Heliyon 9 (2023) e21627

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

SAR model for accurate detection of multi-label arrhythmias from electrocardiograms

Liuyang Yang[a,b,d,1], Yaqing Zheng[a,1], Zhimin Liu[c,1], Rui Tang[b], Libing Ma[d,e], Yu Chen[b], Ting Zhang[d,**], Wei Li[a,*]

[a] The Affiliated Hospital of Kunming University of Science and Technology. The First People's Hospital of Yunnan Province, Kunming, Yunnan, China
[b] Department of Management Science and Information System, Faculty of Management and Economics, Kunming University of Science and Technology, Kunming, Yunnan, China
[c] The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming, Yunnan,China
[d] School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[e] Department of Respiratory and Critical Care Medicine, the Affiliated Hospital of Guilin Medical University, Guilin, Guangxi, China

ARTICLE INFO

Keywords:
SAR model
Multi-label
Arrhythmias
ECG

ABSTRACT

Objective: Arrhythmias are prevalent symptoms of cardiovascular disease, necessitating accurate and timely detection to mitigate associated risks. Detecting arrhythmias from ECGs quickly and accurately holds great significance in preventing heart disease and reducing mortality. This research endeavors to outperform previous studies by developing a scientific neural network model capable of training and predicting ECG signals for 11 categories of arrhythmias, accounting for up to 5 co-existing labels.

Methods: In this study, we initially address the issue of imbalanced datasets by employing Borderline SMOTE and Cluster Centroids techniques during preprocessing. Subsequently, we propose a novel SAR model that combines attention and resnet mechanisms. The dataset is subjected to a 10-fold validation process to train and evaluate the model. Finally, several metrics such as HammingLoss, RankingLoss, F1-score, AUC and Coverage are used to evaluate the model.

Results: By evaluating the results of the tests, the average Hamming Loss is 1.12 %, the average Ranking Loss is 1.17 %, the average Micro F1-score is 98.46 %, the average Micro AUC is 98.76 %, and the average Coverage is 3.2762. The results show that the SAR model outperforms previous related studies on the task of classifying arrhythmia signals with multiple categories and labels.

Conclusion: The SAR model demonstrated excellent performance in accurately classifying multi-category and multi-label arrhythmia signals, affirming its scientific validity. Compared with previous studies, the model achieves a certain improvement in performance, which can help cardiologists to achieve scientific and accurate diagnosis of arrhythmia diseases.

* Corresponding author.
** Corresponding author.
E-mail addresses: yly@stu.kust.edu.cn (L. Yang), zyq15812093227@163.com (Y. Zheng), ych_lzm@163.com (Z. Liu), tangrui@kust.edu.cn (R. Tang), malibing1984@163.com (L. Ma), ychen@kust.edu.cn (Y. Chen), zt0416@126.com (T. Zhang), lw13908701155@163.com (W. Li).

[1] Liuyang Yang, Yaqing Zheng and Zhimin Liu contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.heliyon.2023.e21627
Received 17 January 2023; Received in revised form 6 October 2023; Accepted 25 October 2023
Available online 28 October 2023

1. Introduction

As per the statistics from the World Health Organization (WHO), cardiovascular diseases lead to the maximum number of mortalities globally, resulting in over 10 million deaths annually. Arrhythmia, which is a common problem in cardiology, is when the heart's activity becomes too fast, too slow, or irregular, or that the order of its parts becomes disordered. The swift and precise diagnosis of arrhythmia is tremendously valuable for the therapy of cardiac diseases and saving lives [1,2].

Currently, the identification of arrhythmia primarily depends on the electrocardiogram (ECG) [3,4]. A Surface electrocardiogram is the most simple, cheap and accurate method to diagnose arrhythmia. In actual clinical diagnosis, doctors usually collect a long period of ECG signals to determine whether patients have related heart diseases, which requires cardiologists to spend a long time analyzing patients' ECG, which brings a huge workload to doctors [5,6].

At the same time, the accuracy and speed of arrhythmia detection are very high. Once misdiagnosis or failure to intervene in the best treatment time will delay the patient's condition, which is serious and even life-threatening. Some patients may have more than one type of arrhythmia at the same time. Therefore, we need to establish an accurate automatic analysis model of ECG signals to identify such multi-category and multi-label ECG signals. The application of analytical models can reduce the difficulty and workload of cardiologists, increase the probability of detecting arrhythmia events, and thus improve the survival rate of patients with arrhythmia [7,8].

In the last decade, rapid progress has been made in finding patterns in images and signals using neural networks [9]. Deep learning, an advanced machine learning approach, facilitates the automatic derivation of crucial characteristics via the integration of diverse neural networks encompassing multiple convolutional layers, nonlinear variation layers, pooling layers, and fully connected layers. Deep learning has accomplished remarkable advancements in image recognition, speech recognition, and natural language processing. In ECG analysis, deep learning techniques have exhibited superior classification capabilities over conventional approaches when trained with ample data [10–13]. However, there are still some limitations and challenges in analyzing ECG signals using machine learning methods, such as Ardeti et al. [14] argued that there is a general problem of low accuracy in ECG signal analysis using machine learning methods.

