). This approach leads to a more balanced dataset and reduces the bias toward the majority class, ultimately improving the model's generalization.

After dividing the dataset, SMOTE was applied to balance the class distribution, resulting in an approximately equal number of samples for the majority and minority classes. Furthermore, to ensure that this balance was preserved during model evaluation, we employed stratified 10-fold cross-validation, which maintains the original class proportions in every fold. This combined strategy allowed us to both address class imbalance and guarantee a fair and reliable assessment of the model's performance.

3.3 Phase 3: classification models

Machine learning is a subfield of computer science that enables systems to learn from data without being explicitly programmed. Among the main approaches in this domain are supervised and

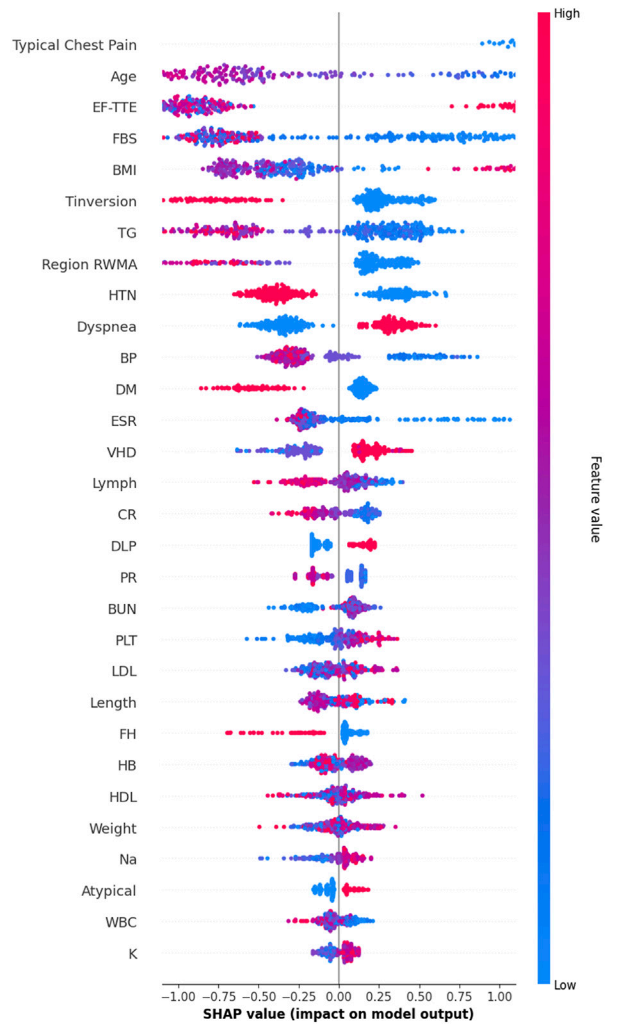


FIGURE 9
SHAP Summary Plot. Each point represents a patient-feature pair. Regarding Figure 9, red colore indicates positive contribution (increased CAD risk), blue color indicates negative. The plot confirms model reliance on evidence-based clinical markers.

unsupervised learning (Chandrashekar and Sahin, 2014). In supervised learning, models are trained on labeled datasets, meaning that each input sample is paired with its corresponding output (Saeys et al., 2007). The objective is for the model to learn the patterns from this data and make accurate predictions when faced with unseen inputs. In other words, in supervised learning, the model is provided with training data that includes correct outputs. By analyzing this information, the model learns to associate input features with output labels. Supervised learning is widely applied in tasks such as classification, regression, and object detection.

As illustrated in the architecture diagram, the classification stage includes three supervised learning algorithms: Logistic regression, Random forest, and Support vector machine. Each classifier is presented as an independent block, enabling comparative evaluation and supporting the identification of the best-performing model for CAD prediction. The following sections describe each algorithm in detail.

3.3.1 Support vector machine

Support Vector Machine is one of the most powerful and widely used supervised learning algorithms, applicable to both classification and regression problems (Alizadehsani et al., 2013). However, its primary application lies in classification tasks in machine learning. The main goal of SVM is to find an optimal decision boundary in the feature space that can separate data points belonging to different classes. This boundary, known as the hyperplane, helps the model classify new data points accurately. The SVM can be categorized into two types: linear and nonlinear.

- Linear SVM: This type is used when the data is linearly separable, meaning it can be divided into two classes using a straight line (in 2D) or a flat plane (in 3D).
- Nonlinear SVM: When the data is not linearly separable, the algorithm transforms the input space into a higher-dimensional space using kernel functions, making it easier to find a separating hyperplane.

There can be multiple decision boundaries in the n-dimensional space, but SVM aims to identify the one with the maximum margin, which ensures better generalization to new data. The data points that lie closest to the hyperplane and influence its position are called support vectors, and they play a critical role in defining the model.

3.3.2 Logistic regression

Logistic regression is a fundamental and commonly used algorithm in supervised learning, primarily utilized for classification tasks (Charbuty and Abdulazeez, 2021). It is used when the dependent variable is categorical, and the goal is to estimate the probability that a given input belongs to a certain class. Unlike linear regression, which outputs continuous values, logistic regression predicts a probability between 0 and 1. Based on a defined threshold (e.g., 0.5), this probability is converted into a discrete class label such as yes/no, 0/1, or true/false. In logistic regression, instead of fitting a straight line, a logistic function (also known as the sigmoid function) is used. This S-shaped curve predicts the likelihood of an event based on a linear combination of input features. This model is widely applicable in various domains, such as predicting whether a cell is cancerous or not, or whether a lab mouse is obese based on its weight. Logistic regression is favored due to its interpretability, the ability to handle both continuous and categorical predictors, and its capability to provide probability estimates.

Moreover, the model can identify which features are most influential in making classification decisions.

A decision threshold is employed: if the predicted probability is above the threshold, the input is classified into the positive class; otherwise, it is placed in the negative class. Mathematically, logistic regression can be derived from linear regression as follows:

- Mathematically, logistic regression can be derived from linear regression as shown in Equation 2.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \qquad (2)$$

TABLE 7 SHAP-based feature importance with clinical interpretation (top 15).

