


Research on atrial fibrillation diagnosis in electrocardiograms based on CLA-AF model

Jiajia Si ¹, Yiliang Bao¹, Fengling Chen^{2,3}, Yue Wang¹, Meimei Zeng¹, Nongyue He¹, Zhu Chen^{1,3}, and Yuan Guo^{1,2,3,*}

¹Hunan Key Laboratory of Biomedical Nanomaterials and Devices, Hunan University of Technology, No. 88 West Taishan Road, Zhuzhou 412007, Hunan, China; ²Department of Cardiovascular Medicine, Zhuzhou Hospital Affiliated to Xiangya School of Medicine, Central South University, No. 116 South Changjiang Road, Zhuzhou 412007, Hunan, China; and

³Hengyang Medical School, University of South China, No. 28 West Changsheng Road, Hengyang 421001, Hunan, China

Received 6 July 2024; revised 4 September 2024; accepted 27 October 2024; online publish-ahead-of-print 27 November 2024

Aims

The electrocardiogram (ECG) is the primary method for diagnosing atrial fibrillation (AF), but interpreting ECGs can be time-consuming and labour-intensive, which deserves more exploration.

Methods and results

We collected ECG data from 6590 patients as YY2023, classified as Normal, AF, and Other. Convolutional Neural Network (CNN), bidirectional Long Short-Term Memory (BiLSTM), and Attention construct the AF recognition model CNN BiLSTM Attention-Atrial Fibrillation (CLA-AF). The generalization ability of the model is validated on public datasets CPSC2018, PhysioNet2017, and PTB-XL, and we explored the performance of oversampling, resampling, and hybrid datasets. Finally, additional PhysioNet2021 was added to validate the robustness and applicability in different clinical settings. We employed the SHapley Additive exPlanations (SHAP) method to interpret the model's predictions. The F1-score, Precision, and area under the ROC curve (AUC) of the CLA-AF model on YY2023 are 0.956, 0.970, and 1.00, respectively. Similarly, the AUC on CPSC2018, PhysioNet2017, and PTB-XL reached above 0.95, demonstrating its strong generalization ability. After oversampling PhysioNet2017, F1-score and Recall improved by 0.156 and 0.260. Generalization ability varied with sampling frequency. The model trained from the hybrid dataset has the most robust generalization ability, achieving an AUC of 0.96 or more. The AUC of PhysioNet2021 is 1.00, which proves the applicability of CLA-AF. The SHAP values visualization results demonstrate that the model's interpretation of AF aligns with the diagnostic criteria of AF.

Conclusion

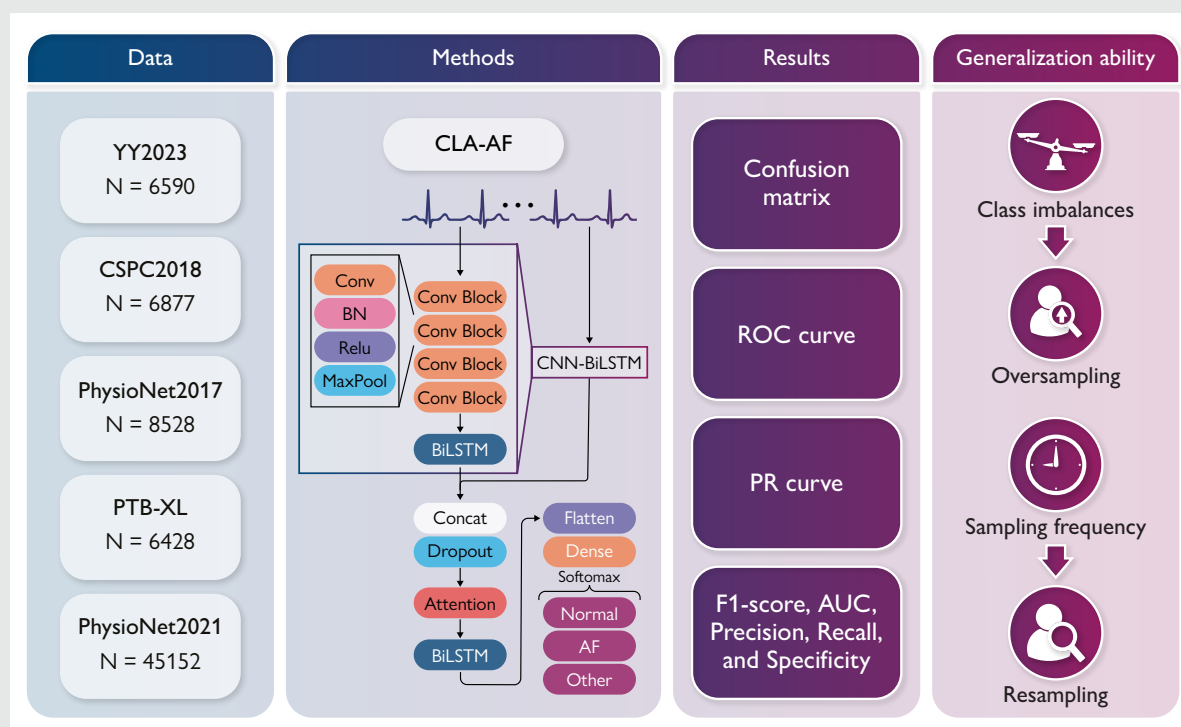
The CLA-AF model demonstrates a high accuracy in recognizing AF from ECG, exhibiting remarkable applicability and robustness in diverse clinical settings.

* Corresponding author. Tel: +86 15200828141, Email: guoyuan0815@163.com

© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Graphical Abstract



Keywords

Atrial fibrillation • Electrocardiograms • Convolutional Neural Network • Bidirectional Long Short-Term Memory • Attention

Introduction

Atrial fibrillation (AF) stands as the prevailing tachyarrhythmia in clinical settings, considerable harm, and suboptimal treatment efficacy.^{1,2} In clinical practice, the primary tool for physicians to diagnose AF is through electrocardiograms (ECGs). However, this process demands significant time and effort, proving less efficient. Hence, there is a compelling need to adopt more effective approaches, such as employing deep learning, for AF diagnosis, which alleviates the burden on health-care professionals and enhances diagnostic accuracy.^{3–5}

Recently, a growing number of automatic AF recognition algorithms have been applied in AF screening, mainly for machine learning and deep learning^{6–11} to recognize AF. Traditional machine learning algorithms focus on processing data, extracting features, and classification. It has been studied to extract features using wavelet analysis and classify them using support vector machine (SVM) and decision tree to achieve high recognition accuracy.¹² Deep learning is a sophisticated machine learning algorithm that solves more complex problems.^{13,14} Models such as Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), and Attention mechanisms are suitable for AF recognition.

The initial success of CNNs in processing ECG data provides a compelling avenue for advancing the scalability and accuracy of automated classification for ECG signals.¹⁵ This success stems from two key advantages: using raw ECG signals as input data for end-to-end deep neural networks without manual rule pre-processing and continuously expanding classification types based on appropriate training set data.

Recent findings indicate that employing a lightweight CNN or a 2D CNN for recognizing AF from ECG yields superior accuracy.^{16,17} Despite the breakthroughs achieved in accuracy by previous studies, challenges related to overfitting and generalization capabilities remain open questions in the current landscape of model research. Thus, further investigation is warranted to develop models exhibiting high accuracy and robust generalization capabilities.

Recurrent neural networks have been widely used in research related to sequential data and are considered one of the most effective models for time series prediction.¹⁸ However, conventional RNNs encounter challenges in handling long-term dependency issues. Long Short-Term Memory (LSTM) addresses these problems and enabling the capture of both long-term and short-term dependencies in sequence data.¹⁹ Long Short-Term Memory mainly deals with time series and applies to recognizing ECGs. Transformer can capture the global dependencies in the sequences through the self-attention mechanism to process the long sequence data efficiently.²⁰ In time series classification, CNN and RNN are often combined because they have complementary advantages that can improve the performance of the model.

Attention mechanisms can automatically learn and selectively focus on the crucial input information, which is pivotal in enhancing model performance and generalization.²¹ Automatic ECG classification using RA-UNET or non-local convolutional block attention modules has been found to have high recognition performance on different datasets.^{22,23} Many recent studies have focused on hybrid models, such as the self-attention-based LSTM-Fully Convolutional Network (LSTM-FCN) model for arrhythmia classification and the hybrid attention-based deep learning network.^{24,25}

To better retain the state of the entire signal before and after and improve the model's generalization, we use bidirectional Long Short-Term Memory (BiLSTM) to extract temporal features and introduce an attention mechanism.

Therefore, to achieve both high accuracy and robust generalization, we initially gathered ECG data and diagnostic information from 6590 patients, forming dataset YY2023. The AF diagnostic model, CLA-AF, was developed using a combination of CNN, BiLSTM, and Attention. The model underwent training and testing on the YY2023 dataset. Subsequently, its generalization performance was validated on open-access datasets, including CPSC2018, PhysioNet2017, and PTB-XL. Finally, additional dataset PhysioNet2021 was added to verify the robustness and applicability of the model in different clinical settings. Enhancements were iteratively applied based on the results to ensure improved generalizability.

Methods

Study design of electrocardiogram diagnosis of AF

Dataset construction and public dataset collection

The dataset YY2023, comprising 6590 ECGs, was sourced from Zhuzhou Hospital Affiliated to Xiangya School of Medicine, Central South University. Initial label files were generated for YY2023 by categorizing the ECGs into three classes based on physician diagnosis: normal ECGs, AF ECGs, and other abnormal ECGs.