The most of existing research has concentrated on developing ECG classifiers for single-label arrhythmia detection, aiming to categorize individual heartbeats into one of several predefined abnormality types. However, single-label models are ill-equipped to handle patients exhibiting multiple concurrent arrhythmia modalities, which is a commonly encountered clinical scenario [15–17]. The multi-label classification task, though better representative of real-world complexity, poses added challenges. With the rise in the number of arrhythmia types, the quantity of plausible label permutations surges exponentially, posing challenges to model training and assessment. Consequently, multi-label ECG classification has received limited attention compared to single-label approaches [18, 19].

In light of the aforementioned challenges, the present study aimed to develop a scientific and accurate solution for multi-label ECG arrhythmia detection, building on prior work [15]. We approached ECG abnormality diagnosis as a multi-label classification problem, and proposed a self-attention residual network (SAR) model tailored for this task. The key contributions of this research are as follows.

(1) The proposed SAR model can accurately classify fine-grained ECG morphological features, supporting recognition of 11 arrhythmia types.
(2) Our approach can handle multi-label ECG signals with up to 5 concurrent abnormality labels, better resembling real-world complexity.
(3) Innovative use of the SAR model to classify ECG signals with 11 categories and 5 labels, achieving better classification results than similar previous studies.

2. Related work

ECG signal based arrhythmia detection serves as a vital instrument extensively utilized in medicine for monitoring cardiovascular illnesses. Effective and precise identification of anomalous ECG signals has constituted a crucial subject probed by medical academia.

At present, Resnet is a powerful neural network structure that is widely used in various tasks [20]. It is also a relatively advanced deep network, which can be leveraged across diverse data modalities spanning images to time series [21]. It has the advantages of improving accuracy easily by increasing depth, being easy to optimize, and superior to other networks. It benefits from a shortcut module that enables the network to plunge deeper while preserving a relatively low complexity, thereby making learning easier. Some researchers have added attention mechanism. It strives to ascertain weights for the constituents, anticipating pivotal elements to possess greater weightage, and non-critical entities hold lower weight. The weight of the components mirrors the respective element's contribution towards the target objective [18].

Some researchers have deployed residual networks to ECG prediction, and the arrhythmia detection algorithm derived from one-dimensional CNN (1D-CNN) with residual blocks has achieved excellent performance [11,20,22]. Jihye [11], Zhu [23] Wang [24] et al. used the resnet model as the basic model of classification, and added one or two extruder blocks and excitation blocks to ordinary resnet to improve the performance, and the results showed that the model had a good performance. Wong et al. proposed staged neural network architecture for automatic coding to extract heartbeat, embed sequences, and train a multi-layer perceptron for classification [25]. Pardasani et al. developed a method based on a 1D-CNN with class-dependent thresholds for identifying arrhythmias from 12-lead ECG [26]. But all of these studies suffer from low metrics. Other researchers have attempted to combine resnet with other algorithmic mechanisms. For example, Rong et al.combined ResNet34 with GRU and simultaneously extracted local and sequence features from ECG for model training [12]. An improved FL function is proposed as a loss function to solve the problem of an unbalanced data set. But it doesn't work very well when it comes to predicting outcomes. Borra et al. designed the decoding workflow of time series classification based on Inception Time, ResNet, and X ResNet, the three most advanced architectures [27]. However, these algorithms have limited interpretability to the learned features.

In the study of attention mechanism, some scholars have also carried out related research. For example, Nan et al. designed a fine-grained multi-label ECG (FM-ECG) framework to detect abnormal cation mechanisms on ECG images through weakly supervised fine-grained classification, which can find potential identification sites adaptively [18]. They are fused only with image-level annotations. Secondly, a recurrent neural network (RNN) is used to deduce ECG label correlation. Feng et al. developed an open-source deep neural network combining ResNet with attention-based model to predict a variety of cardiac abnormalities in 12-lead ECG [21]. Li et al. proposed a new neural network structure to achieve the classification of 9 arrhythmias and added a new attention mechanism to the