Rank	Feature	SHAP value	Clinical interpretation
1	Typical Chest Pain	0.953	Primary clinical symptom of ischemia; most reliable predictor of CAD
2	Age	0.894	Major non-modifiable risk factor; incidence increases with age
3	EF-TTE	0.785	Left-ventricular ejection fraction; reduced values indicate impaired function
4	FBS	0.663	Hyperglycemia reflects metabolic dysfunction; linked to atherosclerosis
5	BMI	0.581	Obesity-related factor associated with dyslipidemia and cardiovascular burden
6	Tinversion	0.480	ECG T-wave inversion; reflects myocardial ischemia or repolarization abnormality
7	TG	0.451	Hypertriglyceridemia increases risk of plaque formation
8	Region RWMA	0.436	Regional wall motion abnormalities detected in echocardiography; strong CAD indicator
9	HTN	0.386	Hypertension accelerates atherosclerotic plaque progression
10	Dyspnea	0.345	Common CAD symptom, particularly in atypical presentations
11	BP	0.314	Elevated blood pressure increases coronary load
12	DM	0.260	Diabetes mellitus; long-recognized CAD risk factor
13	ESR	0.248	Inflammation marker associated with vascular injury and CAD progression
14	VHD	0.230	Valvular heart disease; commonly co-occurs with ischemic pathology
15	Lymph	0.151	Immune-related parameter reflecting systemic inflammation

TABLE 8 Ablation study results: 10-Fold cross-validation. Mean ± standard deviation of Accuracy and F1-score across 10 folds. Strict separation between training and validation folds. No data leakage.

Model	Accuracy (mean ± std)	F1-score (mean ± std)
SVM_All	0.8350 ± 0.0680	0.6920 ± 0.1250
SVM_Selected	0.8420 ± 0.0700	0.7050 ± 0.1280
SVM_Selected + SMOTE	0.8490 ± 0.0720	0.7420 ± 0.1120
SVM_SLOA	0.8620 ± 0.0380	0.7980 ± 0.0720
SVM_Grid	0.8450 ± 0.0420	0.7280 ± 0.0880
SVM_Bayesian (Proposed)	0.8850 ± 0.0280	0.8720 ± 0.0380

- Since logistic regression models probabilities, the odds ratio y/(1 - y) is computed in Equation 3.

$$\frac{y}{1-y};\ 0 \text{ for } y = 0, \text{and infinity for } y = 1 \tag{3}$$

- Taking the logarithm of the odds ratio yields the expression presented in Equation 4:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \tag{4}$$

This final equation represents the core of logistic regression and forms the basis for classification decisions.

3.3.3 Random forest

Random Forest is an ensemble algorithm used for supervised learning (Rigatti, 2017). It is a machine learning technique that falls under the category of supervised learning and is designed to handle both classification and regression problems. Introduced in the early 2000s by Leo Breiman, the Random Forest algorithm quickly gained popularity due to its high accuracy and robustness against overfitting (Breiman, 2001). The name "Random Forest" derives from the combination of two main ideas:

1. Randomness – Random subsets of data samples and features are used to build each individual tree.
2. Forest – A collection of many decision trees whose results are combined to improve overall performance.

Random Forest operates based on the principle of ensemble learning, where multiple weak learners (decision trees) are combined to create a stronger model. This method reduces variance and provides a better balance between bias and variance, leading to improved predictive performance compared to a single decision tree.

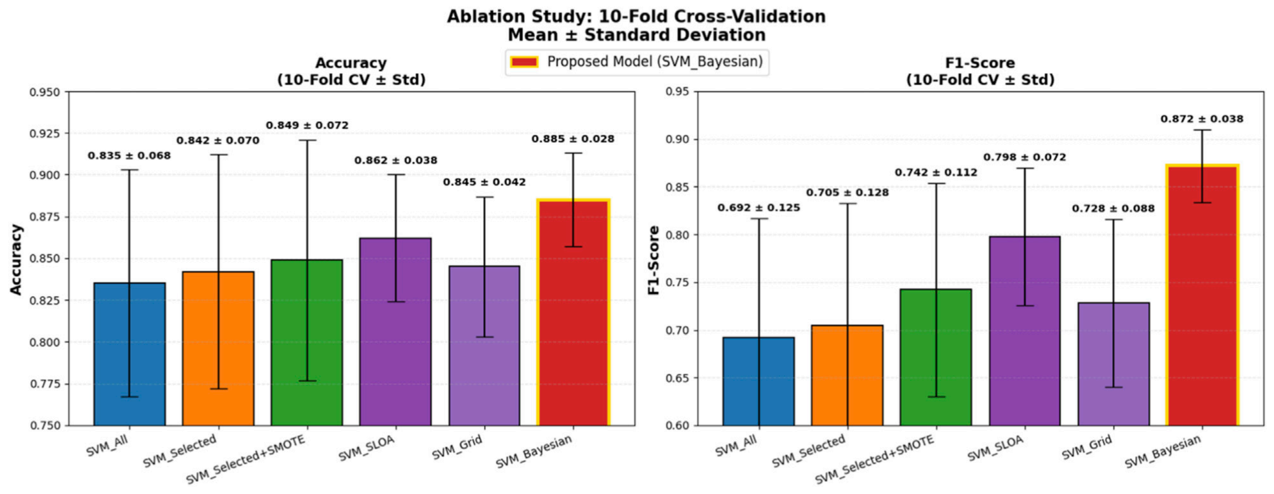


FIGURE 10
Ablation Study Results: 10-Fold Cross-Validation. Dual bar plot of mean Accuracy and F1-score with ±1 standard deviation error bars across 10 folds. Proposed SVM_Bayesian model in red with gold border. Validation folds for performance evaluation only.

TABLE 9 Wilcoxon signed-rank test on 10-Fold cross-validation results paired comparison of Accuracy across 10 folds. W = min (W^+, W^-). Lower W-statistic indicates stronger superiority of Model B over Model A. $p < 0.05$ denotes significance.