Additionally, ECGs from public datasets CPSC2018, PhysioNet2017, PTB-XL, and PhysioNet2021 were similarly classified into these three categories. This study focuses on constructing a three-classification model for ECGs, with the primary objective of diagnosing AF. In the four datasets, YY2023, CPSC2018, PhysioNet2017, PTB-XL, and PhysioNet2021 are 12-lead ECGs with a sampling frequency of 500 Hz, and PhysioNet2017 is a single-lead ECG with a sampling frequency of 300 Hz, and the information of the datasets after re-creating the labels is shown in [Table 1](#) and [Supplementary material online, Table S1](#). YY2023 was used for model research and construction, CPSC2018, PhysioNet2017, and PTB-XL were used to verify the generalization ability of the model, and PhysioNet2021 was used to further verify the robustness and applicability of the model in different clinical settings.

Electrocardiogram data pre-processing

The raw information of YY2023 was stored in XML format and desensitized before use. Since the AF features are most significant on lead II, to strengthen the generalization ability of the constructed model, we only used lead II of 12-lead ECG and single-lead ECG for the study. As shown in [Figure 1](#), the ECG data are first extracted from the XML file and pre-processed: mean normalization, wavelet transform filtering to remove baseline drifts, etc.²⁶ The pre-processed data are stored in npy format. Since the length of ECGs can be different and the input size of the neural network is fixed, we split the ECGs into segments of length 1500 to improve the model's generalization ability.²⁷ Finally, the segmentation is fed into the model for training, testing, and evaluating the model performance.

CLA-AF model architecture for diagnosing AF

In this paper, we present a CLA-AF model (CNN BiLSTM Attention-Atrial Fibrillation) that integrates CNN, BiLSTM, and Attention mechanisms for AF recognition, as depicted in [Figure 2A](#). To determine which model architecture is optimal, multiple model structures were compared, as shown in [Supplementary material online, Table S2](#). Given the continuous nature of ECG segments within and between segments, bidirectional LSTM is employed for effective temporal feature extraction. The CNN block incorporates a one-dimensional convolutional layer for feature extraction, batch normalization to mitigate overfitting, rectified linear unit (ReLU) activation function to enable the network to generate a sophisticated nonlinear representation of the ECG, and MaxPooling (MaxPool) to compress features for manageable network optimization. Concatenation (Concat) operation combines all learned segments of the complete ECG, followed by weight redistribution using Attention. This step aims to identify correlations between the segments and emphasize specific crucial features, as illustrated in [Figure 2B](#). Ultimately, the ECG recognition results are generated through the Dense layer, employing the softmax activation function.

YY2023 is divided into a training set and a test set according to the ratio of 8:2, and the data were pre-processed. A total of 20% of the training set is used as the validation set, and the parameters of the model are adjusted according to the validation set during the training process. The training epoch is formed to 100, the epoch size is set to 64, the optimization process using SGD optimizer, the initial value of the learning rate is set to 0.1, and the minimum value is set to 0.001. The learning rate is 0.1 when epoch ≤ 20 , 0.01 when $20 < \text{epoch} \leq 60$, and 0.001 when epoch > 60 , and the learning rate is updated according to this rule. The model is trained according to the set parameters, then the model is tested using the split test set, and the CLA-AF model performance is analysed based on the model evaluation metrics. The cross-entropy loss function `categorical_crossentropy` is used as a metric to measure the variance of the classification task.

$$\text{loss} = - \sum_{i=1}^3 y_i \log \hat{y}_i.$$

Then, the trained model is tested using the test set after obtaining the trained model. CPSC2018, PhysioNet2017, PTB-XL, and PhysioNet2021 are also split into training and testing sets in an 8:2 ratio to validate the model's generalization performance. It is also considered whether training the model using the composite dataset obtained by integrating the four datasets will give better performance. The overall flow of research is shown in [Figure 1](#).

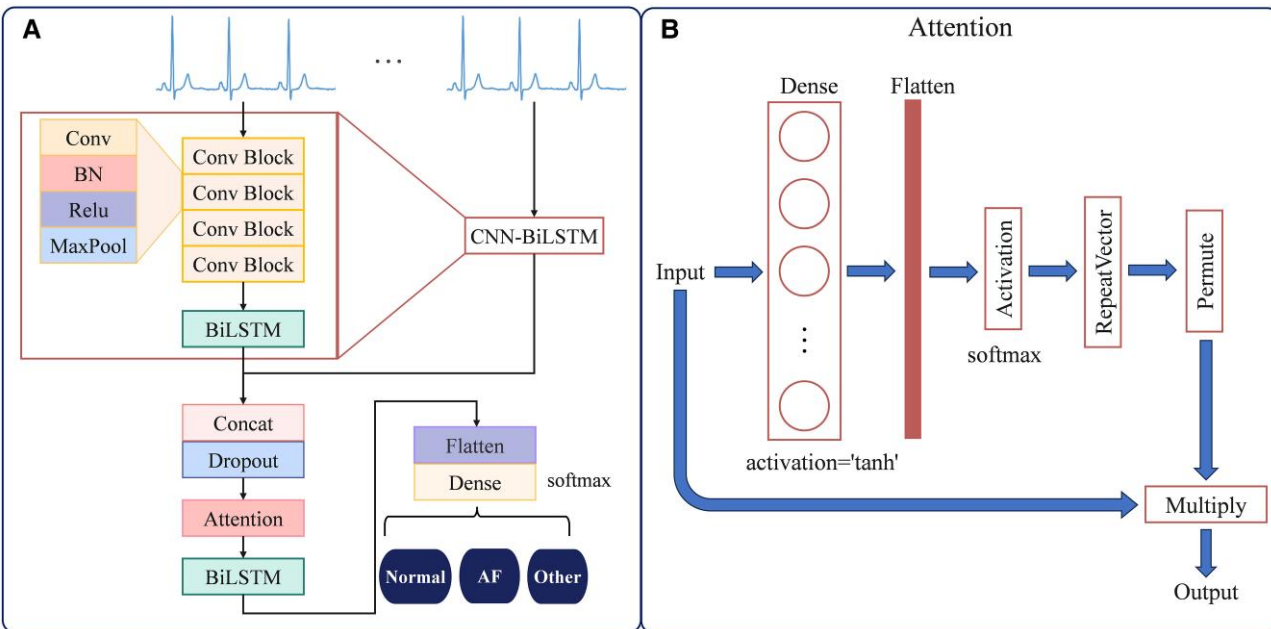
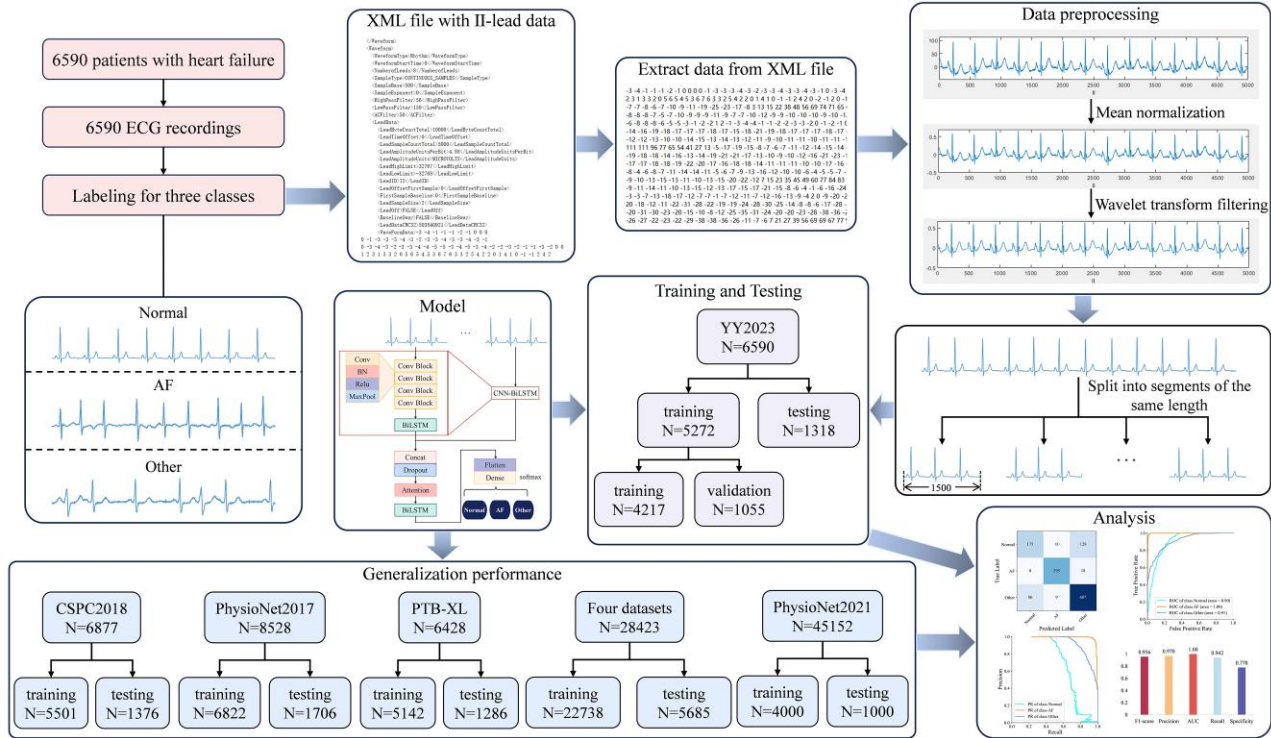
Model evaluation metrics