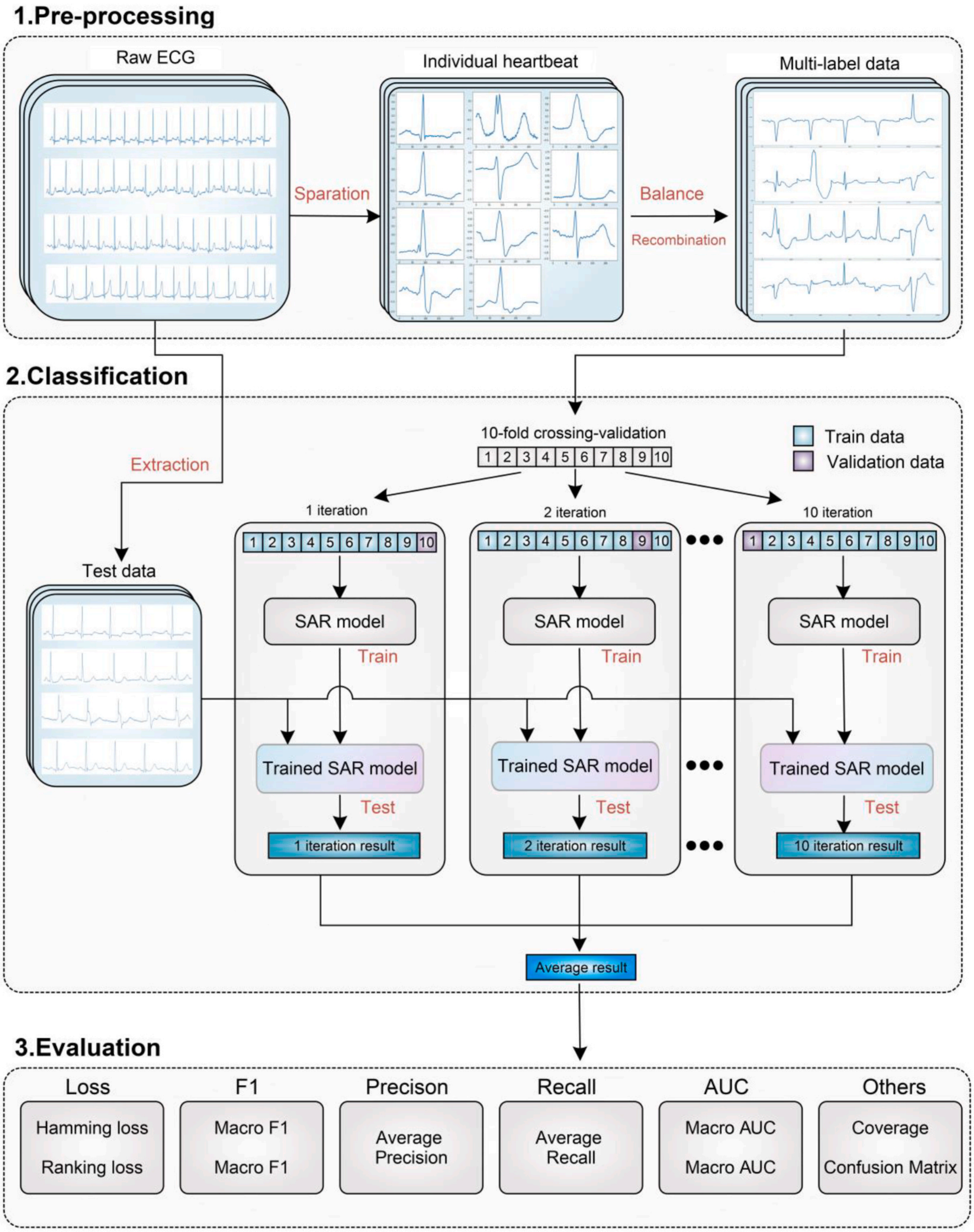


Fig. 1. Flow chart of data pre-processing, model training and result evaluation.

model, which is meaningful for data symmetry [28]. Liu et al. developed a deep 1D-CNN with a residual block and compression-excitation (SE) attention mechanism [29]. Wang et al. proposed a multi-label classification method based on a weighted graph attention network [30]. Li et al. proposed a neural network structure 12-lead ECG multi-label classification algorithm based on resnet, including data denoising, frame segmentation, data balance, and other pre-processing parts, combined with attention-based bidirectional short and long-duration memory (BiLSTM) [31]. Yang et al. adopted the improved stage-based residual network and split attention block residual network, but the performance of ECG, each abnormal recognition needs to be further improved [20].

By reviewing previous studies, ECG arrhythmia detection is usually categorized as a multi-label classification problem. After analyzing the existing literature related to the multi-label classification problem of arrhythmia, We summarized and found the following two main problems.

(1) Classification labels are not detailed enough and rarely cover multiple categories of arrhythmias. For example, in the arrhythmia recognition system developed by Liu, Li et al. , seven arrhythmias are classified and detected [32,33]. Sangha et al. trained convolutional neural networks to recognize six clinical labels defined by physicians covering rhythm and conduction disorders and a hidden gender [34]. Sun et al. proposed a new integrated multi-label classification model that combines seven multi-label classification methods to generate a new classifier [7].

(2) Multi-label classification network performance can be further improved. For example, Ge et al. designed a feature fusion based on multi-label correlation guidance of the ECG abnormal event detection model, the accuracy is 81.6 % [3]. Cai et al. created and trained a new deep learning architecture to model and capture label correlation of graph convolutional network (GCN) for multi-label classification of 12-lead ECG, the final F1-score is 60.3 % [10]. Osnabrugge et al. used a convolutional recurrent neural network (CRNN) to identify cardiac abnormalities in 2, 3, 4, 6, and 12 lead electrocardiogram numbers, the accuracy is 40 % [22].

In summary, we propose a novel SAR model according to the actual situation in this research, which can better cope with the problem of multi-label ECG analysis and achieve better performance.

3. Methodology

As illustrated in Fig. 1, our experiment encompasses three principal components, namely pre-processing, classification, and evaluation.

3.1. Pre-processing

For the dataset, we employed the MIT-BIH dataset, which is internationally accepted as a standard ECG dataset and is the most widely used ECG database in recent years [35,36]. This database is widely employed as experimental data in numerous related studies and has gained broad recognition in the academic sphere. Supplied with trustworthy information, meticulous notes, abundant cases and diverse ECG signal categories, the database can undergo training and examination through investigative algorithms, thereby establishing a robust foundation for detection research. The MIT-BIH dataset was also used to reflect the generalizability of the method in this paper and to facilitate comparison with other researchers' research methods.