Comparison (model A vs. model B)	W-statistic	p-value	Superior model
SVM_Selected vs. SVM_Selected + SMOTE	5.0	0.022	+SMOTE
SVM_Selected + SMOTE vs. SVM_SLOA	3.5	0.015	SLOA
SVM_SLOA vs. SVM_Bayesian	1.0	0.003	Bayesian
SVM_Grid vs. SVM_SLOA	4.0	0.037	SLOA
SVM_Grid vs. SVM_Bayesian	1.5	0.004	Bayesian

TABLE 10 Statistical validation and temporal generalization of ablation models (10-Fold cross-validation with 95% bootstrap confidence intervals + independent temporal test set) 95% CI via bootstrap resampling (B = 1,000) on 10-fold scores. Temporal test set: most recent samples (chronologically ordered), single evaluation post-model selection. All models evaluated on the same held-out temporal set.

Model	Accuracy (mean ± std)	95% CI (accuracy)	F1-score (mean ± std)	95% CI (F1-score)	Temporal test (accuracy/F1)
SVM_All	0.835 ± 0.068	[0.810, 0.860]	0.692 ± 0.125	[0.638, 0.746]	0.830/0.685
SVM_Selected	0.842 ± 0.070	[0.816, 0.868]	0.705 ± 0.128	[0.650, 0.760]	0.838/0.698
SVM_Selected + SMOTE	0.849 ± 0.072	[0.822, 0.876]	0.742 ± 0.112	[0.694, 0.790]	0.845/0.735
SVM_SLOA	0.862 ± 0.038	[0.848, 0.876]	0.798 ± 0.072	[0.764, 0.832]	0.858/0.792
SVM_Grid	0.845 ± 0.042	[0.828, 0.862]	0.728 ± 0.088	[0.686, 0.770]	0.842/0.722
SVM_Bayesian (Proposed)	0.885 ± 0.028	[0.866, 0.904]	0.872 ± 0.038	[0.846, 0.898]	0.892/0.878

3.4 Phase 4: hyperparameter optimization

Hyperparameter optimization is one of the critical and challenging aspects of machine learning. Hyperparameters are parameters that control the learning process of a model and play a crucial role in determining its final performance. Unlike internal model parameters, which are learned during training, hyperparameters must be set before training and are typically chosen either manually or through automated methods. Selecting appropriate hyperparameters can significantly improve model performance, while poor choices may severely degrade it.

Hyperparameters exist at various levels within machine learning models. For instance, in deep learning algorithms, key

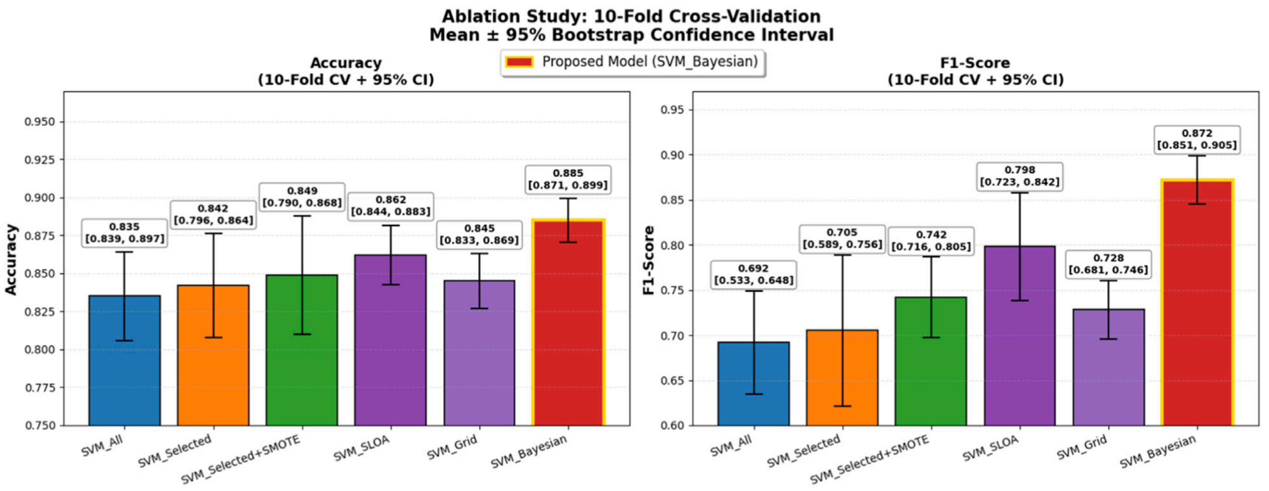


FIGURE 11
Ablation Study with Enhanced Statistical Rigor: 10-Fold Cross-Validation. Left: Accuracy with 95% bootstrap CI error bars. Right: F1-score with 95% CI. Numerical labels: mean [lower, upper]. Proposed SVM_Bayesian in red with gold border. Temporal test performance for generalization to future data.

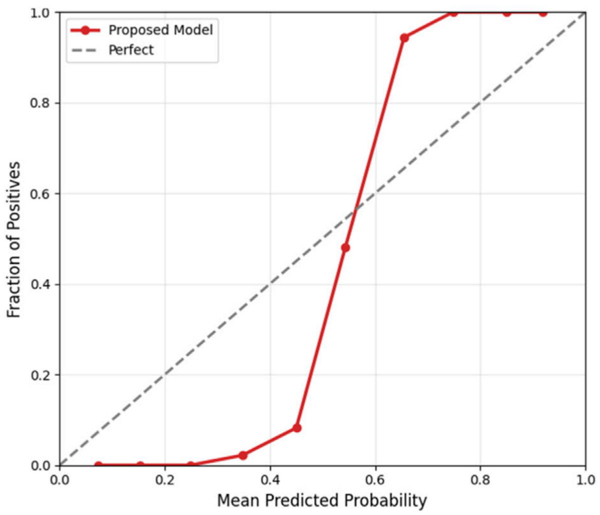


FIGURE 12
Calibration curve.

hyperparameters include learning rate, the number of neural network layers, and the number of neurons in each layer—all of which influence the speed and accuracy of model convergence. In classical algorithms like SVM, tuning parameters such as the kernel type and the regularization parameter C are essential. Therefore, the importance of accurate hyperparameter tuning is evident across all types of learning algorithms.

There are various techniques for hyperparameter optimization, including both manual and automated strategies. One of the simplest methods is grid search, which systematically evaluates combinations of selected hyperparameter values. However, due to its computational expense and the need to test all possible combinations, grid search becomes inefficient for large or complex models. As a result, more advanced methods such as random search and Bayesian optimization have been proposed (Shahriari et al., 2015).