We primarily employ F1-score, Precision, AUC (area under the ROC curve), Recall, and Specificity as evaluation metrics for assessing model performance. We emphasize the F1-score, Precision, and AUC, with the F1-score serving as the harmonized average of Precision and Recall. It effectively gauges the model's capability to accurately and comprehensively detect, with higher values indicating superior performance. The formula for the F1-score is shown below.²⁸

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Table 1 Basic dataset information

Dataset	Frequency	Lead	Length	Normal	AF	Other	All
YY2023	500 Hz	12	10s	1542 (23.40%)	1470 (22.31%)	3578 (54.29%)	6590
CPSC2018	500 Hz	12	6–60s	918 (13.35%)	1098 (15.97%)	4861 (70.68%)	6877
PhysioNet2017	300 Hz	1	9–61s	5050 (59.22%)	738 (8.65%)	2740 (32.13%)	8528
PTB-XL	500 Hz	12	10s	2000 (31.11%)	1587 (24.69%)	2841 (44.20%)	6428
PhysioNet2021	500 Hz	12	10s	7882 (17.46%)	7388 (16.36%)	29 882 (66.18%)	45 152



Area under the ROC curve is a robust metric used for the performance of the classification model, and the higher the AUC, the better the performance. Precision becomes crucial, especially in scenarios with imbalanced proportions of positive and negative samples, as it provides insights into

the accuracy when dealing with uneven sample distributions. Additionally, we calculate the confusion matrix, ROC curve, and PR curve. The confusion matrix facilitates a comparison between the model's predicted labels and the true labels. ROC curves illustrate the trade-off between the true

positive rate and the positive predicted value at different thresholds. It provides an overall assessment of the model's performance, regardless of the specific threshold used for classification. PR curve is a valuable tool for expressing distinct preferences for precision and recall. It emphasizes the model's ability to capture as many true positives as possible, which is crucial when the cost of false negatives is high.

Results

Performance test of CLA-AF model

Figure 3A illustrates the model's loss value evolution across epochs. The loss undergoes a swift decline in the initial stages, converging towards stability as the epoch approaches 100. In Figure 3E, the model exhibits outstanding performance in recognizing AF, with F1-score, Precision, and AUC values of 0.956, 0.970, and 1.00, respectively. These results underscore the model's exceptional ability to accurately identify AF patterns in ECGs, affirming its high performance.

We performed a five-fold cross-validation of the CLA-AF model on YY2023, and the results are shown in Table 2. We employed Stratified K-Fold Cross-Validation, an enhanced version of the K-Fold cross-validation method that works well on unbalanced data. The whole dataset, just like the K-Fold, is divided into K-Folds of equal size, and the proportion of categories in each Fold is guaranteed to be the same as the proportion of the entire dataset. The performance averages for the five-fold cross-validation are also high, implying that the hyperparameter values are excellent. We also conducted the maximum vote on the classification results obtained from each K-Fold test, and the F1-score and Recall improved to 0.973 and 0.984, respectively.

Validation of generalization ability of CLA-AF model architecture

The initial aspect, focused on validating the CLA-AF model architecture, pertains to the overall generalization performance of the proposed algorithm. This involved training and testing on CPSC2018, PhysioNet2017, and PTB-XL datasets, utilizing the same methodology as YY2023. Table 3 and Figure 4 (Pre-oversample) present evaluation metrics for each of the four datasets. The F1-score, Precision, and AUC of YY2023 are 0.956, 0.970, and 1.00, respectively; the F1-score, Precision, and AUC of CPSC2018 are 0.827, 0.812, and 0.97; F1-score, Precision, and AUC of PhysioNet2017 are 0.589, 0.776, and 0.95; and F1-score, Precision, and AUC of PTB-XL are 0.871, 0.907, and 0.98, respectively. Most of all, the AUC on CPSC2018, PhysioNet2017, and PTB-XL reached above 0.95. The generalization performance of the proposed algorithm is shown to be excellent.

Nonetheless, the model's performance slightly decreases on the PhysioNet2017 dataset. The F1-score is only 0.589, and the Recall is lower than 0.5. When the amount of data of a few categories is small, the model is prone to bias, resulting in a low recognition rate for a few categories. Examination of Table 1 reveals that the AF data in PhysioNet2017 are notably limited, leading to a severe imbalance in categories. To address this, we implemented oversampling on all training sets, ensuring a balanced class ratio of 1:1:1. Specifically, the category with the largest number of samples served as the benchmark, and the deficit in categories with fewer samples was addressed by randomly selecting missing samples. This oversampling was exclusively applied to the training and validation sets. A comparison of test results before and

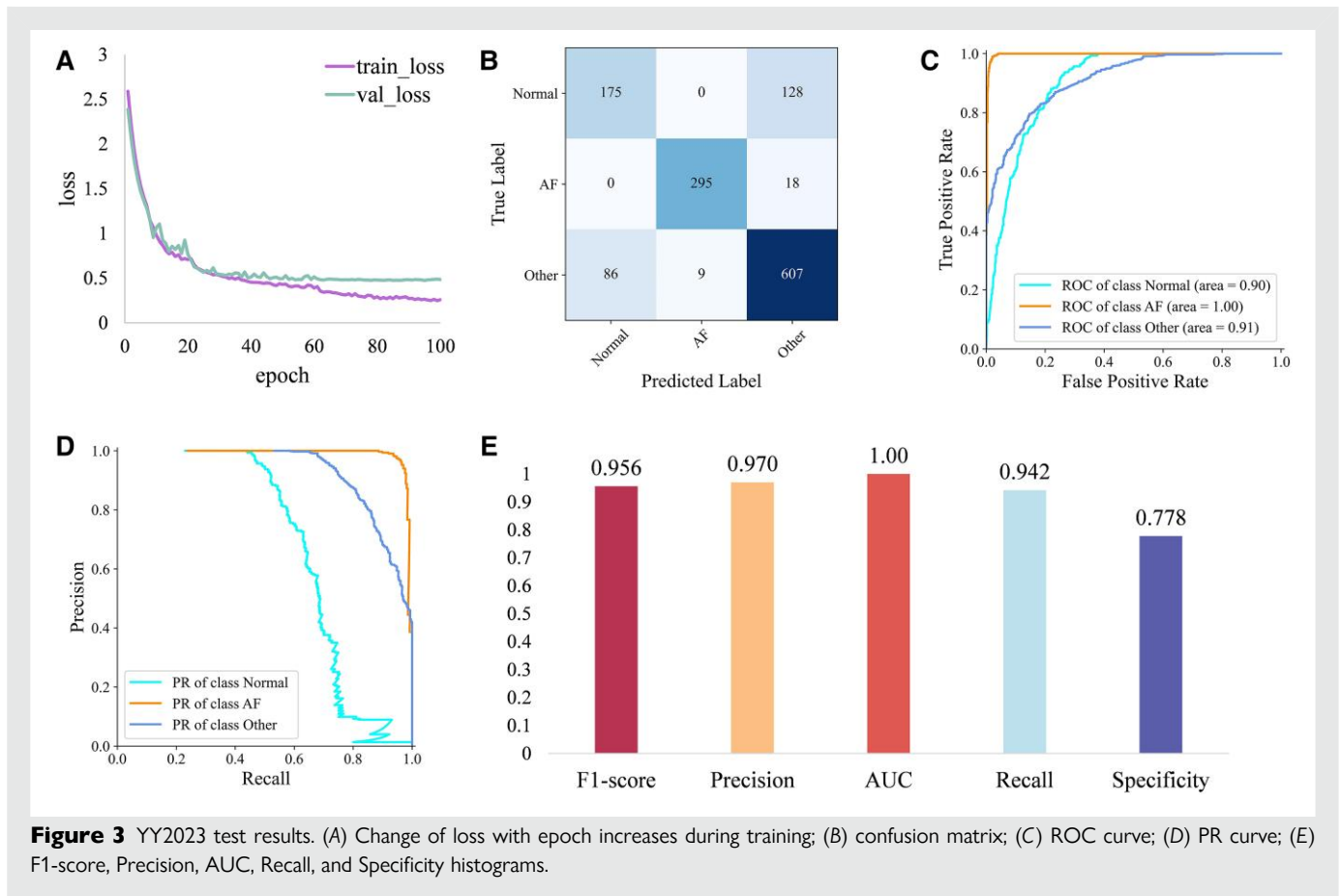


Table 2 Stratified five-fold cross-validation on YY2023

Stratified K-Fold	F1-score	Precision	AUC	Recall	Specificity
1	0.953	0.958	1.00	0.949	0.787
2	0.964	0.950	1.00	0.948	0.781
3	0.955	0.952	1.00	0.958	0.771
4	0.954	0.938	1.00	0.971	0.778
5	0.964	0.948	1.00	0.981	0.767
Mean	0.958	0.949	1.00	0.967	0.777
Post-test vote	0.973	0.963	1.00	0.984	0.788

Data in bold are the highest values.

after oversampling is detailed in [Table 3](#) and [Figure 4](#). Minimal alterations were observed in each evaluation metric for YY2023, CPSC2018, and PTB-XL, while improvements of 0.156 in F1-score, 0.260 in Recall, and 0.02 in AUC were noted for PhysioNet2017. These outcomes highlight the efficacy of balancing techniques in enhancing model performance, particularly for datasets characterized by severe class imbalances.