The database contains 43 available samples, each containing labeled two-lead ECG data with each segment lasting 30 min. The labeled data underwent meticulous annotation on a per ECG cycle basis, with matching labeling documents furnished. In this dataset, 11 main representative heart rate categories are included, namely Normal beat (N), Atrial premature beat (A), Aberrated atrial premature beat (a), Ventricular escape beat (E), Atrial escape beat (e), Fusion of the ventricular and normal beat (F), Nodal (junctional) premature beat (J), Nodal (junctional) escape beat (j), Left bundle branch block beat (L), Right bundle branch block beat (R), and Premature ventricular contraction (V). In the study, these same 11 representative heart rate categories were selected for study. The types and number of each are shown in Table 1.

It is clear from the characteristics of the dataset that arrhythmias are often comorbid, more than one arrhythmia may be present in a

Table 1
Types of arrhythmia and corresponding labels.

Label	Type abbreviation	Type	Amount of beats
0	N	Normal beat	144847
1	A	Atrial premature beat	1454
2	a	Aberrated atrial premature beat	299
3	E	Ventricular escape beat	105
4	e	Atrial escape beat	31
5	F	Fusion of ventricular and normal beat	1603
6	J	Nodal (junctional) premature beat	163
7	j	Nodal (junctional) escape beat	33
8	L	Left bundle branch block beat	16137
9	R	Right bundle branch block beat	14500
10	V	Premature ventricular contraction	13797

given arrhythmia patient. Using the MIT-BIH dataset as an example, we performed a statistical analysis using histograms of marginal distributions.

As shown in Fig. 2, the edge distribution histogram visualizes the types of different heartbeats contained in all 43 available samples of the MIT-BIH dataset. The horizontal axis is the sample number and the vertical axis is the label, while the number of beat types contained in each sample and the total number of samples corresponding to each beat type are summed below and to the right of the scatter plot. For example, Sample 1 (S1) contains three types of beat labels, 0, 1 and 10, which correspond to the three categories of Normal beat (N), Atrial premature beat (A), and Premature ventricular contraction (V) in Table 1. In Sample 12 and Sample 18, there is only one type of heartbeat, all heartbeats are labeled as 0. While in Samples 6, 11, 20, 22, 23, 26, 29, 31 and 37, as many as five different types of labeled heartbeats are included.

Also, we can see from Fig. 2 that the different categories are extremely unevenly distributed. For example, there are 36 samples containing heartbeats with label 0, but only Sample 37 contains heartbeats with label 4. The number of different heartbeats is also extremely unbalanced, and the exact number of heartbeats in different categories is shown in Table 1.

In summary, the target dataset in this study has the following characteristics.

(1) The data have multiple categories. There are 11 different categories of heartbeats.
(2) The data have multiple labels. There may be at most 5 labels at the same time and at least 1 label.
(3) The data of various categories exhibits an extremely skewed distribution.
(4) The quantities across the diverse data categories manifest severe imbalance.

To achieve excellent training and testing results, in the research, the following work is carried out to address the above characteristics of the dataset.

3.1.1. Data segmentation

We utilize the Wave From Database (WFDB) module to process the database, WFDB is used to read annotation files and find the R peak location [2,37], including waveforms, amplitudes, periods, and associated labels, to facilitate the subsequent training. The data preprocessing comprises two primary components. First, extraction is performed. We extract the complete heart waves by an individual as the original data set. Then, segmentation is performed. In the processing of heartbeats, the features of a segment of heartbeat waves are mainly PQRST waves. Accordingly, for heartbeat segmentation, the R-peak constitutes the reference anchor, and a standard interval of 0.3s forward and 0.4s backward is used to separate the heartbeat into discrete instances. Such a beat of length 0.7s can effectively contain the cardinal traits of the heartbeat wave. Ultimately, the discrete heartbeats were mapped to the respective tags per the annotations within the original database. This resulted in several standard heartbeats with labels.

3.1.2. Data balance

In this research, we use the Borderline SMOTE and Cluster Centroids method to deal with the uneven distribution and an unbalanced number of data sets. Borderline-SMOTE (Borderline-Synthetic Minority Oversampling Technique) is an improved oversampling algorithm for SMOTE. The algorithm exclusively leverages the frontier minority instances for fabricating novel examples, thereby enhancing sample distribution [38].