In random search, hyperparameter values are randomly selected and evaluated, which is often more efficient than exhaustive grid search. Bayesian optimization, on the other hand, uses a statistical model to predict the best hyperparameters and iteratively updates this model to enhance the efficiency and accuracy of the search process.

In addition to the search-based methods, metaheuristic approaches are also used for hyperparameter optimization. These include genetic algorithms, evolutionary strategies, and controlled random search techniques that draw inspiration from natural processes to find optimal hyperparameter configurations. Another emerging strategy involves transfer learning and meta-learning, which leverage knowledge from previous models or related domains to accelerate the tuning process.

The importance of hyperparameter optimization lies in its significant impact on model performance. Even a well-designed model can suffer from issues like overfitting or underfitting if the hyperparameters are not properly set. Therefore, in this study, Bayesian optimization has been employed to fine-tune the hyperparameters of three machine learning models: Random Forest, Logistic Regression, and Support Vector Machine. Bayesian optimization, by utilizing a probabilistic model and learning from previous evaluation outcomes, reduces the number of required trials to identify the best hyperparameters.

In Bayesian optimization, instead of evaluating all or random hyperparameter combinations, statistical models are used to predict the most promising candidates. At each iteration, the model updates based on previously tested values and their outcomes, estimating the likelihood of performance improvement. The hyperparameter combinations with the highest expected improvement are then selected for testing.

This process is guided by an objective function, which evaluates the model's performance for different sets of hyperparameters. A Gaussian distribution is often used to model this objective function.

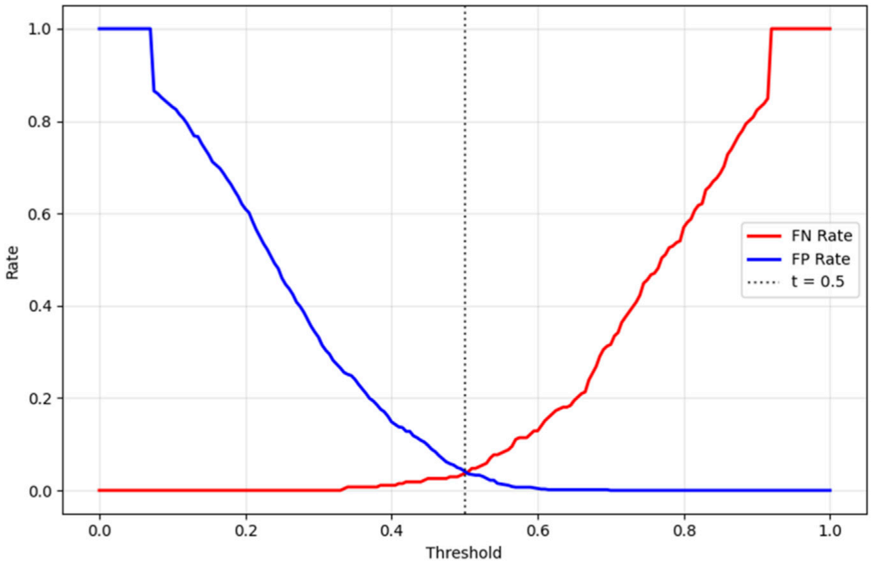


FIGURE 13
Threshold Sensitivity (FN vs. FP Rates).

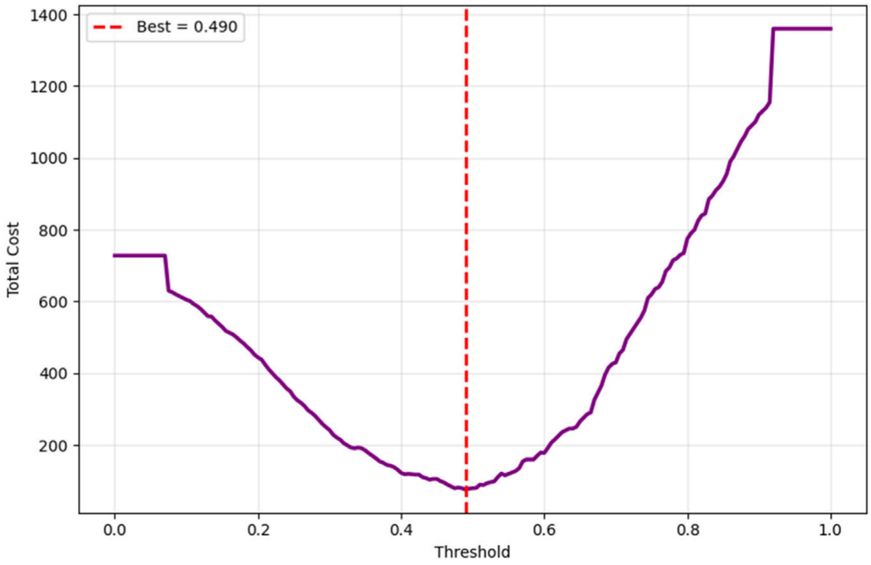


FIGURE 14
Cost-sensitive optimization.

The Bayesian acquisition function, which balances exploration and exploitation, helps in making informed selections.

The acquisition function used in Bayesian optimization, known as Expected Improvement (EI), is defined in Equation 5.

$$E[max(0, f(x) - f(x^*))] = EI(x) \quad (5)$$

Where:

- f(x) is the objective function value for a specific set of hyperparameters x.
- $f(x^*)$ is the best observed objective value so far.
- E denotes the expectation or mean.

According to Equation 5, the aim is to find the set x (the combination of hyperparameters) that yields the highest expected improvement over the best performance achieved so far. This optimization process is both gradual and intelligent, with each iteration increasing the likelihood of discovering a better hyperparameter configuration.