The ability of a trained model to recognize new data

The final aspect involves validating the discriminative power of the model trained on a specific dataset for new data. Initially, the generalization performance of the model trained with YY2023 for new data was assessed on ECGs encompassing all cases of AF from CPSC2018, PhysioNet2017, and PTB-XL. As illustrated in [Figure 5](#), the AF recognition accuracy is 0.923, 0.978, and 0.921 for

Table 3 The different training and test sets correspond to the values of the evaluation metrics

Train dataset	Test dataset	F1-score	Precision	AUC	Recall	Specificity
YY2023	YY2023	0.956	0.970	1.00	0.942	0.778
	CPSC2018	0.746	0.631	0.95	0.913	0.716
	PhysioNet2017	<u>0.163</u>	<u>0.089</u>	<u>0.62</u>	<u>0.978</u>	<u>0.056</u>
	PhysioNet2017 (resample)	<u>0.413</u>	<u>0.270</u>	<u>0.91</u>	<u>0.848</u>	<u>0.272</u>
	PTB-XL	0.794	0.696	0.95	0.924	0.527
CPSC2018	YY2023	0.867	0.958	0.98	0.792	0.752
	CPSC2018	0.827	0.812	0.97	0.843	0.825
	PhysioNet2017	<u>0.181</u>	<u>0.100</u>	<u>0.66</u>	<u>0.964</u>	<u>0.087</u>
	PhysioNet2017 (resample)	<u>0.469</u>	<u>0.334</u>	<u>0.91</u>	<u>0.784</u>	<u>0.297</u>
	PTB-XL	0.799	0.829	0.95	0.772	0.579
PhysioNet2017	YY2023	0.171	0.789	0.64	0.096	0.640
	CPSC2018	0.252	0.535	0.70	0.165	0.746
	PhysioNet2017	0.589	0.776	0.95	0.475	0.823
	PTB-XL	0.176	0.446	0.57	0.109	0.490
PTB-XL	YY2023	0.856	0.964	0.99	0.770	0.422
	CPSC2018	0.810	0.779	0.96	0.843	0.464
	PhysioNet2017	<u>0.192</u>	<u>0.112</u>	<u>0.63</u>	<u>0.669</u>	<u>0.160</u>
	PhysioNet2017 (resample)	<u>0.461</u>	<u>0.331</u>	<u>0.91</u>	<u>0.755</u>	<u>0.504</u>
	PTB-XL	0.871	0.907	0.98	0.838	0.772
YY2023 (1:1:1)	YY2023	0.955	0.927	1.00	0.990	0.767
CPSC2018 (1:1:1)	CPSC2018	0.833	0.811	0.97	0.857	0.839
PhysioNet2017 (1:1:1)	PhysioNet2017	0.775	<u>0.756</u>	<u>0.97</u>	0.734	<u>0.810</u>
PTB-XL (1:1:1)	PTB-XL	0.882	0.876	0.98	0.887	0.774
Four datasets	Four datasets	0.876	0.878	0.98	0.875	0.771
Four datasets (1:1:1)	YY2023	0.951	0.973	1.00	0.930	0.772
	CPSC2018	0.868	0.824	0.98	0.917	0.818
	PhysioNet2017	0.687	0.713	0.96	0.662	0.802
	PTB-XL	0.891	0.899	0.98	0.884	0.664
	Four datasets	0.881	0.864	0.98	0.899	0.765
	YY2023	0.961	0.968	1.00	0.955	0.781
	CPSC2018	0.868	0.816	0.97	0.926	0.786
	PhysioNet2017	0.697	0.669	0.97	0.727	0.796
	PTB-XL	0.899	0.898	0.98	0.901	0.673

The bolded data are test results from the same dataset for both the training and test sets. The underlined data are the comparison of test results before and after the resampling of the PhysioNet2017 test set.

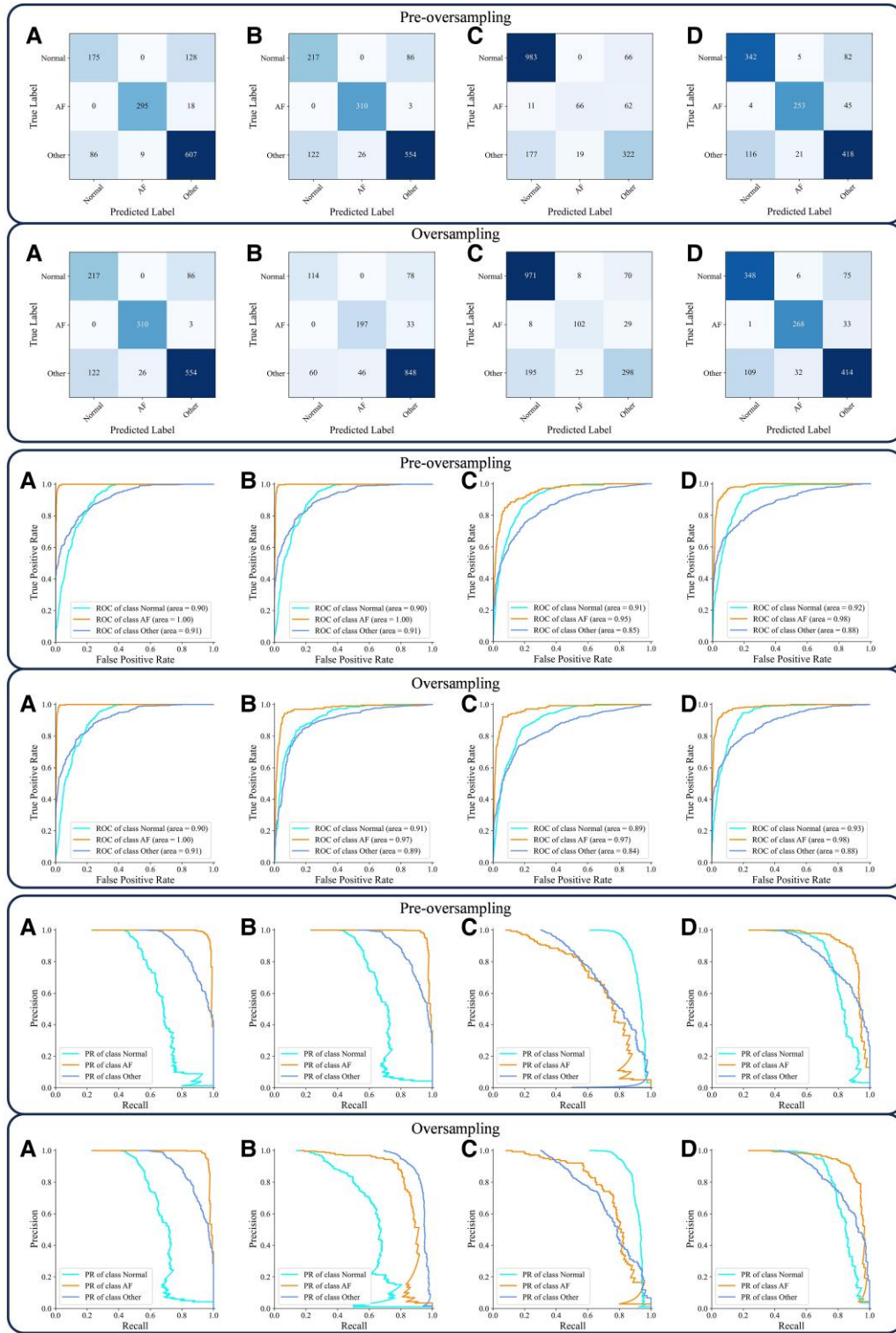


Figure 4 Comparison of the test results before and after oversampling. (A) Confusion matrix, ROC curve, and PR curve comparison before and after oversampling of YY2023; (B) confusion matrix, ROC curve, and PR curve comparison before and after oversampling of CPSC2018; (C) confusion matrix, ROC curve, and PR curve comparison before and after oversampling of PhysioNet2017; (D) confusion matrix, ROC curve, and PR curve comparison before and after oversampling of PTB-XL.

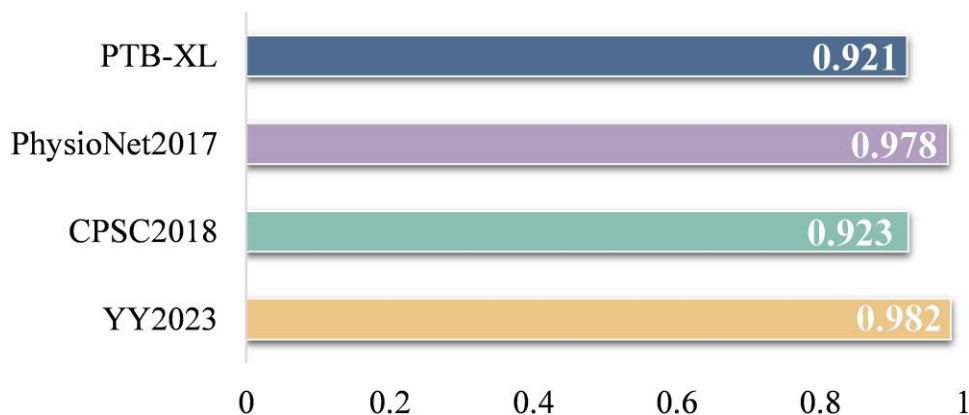


Figure 5 Accuracy of four datasets for AF recognition.