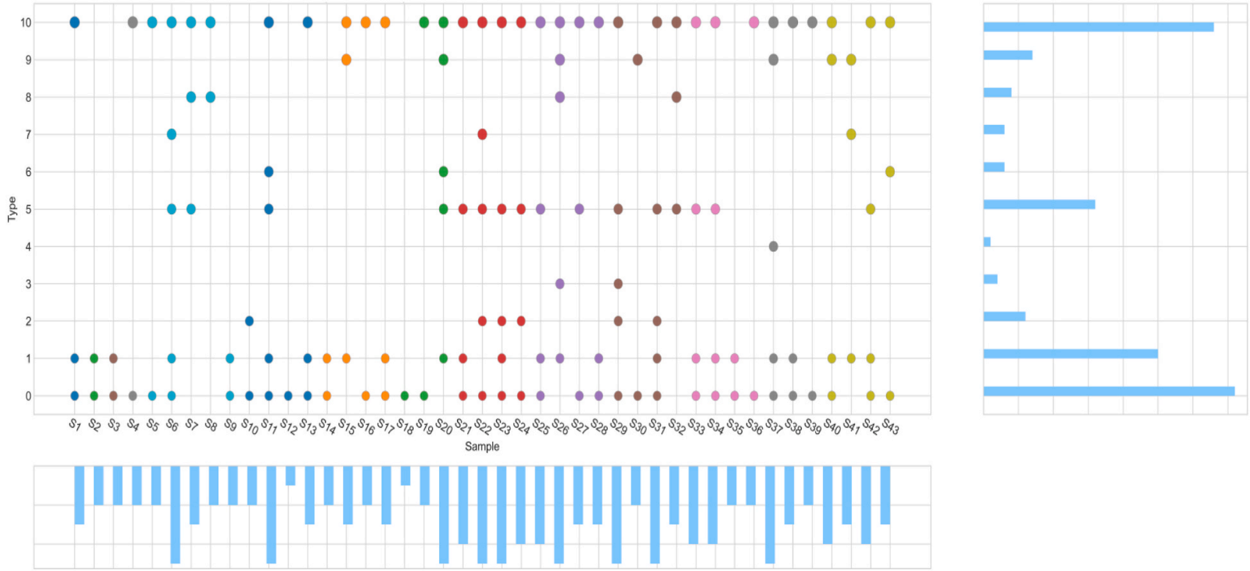


Fig. 2. Types of arrhythmias contained in different samples.

SMOTE represents an approach for interpolating between a limited set of sample categories to generate supplementary examples. For a minority class sample A, the nearest neighbor sample B is randomly selected, and A point C is randomly selected as a new minority class sample from the concatenation of A and B. Specifically, for a minority class sample x_i use the k nearest neighbor method (the value of k needs to be specified in advance) to find the k minority class samples that are closest to x_i, where the distance is defined as the Euclidean distance in the n-dimensional feature space between the samples. Subsequently, one of the k nearest neighbors are randomly chosen to generate a novel sample, the formula is depicted in ①.

$$\mathbf{x}_{\text{new}} = \mathbf{x}_\mathbf{i} + (\widehat{\mathbf{x}_\mathbf{i}} - \mathbf{x}_\mathbf{i}) \times \boldsymbol{\delta} \tag{1}$$

Where $\widehat{x}$ is the selected k nearest neighbor and δ is a random quantity, δ ∈[0,1].

Border-line SMOTE is an improved version of the SMOTE algorithm, which divides the sample into three categories. "Noise": all k-nearest neighbors belong to the majority class; "Danger": more than half of k-nearest neighbors belong to the majority class; "Safe": more than half of k-nearest neighbors belong to the minority class. Border-line SMOTE randomly chooses a sample from the "Danger" condition, applying the SMOTE algorithm to produce a novel example. The “Danger” state denotes frontier minority instances more susceptible to misclassification. Therefore, Border-line SMOTE exclusively synthesizes examples from minority categories adjoining the “Border”, whereas SMOTE uniformly handles all minority classes.

Cluster Centroids is a method to reduce the number of target samples by k-means clustering, as shown in Fig. 3. First, the target samples are clustered into different classes by k-means method, and the centroids of the target data are calculated, then, the data farthest from the centroids are removed. Finally, the downsampled data are obtained. Therefore, each category will be synthesized with the centroids of the k-means method rather than the original examples. Cluster Centroids method provides an efficient way to represent the reduction in the number of samples in the data clusters, however, the approach requires the data to be grouped into clusters. Moreover, the centroid quantity should be defined to enable the downsampled clusters to represent the original groups.

3.1.3. Data recombination

Since in the original sample, there is often one same heartbeat repeated many times in a row, namely there will be a large amount of single-label data, which will affect the model training. To improve the adaptability of the model, we need to randomly reorganize the heartbeats in all samples in the model after they are completely broken up. Since at most 5 labels appear in the same sample in the original sample, we also combine the individual heartbeats in groups of 5 to form the new features. The new features have at least 1 label and may have at most 5 labels in different combinations. Since we want to identify 11 categories, there will in principle be $\mathbf{C}_{11}^{1} + \mathbf{C}_{11}^{2} + \mathbf{C}_{11}^{3} + \mathbf{C}_{11}^{4} + \mathbf{C}_{11}^{5} = 1023$ combinations in total. In the meantime, we also transformed the labels of the reorganized data by converting multi-category labels to multi-hot labels. In Fig. 4, we show some of the restructured multi-label data features and their multi-hot labels. After converting the data labels to multi-hot form, the presence or absence of 11 types of heartbeats in the feature is represented by an 11-bit binary number arranged from left to right. For example, in the 1 label, there are only normal heartbeats.

3.2. Model structure

In the SAR model construction part, the network model is designed to enable accurate identification of multi-label ECG data, and to ensure that all evaluation metrics are optimal while upholding the concept of science and effectiveness, the model structure is shown in Fig. 5.

In the model, the multi-label data are first fed into the GlobalPooling layer to connect the data and the model and reduce the data dimensionality, and then, divided into four paths. Path 1, directly connected to SelfAttention Block, which contains one self-attention network layer; path 2, directly connected to ResNet Block after ZeroPadding, Convolution, and GlobalPooling layers; path 3, connected to SelfAttention Block and ResNet Block, and adding two residual connections; path 4 concatenate directly with the results of path 1, path 2 and path 3 without connecting any modules.