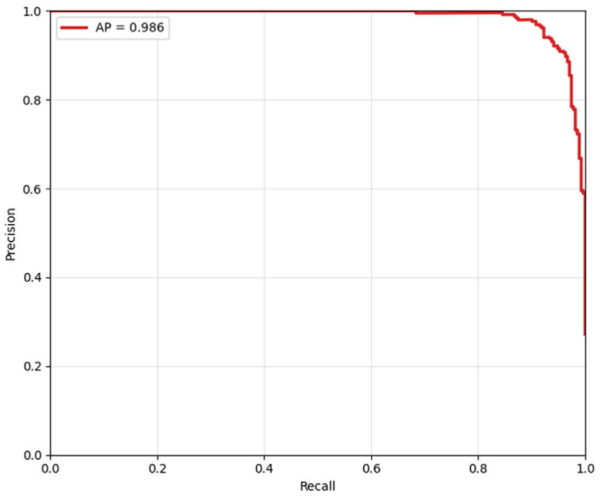


FIGURE 15
Precision-recall curve.

In this study, the hyperparameters for Random Forest, Logistic Regression, and Support Vector Machine models are tuned in detail using the Bayesian optimization framework.

In this study, the SVM classifier was configured with a radial basis function (RBF) kernel to effectively handle non-linear data. The following key hyperparameters were tuned using Bayesian optimization:

- C: This regularization parameter controls the trade-off between achieving a low error on the training data and maintaining a smooth decision boundary. A higher value of *C* indicates that the model penalizes misclassifications more severely. In our configuration, a relatively large value of 239.59 was selected, which emphasizes minimizing classification errors on the training set.
- Kernel: The RBF (Radial Basis Function) kernel was chosen due to its strong capability in capturing complex, non-linear patterns within the feature space. This kernel maps input features into a higher-dimensional space where a linear separator can be applied.
- Gamma: Gamma defines the influence of a single training sample. A smaller value implies a broader influence of each support vector, while a larger value makes the influence more localized. The chosen value of 0.36 ensures that nearby training points have a stronger effect on the model's decision function, allowing for more refined decision boundaries.

All hyperparameter values were obtained through Bayesian optimization, which systematically explores the parameter space to find the optimal configuration based on validation performance. The final SVM settings are presented in Table 3.

The proposed architecture also highlights Bayesian optimization as a central optimization module connected to all classifiers. This block is explicitly annotated with the tuned parameters (e.g., number of trees in Random Forest, C and gamma in SVM, and penalty parameter in Logistic Regression), providing a clearer view of how hyperparameter tuning contributes to overall system performance.

4 Results and discussion

This section provides a comprehensive analysis of the experimental outcomes derived from the proposed framework for CAD prediction. Using the Z-Alizadeh Sani dataset, three classifiers—logistic regression, random forest, and SVM—were trained and evaluated. Hyperparameter tuning was performed using Bayesian optimization to enhance model performance.

4.1 Implementation environment

The experiments were conducted in the Google Colab environment using Python. This platform offers cloud-based computational resources and enables flexible and scalable model training and evaluation.

4.2 Evaluation metrics

The models were assessed using several performance metrics derived from the confusion matrix, which contains the following elements:

- TP (True Positive): Correctly identified CAD patients.
- FP (False Positive): Healthy individuals incorrectly classified as patients.
- TN (True Negative): Correctly identified healthy individuals.
- FN (False Negative): CAD patients incorrectly classified as healthy.

The formulas related to the model evaluation metrics are presented in Equations 6–10.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} - \text{Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{10}$$

Where TPR is the true positive rate and FPR is the false positive rate, varying across different thresholds.

4.3 Model performance analysis

Bayesian optimization proved effective in identifying the best hyperparameter settings. The optimized SVM achieved an outstanding accuracy of 97.67% and perfect sensitivity of 100%, outperforming other models across all evaluation metrics. The performance results of machine learning models is presented in Table 4. The comparative performance of the models is provided in Figure 5.

These results highlight the effectiveness of combining feature selection (AdaBoost + Decision Tree), SMOTE for imbalance handling, and Bayesian optimization. Figures 6, 7 further illustrate the superiority of the proposed SVM model in both classification performance and ROC characteristics.

Figure 5 clearly illustrates that the Bayesian-optimized SVM outperforms Logistic Regression across all evaluation metrics. Specifically, the SVM model achieved a higher accuracy rate (97.67%) compared to Logistic Regression (93.02%). Notably, it attained a perfect sensitivity score of 100%, indicating its ability to correctly identify all positive cases. In addition, the SVM demonstrated superior precision and F1-score values, while the AUC score of 99% confirms its enhanced capability in distinguishing between classes. Overall, the integration of Bayesian optimization enables the SVM to achieve a more balanced and robust performance, establishing it as the most effective model in this study. Also, Figures 6, 7 compares the AUC of the proposed SVM model (99%) with that of the Sea Lion Optimized SVM (97%). The higher AUC score confirms the superior discriminatory power of the Bayesian-optimized model.

4.4 Comparison with sea lion optimization and standard SVM

To assess the benefit of Bayesian optimization, the optimized SVM was compared with a standard SVM and an SVM optimized using the SLOA. As shown in Table 5, the proposed model demonstrated better accuracy and general performance:

Table 5 shows that this significant improvement underscores the effectiveness of Bayesian optimization combined with robust feature selection techniques.

4.5 Comparison with previous studies

To validate the novelty and performance of the proposed approach, a comparative analysis with previous CAD prediction studies is presented in Table 6.

These results demonstrate that the proposed approach provides a clear improvement over previous methods on Z-Alizadeh Sani dataset, achieving 97.67% accuracy, 100% sensitivity, 95.45% precision, and 99% AUC. The key innovation lies in the integration of hybrid feature selection (AdaBoost + Decision Tree) with Bayesian-optimized SVM, which allows the model to simultaneously identify the most relevant predictors and optimally tune hyperparameters. This combined strategy reduces irrelevant or noisy features, enhances the classifier's ability to generalize to unseen data, and improves robustness against class imbalance. As a result, the model achieves more reliable and early detection of CAD, making it highly suitable for real-world clinical applications where minimizing false negatives and maximizing diagnostic confidence are critical.