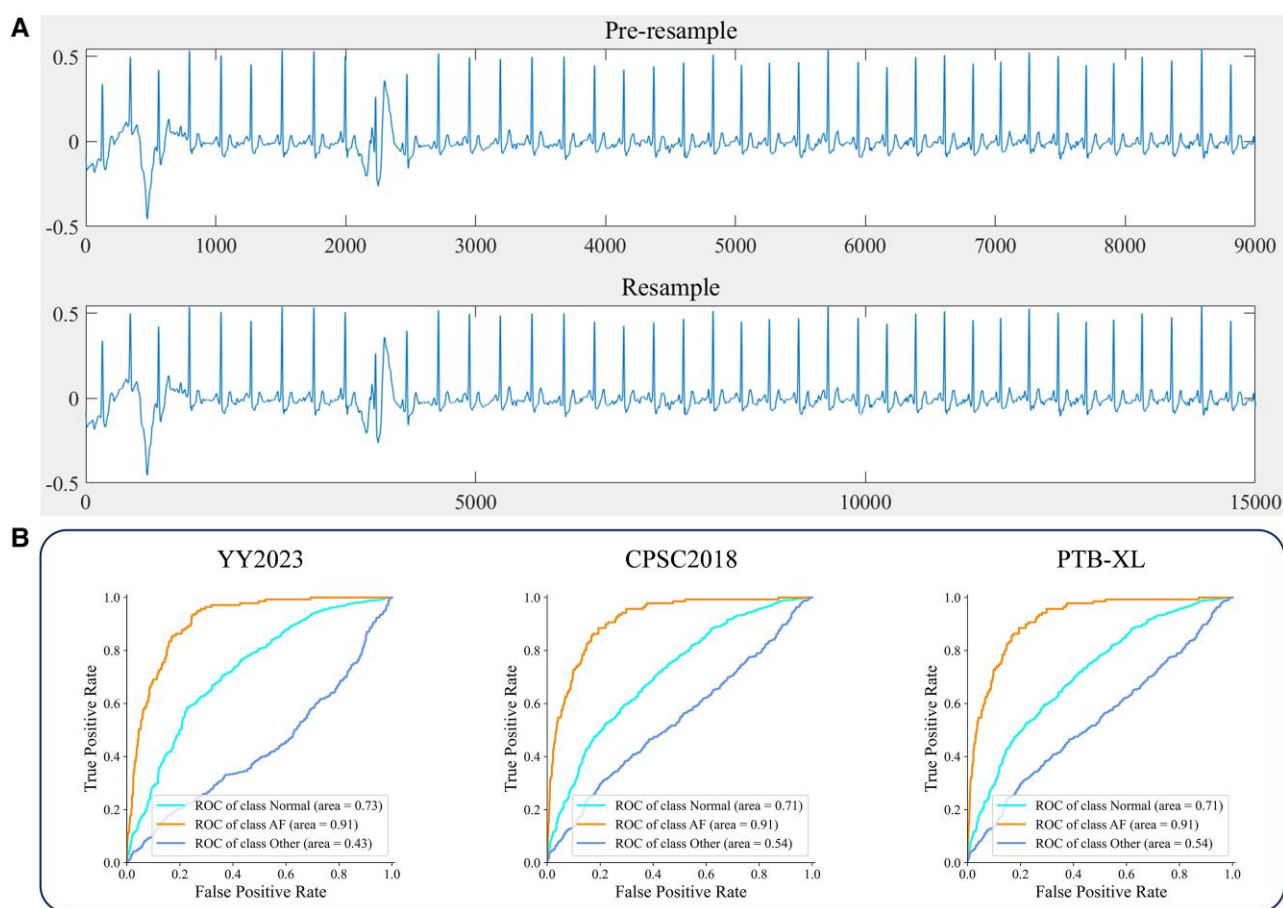


Figure 6 Test results after PhysioNet2017 resampling. (A) Comparison of PhysioNet2017 signals before and after resampling; (B) ROC curves after PhysioNet2017 resampling as a test set under the training set YY2023, CPSC2018, and PTB-XL, respectively.

CPSC2018, PhysioNet2017, and PTB-XL, respectively. These results affirm the model's robust generalization performance for AF.

However, it is crucial to note that the above assessments are specific to AF data. The presence of other classes in the test set might influence results, warranting further validation of prediction outcomes. We

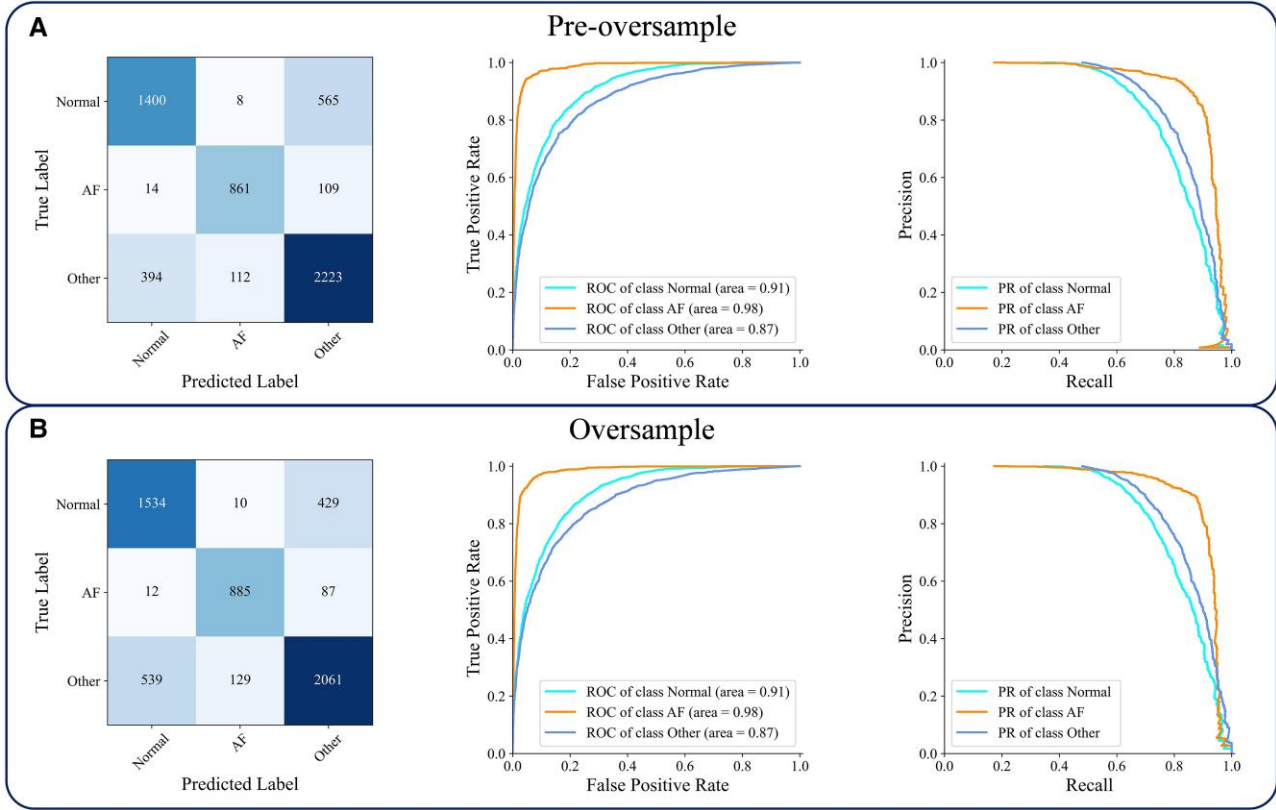


Figure 7 Training test results before and after oversampling of the hybrid dataset. (A) Before oversampling; (B) after oversampling.

cross-tested these four datasets. Table 3 presents values for each evaluation metric corresponding to the training and test sets.

Notably, models trained on YY2023, CPSC2018, and PTB-XL struggled to predict PhysioNet2017 data effectively. The AUC was 0.62, 0.66, and 0.63, respectively. Given PhysioNet2017's sampling frequency of 300 Hz, which is lower than the recommended 500 Hz or more for digitized circuit electrocardiographs to ensure high-fidelity recording of ECG waveforms, resampling was performed. This method utilizes the Fourier method to resample the original signal x , resulting in N samples along the axial direction. The starting value of the resampled signal is the same as that of x , but the spacing of the samples is adjusted accordingly ($\text{len}(x)/N * (x \text{ spacing})$). Figure 6A compares ECG signals before and after resampling, illustrating that resampling did not affect the ECG waveform. Subsequent tests using the resampled PhysioNet2017 test set, as shown in Figure 6B, revealed a substantial improvement in test results, with an AUC of 0.91. As shown in the underlined section of Table 3, the specific gravity of each index was greatly improved after resampling and before sampling. This underscores the significance of frequency adaptation between the training and test sets.