In path 1, given the multi-labeled input data constitute one-dimensional temporal sequences, an effective self-attention architecture is employed, chiefly enabling the model to concentrate on and derive cardinal traits while discounting less relevant

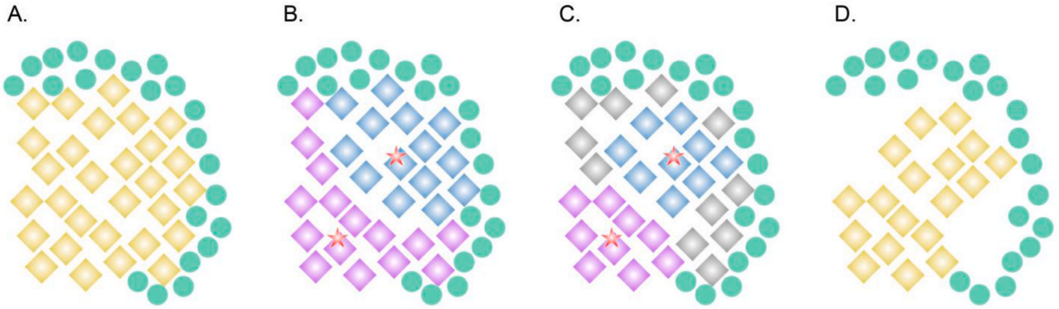


Fig. 3. Working principle of Cluster Centroids method. A: Raw data; B: Calculate centroids; C: Select furthest points; D: Resampled data.

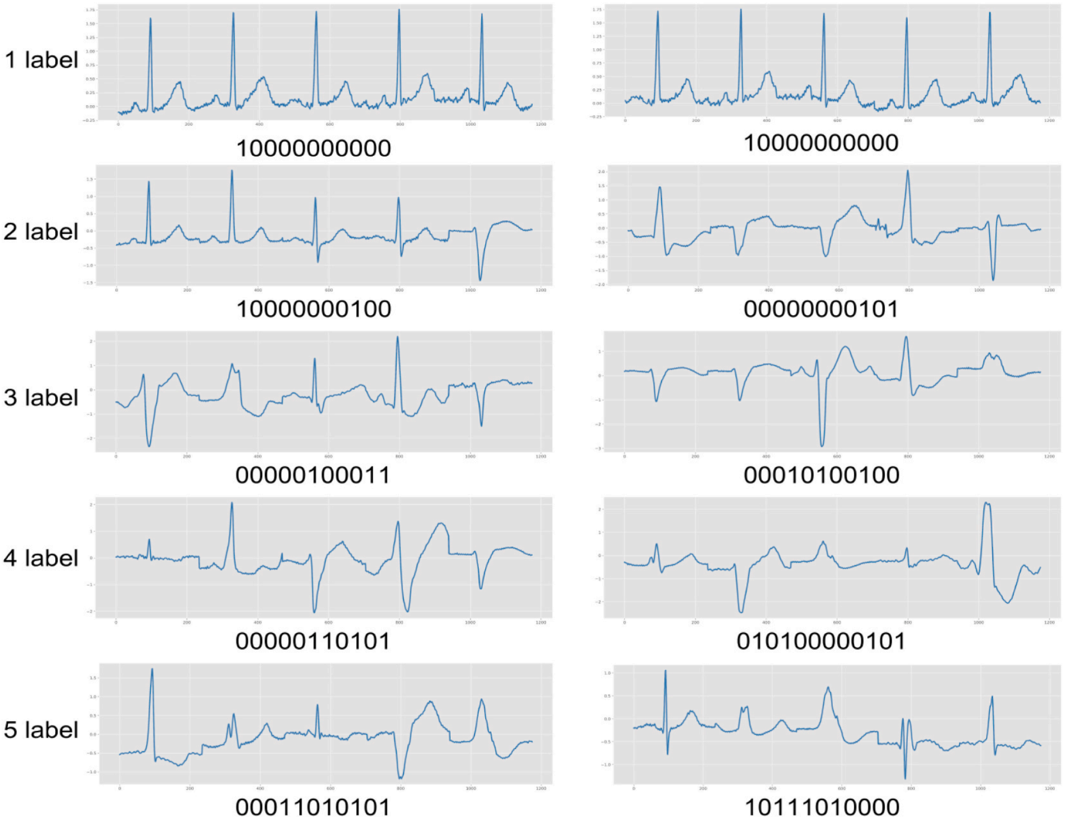


Fig. 4. Recombined multi-label data features and multi-hot labels.

characteristics.

In path 2, as the ResNet Block constitutes a deep neural network, the data may acquire an elongated feature vector post the ResNet Block layer. To regulate the ultimate vector length, a ZeroPadding layer is set at the beginning of path 2, padding with 0 to govern the post-convolution vector length. In the ResNet Block, we set 30 Convolution1D layers to extract the data features, in which several short-circuit connections are added to reduce the model complexity to reduce overfitting, and prevent gradient vanishing, as shown in Fig. 5.

When paths 1, 2, 3, and 4 are concatenated, after several Dense, Dropout, and GlobalPooling layers, finally in the Dense layer, a Sigmoid activation function is used to predict the probability of each label from 0 to 10, and the prediction results in a probability value of 0–1 for each label.

4. Experiment

This work operates on a deep learning framework of Keras with Tensorflow as the backend. The workstation used consists of 6144 MB GPU (NVIDIA GeForce RTX-3070), Intel i7–11700 K processor (3.60 GHz) and 32 GB RAM.

In the training process, we divided the data set into train set, validation set and test set. The number of data sets in each part is shown in Table 2.