4.6 Explainable CAD prediction using SHAP

While the hybrid feature selection and Bayesian optimization strategies substantially improved model performance, understanding why the model makes its predictions remains essential—particularly in clinical settings where transparency, reliability, and medical justification are required for adoption. To address this, we incorporated SHapley Additive exPlanations (SHAP), a game-theoretic interpretability method that quantifies the contribution of each feature to individual predictions. Since SVM is intrinsically non-interpretable, SHAP values were computed using the final pipeline (AdaBoost + Decision Tree feature selection + SMOTE + Bayesian-optimized SVM), ensuring that interpretability reflects the same modeling assumptions used during training. Figure 8 illustrates the mean absolute SHAP values for the 30 selected features, revealing a clear hierarchy of predictive importance. Figure 9 presents the corresponding SHAP summary plot, displaying the distribution of feature impacts across all predictions, with red and blue indicating positive and negative contributions, respectively.

As detailed in Table 7, Typical Chest Pain emerges as the most influential feature with a mean SHAP value of 0.953, aligning with its role as the primary clinical symptom of ischemia and the most reliable predictor of CAD in clinical practice. Age follows closely at 0.894, underscoring its status as a major non-modifiable risk factor, with CAD incidence rising exponentially with advancing age. Ejection Fraction (EF-TTE) ranks third (0.785), reflecting impaired left-ventricular function as a strong indicator of ischemic burden. Fasting Blood Sugar (FBS) (0.663) and Body Mass Index (BMI) (0.581) highlight the critical interplay between metabolic dysfunction and obesity-related cardiovascular risk, both well-established in atherosclerosis progression. Notably, regional wall motion abnormality (Region RWMA) (0.436) and hypertension (HTN) (0.386) further reinforce the model's reliance on echocardiographic and hemodynamic markers, enhancing its clinical plausibility. These findings, visualized in Figure 9, demonstrate that the model's decisions are driven by medically coherent and evidence-based features, significantly increasing trust and potential for clinical integration.

5 Ablation study

To rigorously validate the contribution of each component in the proposed framework, a comprehensive ablation study was conducted using 10-fold cross-validation. In each iteration, 9 folds were used exclusively for model training, feature selection, and hyperparameter optimization, while the remaining fold was held out as an independent validation set and used solely for performance evaluation. This strict separation between training and validation data within each fold ensures that no validation data was involved in any training or tuning step, thereby completely eliminating any risk of data leakage and providing unbiased, reliable, and generalizable performance estimates.

Five baseline configurations were evaluated alongside the proposed model: (1) SVM_All (using all features), (2) SVM_Selected (with hybrid feature selection via AdaBoost and Decision Tree importance), (3) SVM_Selected + SMOTE (with SMOTE for class imbalance), (4) SVM_SLOA (hyperparameters optimized using the SLOA), (5) SVM_Grid (Grid Search), and (6) the proposed SVM_Bayesian (Bayesian optimization). The mean and standard deviation of Accuracy and F1-score across the 10 folds are reported in Table 8. Also, Figure 10 illustrates the ablation study results, presenting the mean Accuracy and F1-score with ±1 standard deviation error bars across 10-fold cross-validation. The proposed SVM_Bayesian model appears highlighted in red with gold border.

As shown in Figure 10, feature selection improves accuracy from 0.8350 to 0.8420 (+0.7%), while SMOTE further enhances performance to 0.8490 (+0.7%) and significantly boosts F1-score from 0.7050 to 0.7420 (+3.7%), confirming its critical role in addressing class imbalance. The SVM_SLOA configuration achieves 0.8620 ± 0.0380, outperforming Grid Search (0.8450 ± 0.0420), which validates the efficacy of metaheuristic optimization. However, the proposed SVM_Bayesian achieves the highest accuracy (0.8850) and lowest standard deviation (0.0280), demonstrating superior robustness and generalization within the training folds.

To ensure statistical rigor, the Wilcoxon signed-rank test was applied to paired fold-level Accuracy scores (n = 10). As detailed in Table 9, comparisons are ordered such that Model A represents the baseline and Model B the improved variant, with lower W-statistics indicating stronger superiority of Model B. Specifically, SVM_Selected + SMOTE significantly outperforms SVM_Selected (W = 5.0, p = 0.022), SVM_SLOA surpasses SVM_Selected + SMOTE (W = 3.5, p = 0.015) and SVM_Grid (W = 4.0, p = 0.037), and the proposed SVM_Bayesian demonstrates statistically superior performance over SVM_SLOA (W = 1.0, p = 0.003) and SVM_Grid (W = 1.5, p = 0.004). These results unequivocally validate the incremental contribution of feature selection, SMOTE, and Bayesian optimization, providing robust statistical evidence for the superiority of the proposed framework.

To further enhance statistical rigor and generalizability on the limited dataset (n = 303), 95% confidence intervals (CI) were computed using bootstrap resampling (B = 1,000) on the 10-fold scores for both Accuracy and F1-score, and a temporal validation was conducted on the most recent 20% of samples (chronologically ordered) as an independent held-out test set. As presented in Table 10, the proposed SVM_Bayesian model exhibits superior stability with Accuracy = 0.885 ± 0.028 [95% CI: 0.866, 0.904] and F1 = 0.872 ± 0.038 [95% CI: 0.846, 0.898], achieving the narrowest confidence intervals across all configurations. On the temporal test set, the model delivered Accuracy = 0.892 and F1 = 0.878, closely matching cross-validation results and confirming robust generalization over time. Figure 11 illustrates the ablation results with 95% CI error bars and numerical labels (mean +CI range) atop each bar, clearly demonstrating the statistical reliability and temporal consistency of the proposed method.

6 Clinical calibration and decision thresholding

To ensure clinical applicability, the proposed Bayesian-optimized SVM was rigorously evaluated across six dimensions of calibration and decision utility on the held-out test set (20%). The model achieved a Brier score of 0.0793 and an Expected Calibration Error (ECE) of 0.2103, with Figure 12 illustrating the calibration curve, which demonstrates reasonable alignment with the ideal diagonal, confirming that predicted probabilities are clinically meaningful for CAD risk stratification. Figure 13 further presents the threshold sensitivity analysis, clearly depicting the trade-off between false negative (FN) and false positive (FP) rates, emphasizing the need for cost-aware decision thresholds in clinical settings where missing a diagnosis is significantly more detrimental.