Generalization performance verification of hybrid training set

Building upon the results above, we posit that the model trained on the hybrid dataset may exhibit superior performance and enhanced generalization. The composite dataset from the four sources was initially combined and then divided into training and testing sets at an 8:2 ratio. Subsequent training, testing, and model evaluation were conducted, as illustrated in Figure 7A. The category ratio of the merged dataset is

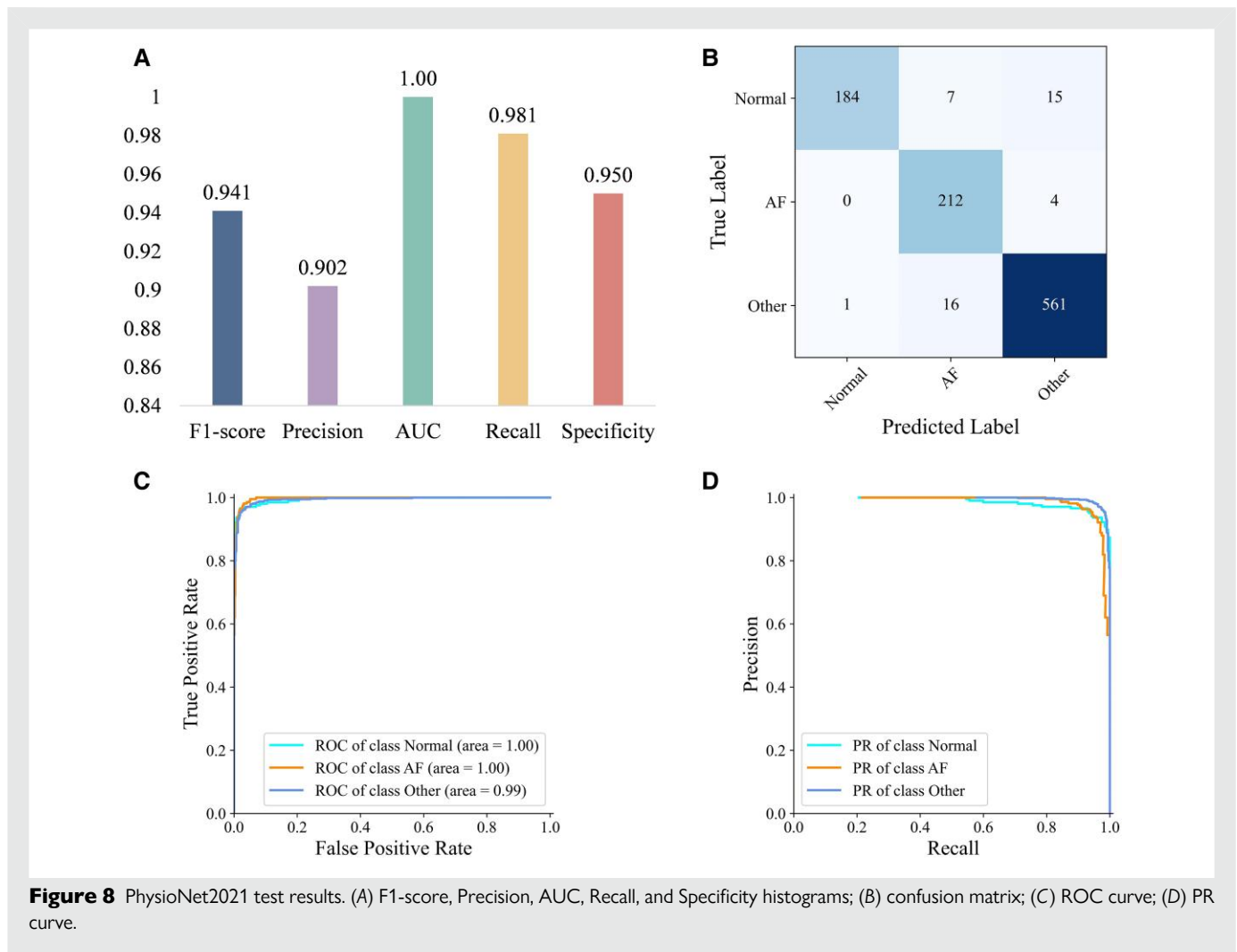
~2:1:3, with the training set also undergoing oversampling, as depicted in Figure 7B. Before oversampling, the F1-score, Precision, and AUC were 0.876, 0.878, and 0.98, respectively. Following oversampling, these metrics became 0.881, 0.864, and 0.98, respectively. The marginal difference in model performance before and after oversampling is negligible, and the achieved AUCs remain consistently high at 0.98, effectively identifying AF.

Testing was subsequently conducted on separate test sets for each of the four datasets, with the results presented in Table 3 and Figure 7. Whether it pertains to PhysioNet2017 at 300 Hz or the other three datasets at 500 Hz, the model derived from training on the hybrid dataset demonstrates proficient AF prediction. This observation underscores its robust generalization capabilities, making it applicable across diverse datasets and varied application scenarios.

The aforementioned results underscore the applicability of the CLA-AF model across diverse datasets, demonstrating robust AF recognition capabilities. It is evident that the model performs well across various datasets. Once again, this emphasizes specific requirements for the training set, with the hybrid dataset proving superior in terms of training efficacy and heightened generalization abilities.

Additional validation on more diverse external datasets

To further validate the robustness and applicability of the CLA-AF model and its ability to detect AF on more datasets, we added the PhysioNet2021 dataset. The dataset includes annotated 12-lead electrocardiogram records from six sources in four countries on three continents. We classified it into Normal, AF, and Other, using the same pre-processing methods. We randomly selected 5000 ECGs from



PhysioNet2021 for training and testing. The values of each index, confusion matrix, ROC curve, and PR curve are shown in Figure 8. The F1-score was 0.941, Precision was 0.902, AUC was 1.00, Recall was 0.981, and Specificity was 0.950.

And to verify the existence of overfitting in our model, we used PhysioNet2021 for additional validation of the model trained with YY2023. A total of 1000 AF ECGs were randomly selected for testing, and the accuracy was 0.952, which proved the generalization ability of the model, and AF ECGs from multiple countries could still be well recognized. From a wide range of data sources in PhysioNet2021, we can infer that CLA-AF can also maintain high AF recognition accuracy in different clinical settings, and its robustness and applicability are strong.

Model interpretability

In the realm of deep learning, models are frequently regarded as 'black boxes,' and the CLA-AF model is no exception. Due to its intricate internal mechanisms, we are unable to elucidate its decision-making process. Our CLA-AF model has demonstrated exceptional performance in diagnosing AF; however, we remain uncertain about which specific features capture the model's attention. To address this, we employed the SHapley Additive exPlanations (SHAP) method to interpret the model's predictions.²⁹ The SHAP method assigns a value to each feature that quantifies how much the absence of that feature would alter the model's predictions, essentially indicating each feature's influence on those predictions.

We computed SHAP values for all three ECGs and visualized both these values and their corresponding ECG signals. As illustrated in Figure 9, this provides an interpretation of the model's prediction results based on ECG instances from seven distinct patients, showcasing only one ECG segment per patient. Features with significant contributions are highlighted in pink. Three of the AF ECGs showed that the model mainly focused on the recognition of irregular QRS wave clusters and f waves formed after the absence of P waves, an observation consistent with the diagnostic criteria for AF. Conversely, a normal ECG emphasizes both waveform presence and regularity, while other anomalies focus specifically on waves that deviate from normal patterns.

Discussion

In this study, the AF recognition model CLA-AF was constructed, trained, and tested on our YY2023 dataset, and the model performed well. CLA-AF is a two-stage model composed of CNN, BiLSTM, and Attention. Then, training tests were conducted on CPSC2018, PhysioNet2017, and PTB-XL to validate the generalization ability of the model architecture and show its applicability to different datasets, where it was found that class imbalance of the datasets made the training worse, while the performance improved after oversampling. The generalization ability of the trained model to new data was further explored by testing the effect of AF recognition between different datasets, and it was found that the frequency of ECG sampling is a crucial factor affecting