During training, loss is "binary cross-entropy", optimizer is "Adam", learning rate is 0.0001 and batch size is 512. Due to the limited computing power of the computer, we train a total of 100 epochs. The training history is shown in Fig. 6.

To ensure a complete and balanced training of all training data and to evaluate the training effect in real-time, we used the 10-Fold cross-validation method.

After training, we tested the trained model with untrained data from raw data, and selected some representative test results for simple visualization in Fig. 7.

5. Evaluation

5.1. Evaluation metrics

To evaluate SAR model's performance, we use various evaluation methods such as Hamming Loss, Ranking Loss, AUC, Coverage,

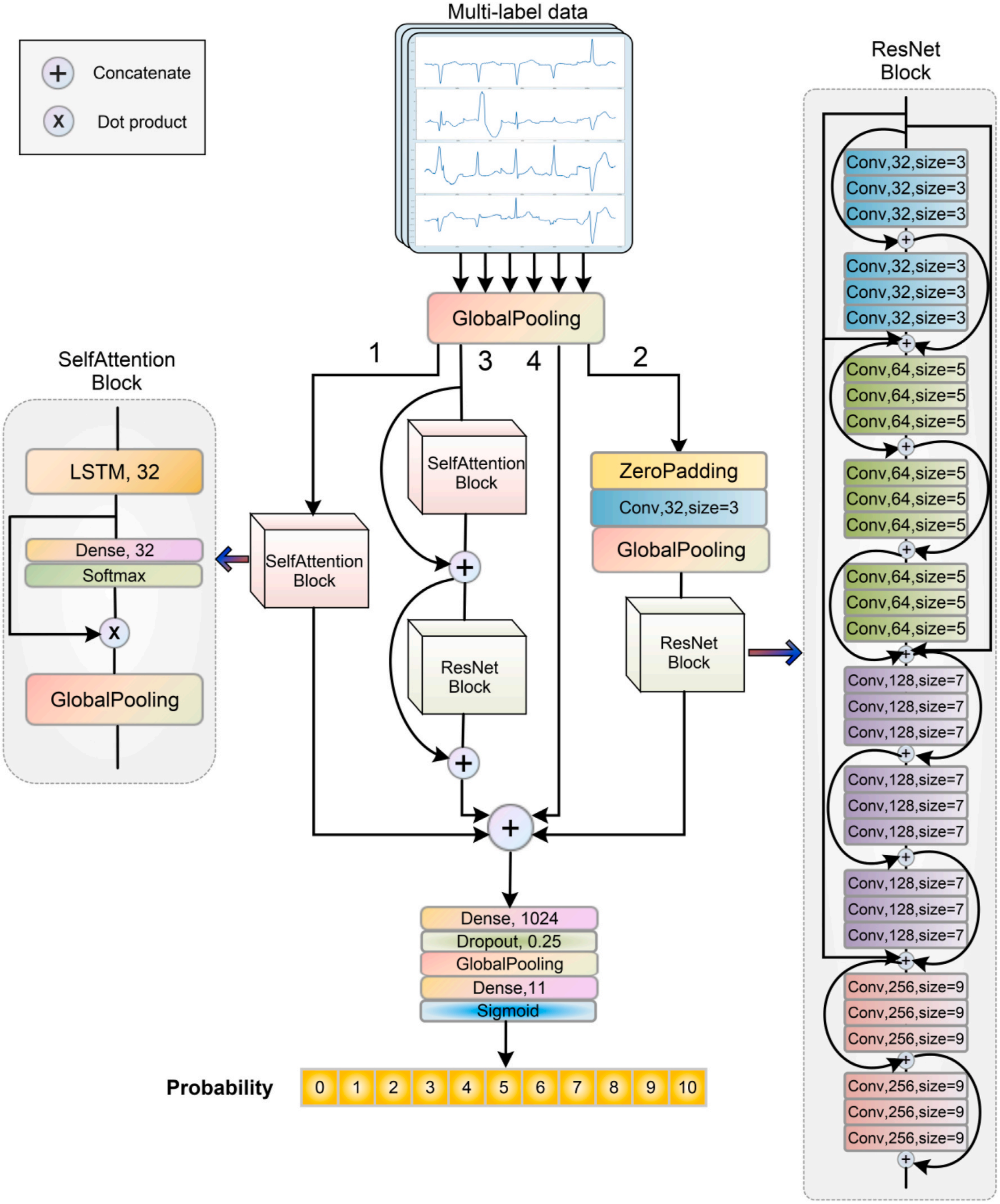


Fig. 5. SAR model structure.

Table 2
Sample parameters of each part.