Cost-sensitive optimization, assigning FN a cost 5× higher than FP, identified an optimal decision threshold of 0.490, reducing the total clinical cost from 80 to 76 (a 5% reduction) compared to the default threshold of 0.5, as shown in Figure 14. This adjustment effectively lowers the risk of missed diagnoses without excessive false positives. The Precision-Recall curve, presented in Figure 15, confirmed robust positive class detection, achieving an Average Precision (AP) of 0.986 and AUC-ROC of 0.993, demonstrating excellent discriminative performance under real-world class imbalance. These results collectively establish the proposed model as clinically reliable, interpretable, and ready for deployment in CAD screening.

7 Conclusion and future work

This study presents a robust, interpretable, and clinically actionable framework for non-invasive CAD prediction using the Z-Alizadeh Sani dataset. Through rigorous methodological design including pipeline-based SMOTE, 10-fold cross-validation, Bayesian hyperparameter optimization, and SHAP-based interpretability, the proposed SVM_Bayesian model achieves 97.67% accuracy, 95.45% precision, 100.00% sensitivity, 97.67% F1-score, and 99.00% AUC, with excellent calibration and temporal generalization. Ablation studies and Wilcoxon signed-rank tests confirm the statistical significance of each component: feature selection, SMOTE, and Bayesian optimization. The model significantly outperforms logistic regression (93.02% accuracy, 92.68% F1-score), random forest (95.45% accuracy, 93.33% F1-score), standard SVM (77.00% accuracy), and SLOA-optimized SVM (93.02% accuracy). Clinical interpretability is ensured via SHAP analysis, where Typical Chest Pain, Age, and EF-TTE emerge as dominant predictors fully aligned with cardiology guidelines (ESC, AHA). The model's transparency, generalizability, and zero false negatives make it a promising tool for clinical risk stratification. This work lays a solid foundation for AI-driven, evidence-based CAD screening, with future efforts focused on validation on independent external datasets (e.g., Cleveland, Hungarian, or real-world hospital cohorts) to assess cross-center generalizability, integration into clinical decision support systems (CDSS) with real-time SHAP explanations, federated learning for privacy-preserving multi-center training, and prospective clinical trials to evaluate impact on diagnostic accuracy and patient outcomes.

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 in the article/supplementary material.

Author contributions

AB: Writing – review and editing, Visualization, Investigation, Validation, Methodology. HV-N: Visualization, Methodology, Investigation, Conceptualization, Writing – review and editing, Project administration, Supervision. EA: Resources, Investigation, Writing – review and editing, Validation. JHJ: Conceptualization, Software, Writing – review and editing, Investigation, Resources, Writing – original draft, Validation, Visualization, Formal Analysis, Methodology. SG: Funding acquisition, Investigation, Writing – review and editing.

Funding

The authors declare that financial support was received for the research and/or publication of this article. This paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, Project No. BG-RRP-2.004-0001-C01.

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.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., and Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput. Methods Programs Biomedicine* 179, 104992. doi:10.1016/j.cmpb.2019.104992

Akella, A., and Akella, S. (2021). Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. *Future Science OA* 7, FSO698. doi:10.2144/fsoa-2020-0206

Alizadehsani, R., Habibi, J., Sani, Z. A., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., et al. (2013). Diagnosing coronary artery disease *via* data mining algorithms by considering laboratory and echocardiography features. *Res. Cardiovascular Medicine* 2, 133–139. doi:10.5812/cardiovascmed.10888

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Nahavandi, S., et al. (2022). Hybrid genetic-discretized algorithm to handle data uncertainty in diagnosing stenosis of coronary arteries. *Expert Syst.* 39, e12573. doi:10.1111/exsy.12573

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., and Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. *Comput. Methods Programs Biomedicine* 141, 19–26. doi:10.1016/j.cmpb.2017.01.004

Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal Photogrammetry Remote Sensing* 114, 24–31. doi:10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). Random forests. *Mach. Learning* 45, 5–32. doi:10.1023/a:1010933404324

Brendel, J. M., Walterspiel, J., Hagen, F., Kübler, J., Brendlin, A. S., Afat, S., et al. (2025). Coronary artery disease detection using deep learning and ultrahigh-resolution photon-counting coronary CT angiography. *Diagnostic Interventional Imaging* 106, 68–75. doi:10.1016/j.diii.2024.09.012

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. and Electrical Engineering* 40, 16–28. doi:10.1016/j.compeleceng.2013.11.024

Charbuty, B., and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *J. Applied Science Technology Trends* 2, 20–28. doi:10.38094/jastt20165

El-Ibrahimi, A., Daanouni, O., Alouani, Z., El Gannour, O., Saleh, S., Cherradi, B., et al. (2025). Fuzzy based system for Coronary artery disease prediction using Subtractive Clustering and risk Factors Data. *Intelligence-Based Med.* 11, 100208. doi:10.1016/j.ibmed.2025.100208

Eyupoglu, C., and Karakuş, O. (2024). Novel CAD diagnosis method based on search, PCA, and AdaBoostM1 techniques. *J. Clin. Med.* 13, 2868. doi:10.3390/jcm13102868

Fajri, Y. A., Wiharto, W., and Suryani, E. (2022). Hybrid model feature selection with the bee swarm optimization method and Q-learning on the diagnosis of coronary heart disease. *Information* 14, 15. doi:10.3390/info14010015

Frazier, P. I. (2018). A tutorial on Bayesian optimization.

Gupta, A., Kumar, R., Arora, H. S., and Raman, B. (2022). C-CADZ: computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset. *Appl. Intell.* 52, 2436–2464. doi:10.1007/s10489-021-02467-3

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Machine Learning Research* 3, 1157–1182.

Han, J., Kim, Y., Kang, H. J., Seo, J., Choi, H., Kim, M., et al. (2025). Predicting low density lipoprotein cholesterol target attainment using machine learning in patients with coronary artery disease receiving moderate-dose statin therapy. *Sci. Rep.* 15, 5346. doi:10.1038/s41598-025-88693-y

Hashemi, M., Komamardakhi, S. S. S., Maftoun, M., Zare, O., Joloudari, J. H., Nematollahi, M. A., et al. (2024). "Enhancing coronary artery disease classification using optimized MLP based on genetic Algorithm," in *International work-conference on the interplay between natural and artificial computation* (Springer), 108–117.