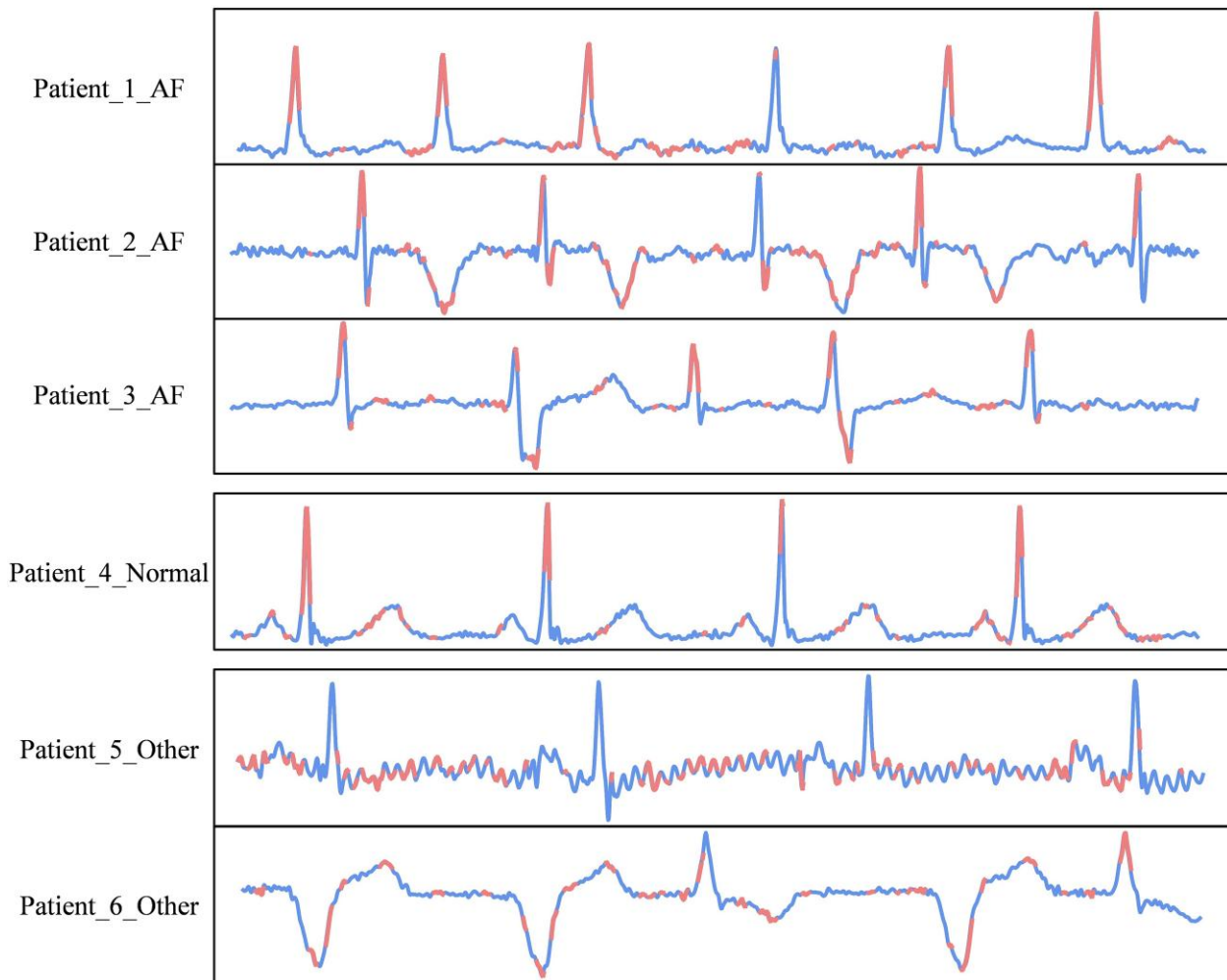


Figure 9 Model interpretation of ECG examples from six different patients. Patients 1–3 were AF, Patient 4 was Normal, and Patients 5 and 6 were other abnormalities.

the generalization ability, with performance improving after resampling. Finally, we merge the four datasets and then retrain them for testing. We find that the model obtained from the hybrid dataset has the best generalization ability and can accurately identify AF from different datasets. The SHAP values visualization results demonstrate that the model's interpretation of AF aligns with the diagnostic criteria of AF.

The ECG is one of the most essential tools for diagnosing and treating patients with AF. Electrocardiogram data can be either 1D amplitude data or a 2D image, plotted to form an ECG map, which can be treated as image processing. Li et al.³⁰ have used a novel ECG feature map for AF recognition practice by drawing the variation in the length of RR intervals into a grid map to form a feature map and then using SVM to classify. The raw data generated by the ECG machine are the amplitude data of the ECG signal; using image recognition, one needs to draw an ECG image first, then calculate the length of the RR interval and draw it to the graph to form the feature map, which is a complicated process. The accuracy of AF recognition based on the RR interval only may need to be improved. Therefore, modelling with raw ECG signals is more accurate, and the procedure is more straightforward.

Long Short-Term Memories are inherently designed to predict the output of the next moment based on the temporal information from previous

instances. Still, in certain problems, the current output may not only be influenced by the past state but may also be correlated with future states. Faust et al.³¹ used LSTM with RR interval signals, achieving an impressive overall average accuracy of 98.6% on the MIT-BIH dataset. Despite the high accuracy, the model's applicability to other datasets remains uncertain. And BiLSTM better captures the dependencies in the sequences by providing two independent input sequences for the same layer of LSTM, processed sequentially and in reverse order, and merging them after obtaining two more informative sequences.³² Therefore, we used BiLSTM to identify ECG more accurately, achieving F1-score, Precision, and AUC of 0.956, 0.970, and 1.00, respectively.

For generalization enhancement, we not only segment the ECG into segments of the same length and use bidirectional LSTM but also add Attention to filter features in the network to focus on more critical features while avoiding overfitting. The model constructed by Wang et al.²³ consists of a 33-layer CNN architecture and an NCBAM module, with an average F1-score of 0.966 and an AUC of 0.93 on the PTB-XL dataset. Although the generalization ability is more robust, the CLA-AF model outperforms this study with an AUC of 0.98 on PTB-XL. We also validated it on the CPSC2018 and PhysioNet2017 datasets, with AUC of 0.97 and 0.95, respectively, demonstrating the strong generalization ability of CLA-AF. Validation on PhysioNet2021

Table 4 Comparison with existing studies

Author	Method	Dataset	Lead	F1-score	AUC	Precision	Interpretability
M. Kent ³⁷	TD and FD + CNN	PTB-XL	12	—	0.98	—	—
Y.-Y. Jo ³⁸	XDM	PTB-XL + Georgia + Chapman + CPSC	12	0.96	0.97	0.961	+
A. Sbrrollini ³⁹	CNN	PhysioNet2021	12	—	0.97	—	—
Y. Dong ⁴⁰	MBSF-Net	CPSC2018	12	0.84	—	0.837	—
S. Choi ⁴¹	LSTM + DL	PTB-XL + China dataset	1	0.94	0.98	0.93	+
G.-W. Yoon ⁴²	GAN + ResNet	MUSE + PTB-XL	1	0.81	—	0.76	—
C.-H. Hsieh ⁴³	end-to-end CNN	PhysioNet2017	1	0.78	—	0.78	—
Our study	CNN + BiLSTM + Attention	YY2023	1	0.96	1.00	0.97	+

from a wide range of data sources shows that CLA-AF can also maintain high AF recognition accuracy in different clinical settings, and its robustness and applicability are strong.

Shu et al.³³ used an automated system combining CNN and LSTM to diagnose five arrhythmia ECGs and achieved 98.10% accuracy on the MIT-BIH. Yildirim et al.³⁴ proposed a new model for wavelet sequences based on bidirectional LSTM that also recognized five arrhythmias on the MIT-BIH, achieving a high recognition rate. These studies proved the feasibility of combining CNN and bidirectional LSTM to identify ECG to make the AF diagnosis faster and more accurate, and the model has a wider scope of application. Despite the combination of CNN and Transformer layers can also capture local and global dependencies in the input data, the previous study by Kim et al.³⁵ still concluded that CNN + LSTM is superior to CNN + Transformer for AF recognition. We added Attention to constructing the AF diagnostic model CLA-AF based on combining CNN and bidirectional LSTM, and the experimental results also proved that it has a high accuracy rate and a strong generalization ability. However, we only need to identify the single-lead to achieve an accuracy rate of 0.970 without needing all the data in the 12-lead. This improves efficiency and a higher detection effect with less data volume.

When the categories of the dataset are severely imbalanced, the model is focused more on the categories with larger sample sizes, thereby reducing the learning effect of a small number of sample categories. Liu et al.³⁶ improved 0–0.06 in the F1-score by enlarging the sample sizes of a few categories. After oversampling the PhysioNet2017 training set, the F1-score and Recall improved by 0.156 and 0.260, respectively. We believe that oversampling the training set is required when the ratio of fewest to most classes is <15%. Different sampling frequencies of ECGs affect the prediction results, and after resampling the PhysioNet2017 test set, the accuracy improves substantially. Results from jointly training the four datasets underscore the achievement of an AUC exceeding 0.96 for each dataset. Therefore, if the training set contains all the data with different sampling frequencies and class ratios, the model obtained from training on this hybrid dataset may have better generalization ability.

Comparison with studies using ECG to detect AF, as shown in Table 4, although there is good accuracy in the literature, most of them are based on 12-leads, and we achieved a satisfactory diagnosis using a single lead. We can observe that the studies with higher F1-score and AUC use 12-leads for identification, while Choi et al.⁴¹ achieved an AUC of 0.98 using only one lead, demonstrating the efficacy of LSTM for AF identification. We used the same lead II as this study, but our F1-score, AUC, and Precision metrics were superior to that study. On the same dataset, we achieved 12-lead equivalents using a single lead as in Kent's PTB-XL and Dong's CPSC2018, and we even achieved an AUC of 1.00 on PhysioNet2021, superior to Sbrrollini's 0.97.^{37–43} The F1-score and AUC of our CLA-AF model

are not only greater in each method with 12-leads but also more prominent in each method with a single lead.

Jo et al.³⁸ require doctors to manually annotate features such as the presence of P waves and PR intervals for each ECG. Choi et al.⁴¹ segmented the ECG into bands such as PreQ and QRS. Sbrrollini et al.³⁹ converted the ECG signals from 1D to 2D greyscale images, but we save time and resources by not requiring additional data conversion and heartbeat segmentation. We validated the generalization ability on several datasets from different sources, and also tried to interpret the predictions of the model using the SHAP method. In conclusion, a multifaceted comparison in terms of metrics such as AUC, pre-processing work, generalization ability, and interpretability revealed the high recognition rate of the CLA-AF model for AF in a single lead and proving the effectiveness of our method.

Consequently, the CLA-AF model we devised proves accurate and effective for diagnosing AF. It showcases enhanced generalization across diverse datasets, introducing a novel perspective to further improve the efficiency and accuracy of AF diagnosis and establishing a research foundation for predicting AF risk.