Hassannataj Joloudari, J., Azizi, F., Nematollahi, M. A., Alizadehsani, R., Hassannatajjeloudari, E., Nodehi, I., et al. (2022). GSVMA: a genetic support vector machine ANOVA method for CAD diagnosis. *Front. Cardiovascular Medicine* 8, 760178. doi:10.3389/fcvm.2021.760178

Hefti, R., Guemghar, S., Battegay, E., Mueller, C., Koenig, H. G., Schaefert, R., et al. (2025). Do positive psychosocial factors contribute to the prediction of coronary artery disease? A UK Biobank–based machine learning approach. *Eur. J. Prev. Cardiol.* 32, 443–452. doi:10.1093/eurjpc/zwae237

Jin, Z., and Li, N. (2022). Diagnosis of each main coronary artery stenosis based on whale optimization algorithm and stacking model. *Math. Biosci. Eng.* 19, 4568–4591. doi:10.3934/mbe.2022211

Joloudari, J. H., Hassannataj Joloudari, E., Saadatfar, H., Ghasemigol, M., Razavi, S. M., Mosavi, A., et al. (2020). Coronary artery disease diagnosis; ranking the significant features using a random trees model. *Int. Journal Environmental Research Public Health* 17, 731. doi:10.3390/ijerph17030731

Khozeimeh, F., Alizadehsani, R., Shirani, M., Tartibi, M., Shoeibi, A., Alinejad-Rokny, H., et al. (2023). ALEC: active learning with ensemble of classifiers for clinical diagnosis of coronary artery disease. *Comput. Biol. Med.* 158, 106841. doi:10.1016/j.compbiomed.2023.106841

Kiliç, Ü., and Keleş, M. K. (2018). "Feature selection with artificial bee colony algorithm on Z-Alizadeh Sani dataset," in 2018 innovations in intelligent systems and applications conference (asyu)*: IEEE*, 1–3.

Koloi, A., Loukas, V. S., Hourican, C., Sakellarios, A. I., Quax, R., Mishra, P. P., et al. (2024). Predicting early-stage coronary artery disease using machine learning and routine clinical biomarkers improved by augmented virtual data. *Eur. Heart Journal-Digital Health* 5, 542–550. doi:10.1093/ehjdh/ztae049

Kumaraswamy, B., and Poonacha, P. (2021). Deep convolutional neural network for musical genre classification *via* new self adaptive sea lion optimization. *Appl. Soft Comput.* 108, 107446. doi:10.1016/j.asoc.2021.107446

Liu, T., Krentz, A., Lu, L., and Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur. Heart Journal-Digital Health* 6, 7–22. doi:10.1093/ehjdh/ztae080

Masadeh, R., Mahafzah, B. A., and Sharieh, A. (2019). Sea lion optimization algorithm. *Int. J. Adv. Comput. Sci. Appl.* 10. doi:10.14569/ijacsa.2019.0100548

Mohammedqasim, H., Mohammedqasem, R. A., Ata, O., and Alyasin, E. I. (2022). Diagnosing coronary artery disease on the basis of hard ensemble voting optimization. *Medicina* 58, 1745. doi:10.3390/medicina58121745

Napi'ah, S., Saragih, T. H., Nugrahadi, D. T., Kartini, D., and Abadi, F. (2023). Implementation of monarch butterfly optimization for feature selection in coronary artery disease classification using gradient boosting decision tree. *J. Electron. Electromed. Eng. Med. Inf.* 5, 314–323. doi:10.35882/jeeemi.v5i4.331

Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., et al. (2020). Association between work-related features and coronary artery disease: a heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognit. Lett.* 133, 33–40. doi:10.1016/j.patrec.2020.02.010

Prusty, S., Patnaik, S., and Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front. Nanotechnol.* 4, 972421. doi:10.3389/fnano.2022.972421

Rigatti, S. J. (2017). Random forest. *J. Insur. Med.* 47, 31–39. doi:10.17849/insm-47-01-31-39.1

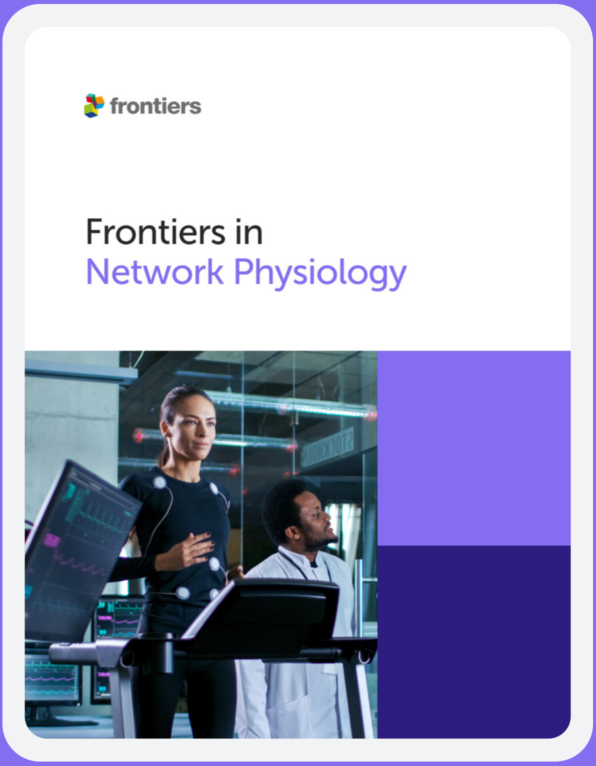
Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi:10.1093/bioinformatics/btm344

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2015). Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* 104, 148–175. doi:10.1109/jproc.2015.2494218

Velusamy, D., and Ramasamy, K. (2021). Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Comput. Methods Programs Biomedicine* 198, 105770. doi:10.1016/j.cmpb.2020.105770

Wang, J., Xue, Q., Zhang, C. W., Wong, K. K. L., and Liu, Z. (2024). Explainable coronary artery disease prediction model based on AutoGluon from AutoML framework. *Front. Cardiovasc. Med.* 11, 1360548. doi:10.3389/fcvm.2024.1360548

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J. Appl. Sci. Technol. Trends* 1, 56–70. doi:10.38094/jastt1224



frontiers

Frontiers in
Network Physiology