Conclusion

In this study, we developed a CLA-AF model using CNN, bidirectional LSTM, and Attention to accurately identify AF in ECG. The results demonstrate the model's effectiveness in AF diagnosis, and its robust generalization ability was validated across diverse datasets, including CPSC2018, PhysioNet2017, and PTB-XL. We addressed challenges such as class imbalance and sampling frequency, finding that oversampling and resampling are effective solutions. Notably, training on a hybrid dataset yielded the best generalization performance, making the model applicable to different datasets. Finally, the robustness and applicability of the CLA-AF model in different clinical settings were verified in PhysioNet2021. And the SHAP values visualization results demonstrate that the model's interpretation of AF aligns with the diagnostic criteria of AF. This research not only introduces a valuable method for AF identification in ECG but also lays the groundwork for algorithms and research directions, contributing to precise treatment and risk prediction for AF patients.

Limitation

In this study, we developed a novel CLA-AF model and evaluated its accuracy and generalization ability in detecting AF, achieving promising results. However, several limitations remain. In clinical practice, the ability to predict AF may be more valuable than merely diagnosing it. Our study focused exclusively on diagnosis, without exploring the potential for prediction, which could be more meaningful if combined with

specific clinical data. Moreover, future research should aim to validate it across a broader range of datasets from diverse clinical settings.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Acknowledgements

During the preparation of this work, we used ChatGPT in order to improve the readability and language of the manuscript. After using this tool or service, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author contribution

Conceptualization, Y.G. and Z.C.; Methodology, Y.G. and Z.C.; Software, J.S.; Validation, J.S. and Y.B.; Formal analysis, J.S.; Investigation, F.C., Y.W., and M.Z.; Resources, Y.G., Z.C., and N.H.; Data curation, Y.B. and F.C.; Writing—original draft, J.S.; Writing—review & editing, Y.G. and Z.C.; Visualization, J.S.; Supervision, Y.G., Z.C., and N.H.; Project administration, Y.G.; Funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (grant no. 82372581), Natural Science Foundation of Hunan Province, China (grant no. 2022JJ30082), and Health Research Project of Hunan Provincial Health Commission (grant number: W20242007).

Conflict of interest: none declared.

Data availability

The code of our model is available at <https://github.com/Hrdw615/my-CLA-AF>. CPSC2018, PhysioNet2017, PTB-XL, and PhysioNet2021 are available at <http://2018.icbeb.org/Challenge.html>, <https://www.physionet.org/content/challenge-2017/1.0.0/>, <https://www.kaggle.com/datasets/arjunascagnetto/ptb-xl-atrial-fibrillation-detection/data>, and <https://physionet.org/content/challenge-2021/1.0.3/#files>, respectively. The YY2023 used in this study will be shared upon reasonable request to the corresponding author.

References

- Baman JR, Passman RS. Atrial fibrillation. *JAMA* 2021;**325**:2218.
- Otto CM. Heartbeat: diagnosis of subclinical atrial fibrillation by physicians and patients. *Heart* 2019;**105**:809–811.
- Al Rahhal MM, Bazi Y, AlHichri H, Alajlan N, Melgani F, Yager RR. Deep learning approach for active classification of electrocardiogram signals. *Inf Sci* 2016;**345**:340–354.
- Sánchez de la Nava AM, Atienza F, Bermejo J, Fernández-Avilés F. Artificial intelligence for a personalized diagnosis and treatment of atrial fibrillation. *Am J Physiol Heart Circ Physiol* 2021;**320**:H1337–H1347.
- Feeny AK, Chung MK, Madabhushi A, Attia ZI, Cikes M, Firouzian M, et al. Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology. *Circ Arrhythm Electrophysiol* 2020;**13**:e007952.
- Ladavich S, Ghoraani B. Rate-independent detection of atrial fibrillation by statistical modeling of atrial activity. *Biomed Signal Process Control* 2015;**18**:274–281.
- Babaeizadeh S, Gregg RE, Helfenbein ED, Lindauer JM, Zhou SH. Improvements in atrial fibrillation detection for real-time monitoring. *J Electrocardiol* 2009;**42**:522–526.
- Wegner FK, Plagwitz L, Doldi F, Ellermann C, Willy K, Wolfes J, et al. Machine learning in the detection and management of atrial fibrillation. *Clin Res Cardiol* 2022;**111**:1010–1017.
- Zhang K, Sun M, Han TX, Yuan X, Gu L, Liu T. Residual networks of residual networks: multilevel residual networks. *IEEE Trans Circuits Syst Video Technol* 2018;**28**:1303–1314.
- Parvaneh S, Rubin J, Rahman A, Conroy B, Babaeizadeh S. Analyzing single-lead short ECG recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation. *Physiol Meas* 2018;**39**:084003.
- Sannino G, De Pietro G. A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Gener Comput Syst* 2018;**86**:446–455.
- Li C, Zheng C, Tai C. Detection of ECG characteristic points using wavelet transforms. *IEEE Trans Biomed Eng* 1995;**42**:21–28.
- Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. *Eur Heart J* 2019;**40**:2058–2073.
- Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. *Eur Heart J* 2021;**42**:4717–4730.
- Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw Learn Syst* 2022;**33**:6999–7019.
- Lai D, Zhang X, Bu Y, Su Y, Ma CS. An automatic system for real-time identifying atrial fibrillation by using a lightweight convolutional neural network. *IEEE Access* 2019;**7**:130074–130084.
- Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. *Comput Biol Med* 2018;**93**:84–92.
- Cossu A, Carta A, Lomonaco V, Bacciu D. Continual learning for recurrent neural networks: an empirical evaluation. *Neural Netw* 2021;**143**:607–627.
- Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* 2019;**31**:1235–1270.
- Yang MU, Lee DI, Park S. Automated diagnosis of atrial fibrillation using ECG component-aware transformer. *Comput Biol Med* 2022;**150**:106115.
- Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal* 2022;**79**:102444.
- Xu P, Liu H, Xie X, Zhou S, Shu M, Wang Y. Interpatient ECG arrhythmia detection by residual attention CNN. *Comput Math Methods Med* 2022;**2022**:2323625.
- Wang J, Qiao X, Liu C, Wang X, Liu Y, Yao L, et al. Automated ECG classification using a non-local convolutional block attention module. *Comput Methods Programs Biomed* 2021;**203**:106006.
- Park J, Lee K, Park N, You SC, Ko J. Self-attention LSTM-FCN model for arrhythmia classification and uncertainty assessment. *Artif Intell Med* 2023;**142**:102570.
- Jiang M, Gu J, Li Y, Wei B, Zhang J, Wang Z, et al. HADLN: hybrid attention-based deep learning network for automated arrhythmia classification. *Front Physiol* 2021;**12**:683025.
- Martis RJ, Acharya UR, Min LC. ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. *Biomed Signal Process Control* 2013;**8**:437–448.
- Aiwiscal. CpSC scheme. Available from: https://github.com/Aiwiscal/CPSC_Scheme (2018). (7 2023).
- Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009;**45**:427–437.
- Zhang D, Yang S, Yuan X, Zhang P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *iScience* 2021;**24**:102373.
- Li Y, Tang X, Wang A, Tang H. Probability density distribution of delta RR intervals: a novel method for the detection of atrial fibrillation. *Australas Phys Eng Sci Med* 2017;**40**:707–716.
- Faust O, Shenfield A, Kareem M, San TR, Fujita H, Acharya UR. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. *Comput Biol Med* 2018;**102**:327–335.
- Wang H, Zhang Y, Liang J, Liu L. DAFA-BiLSTM: deep autoregression feature augmented bidirectional LSTM network for time series prediction. *Neural Netw* 2023;**157**:240–256.
- Oh SL, Ng EYK, Tan RS, Acharya UR. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Comput Biol Med* 2018;**102**:278–287.
- Yildirim Ö. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Comput Biol Med* 2018;**96**:189–202.
- Kim Y, Lee M, Yoon J, Kim Y, Min H, Cho H, et al. Predicting future incidences of cardiac arrhythmias using discrete heartbeats from normal sinus rhythm ECG signals via deep learning methods. *Diagnostics (Basel)* 2023;**13**:2849.
- Liu H, Zhao Z, Chen X, Yu R, She Q. Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records. *Comput Methods Programs Biomed* 2020;**196**:105639.
- Kent M, Vasconcelos L, Ansari S, Ghanbari H, Nenadic I. Fourier space approach for convolutional neural network (CNN) electrocardiogram (ECG) classification: a proof-of-concept study. *J Electrocardiol* 2023;**80**:24–33.
- Jo YY, Kwon JM, Jeon KH, Cho YH, Shin JH, Lee YJ, et al. Detection and classification of arrhythmia using an explainable deep learning model. *J Electrocardiol* 2021;**67**:124–132.

39. Sbröllini A, Tomassini S, Emaldi E, Marcantoni I, Morettini M, Dragoni AF, et al. Multiclass convolutional neural networks for atrial fibrillation classification. *Annu Int Conf IEEE Eng Med Biol Soc* 2022;**2022**:1288–1291.
40. Dong Y, Cai W, Qiu L, Guo Y, Chen Y, Zhang M, et al. Detection of arrhythmia in 12-lead varied-length ECG using multi-branch signal fusion network. *Physiol Meas* 2022;**43**:105009.
41. Choi S, Choi K, Yun HK, Kim SH, Choi HH, Park YS, et al. Diagnosis of atrial fibrillation based on AI-detected anomalies of ECG segments. *Heliyon* 2024;**10**:e23597.
42. Yoon GW, Joo S. Classification feasibility test on multi-lead electrocardiography signals generated from single-lead electrocardiography signals. *Sci Rep* 2024;**14**:1888.
43. Hsieh CH, Li YS, Hwang BJ, Hsiao CH. Detection of atrial fibrillation using 1D convolutional neural network. *Sensors (Basel)* 2020;**20**:2136.