Category	Train	Validation	Test	Total
Amount	11099	2775	3469	17343

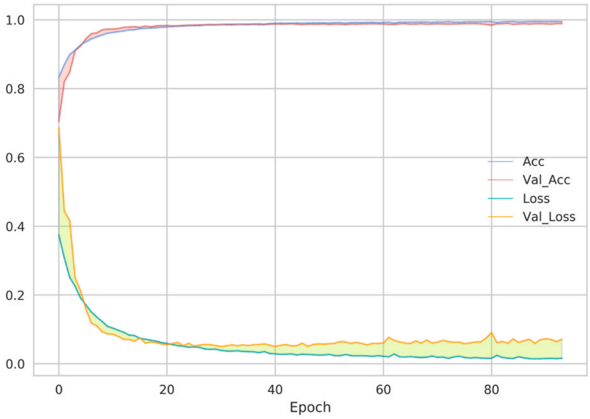
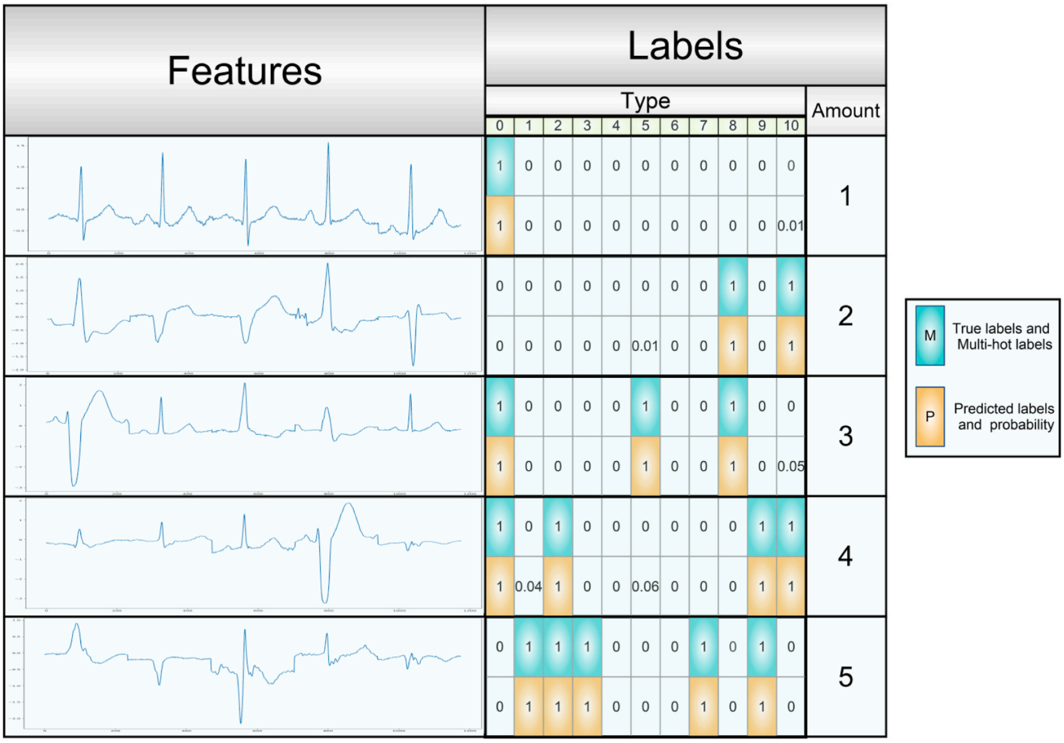


Fig. 6. The training and validation curves of the model: Accuracy and Loss.



Type 0	1	2	3	4	5	6	7	8	9	10	Amount
1	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0.01	
0	0	0	0	0	0	0	0	1	0	1	2
0	0	0	0	0	0.01	0	0	1	0	1	
1	0	0	0	0	1	0	0	1	0	0	3
1	0	0	0	0	1	0	0	1	0	0.05	
1	0	1	0	0	0	0	0	0	1	1	4
1	0.04	1	0	0	0.06	0	0	0	1	1	
0	1	1	1	0	0	0	1	0	1	0	5
0	1	1	1	0	0	0	1	0	1	0	

M True labels and Multi-hot labels

P Predicted labels and probability

Fig. 7. Partial representation of test results.

Precision, Recall, F1-score, etc [7,12,38,39]. The performance of the model is fully evaluated by these evaluation methods.

(1) Hamming Loss

$$\text{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}h(x_i)\Delta Y_i \tag{2}$$

Δ signifies the symmetric variance between two configurations. For instance, the symmetric difference between {1,2,3} and {3,4} is {1,2,4}. This index indicates the misclassification of the sample on a single label, namely the absence of relevant labels in the predicted label set or the presence of irrelevant labels in the predicted label set.

(2) Ranking Loss

$$\textbf{Ranking Loss} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\frac{\mathbf{1}}{|\mathrm{Yi}||\overline{\mathbf{Yi}}|}|\{(\mathbf{y}',\mathbf{y}'')||\mathbf{f}(\mathbf{x_i},\mathbf{y}')\leq\mathbf{f}(\mathbf{x_i},\mathbf{y}'')|\tag{3}$$

Where $\overline{\mathrm{Y}}\mathrm{i}$ is the complementary set of Yi in the label space, and $(\mathrm{y}',\mathrm{y}'')\in\mathrm{Yi} * \overline{\mathrm{Y}}\mathrm{i}$. This metric examines sorting mistakes in the category label sequences, whereby irrelevant labels precede applicable ones. Lower values indicate enhanced performance, with the optimum being 0.

(3) AUC

$$\mathrm{AUC} = \int_0^1 \mathrm{TPRdFPR} = \int_{\mathrm{P}}^{\mathrm{P_p}} *\frac{\mathrm{N_{p,p+\Delta p}}}{\mathrm{N}} = \frac{1}{\mathrm{P}*\mathrm{N}}\int \mathrm{P_p} * \mathrm{N_{p,p+\Delta p}} = \frac{1}{\mathrm{P}*\mathrm{N}}\int_0^1 \mathrm{P_p}\mathrm{dN_p}\tag{4}$$

AUC (Area Under Curve) is the area enclosed by ROC Curve and X coordinate axis. The full name of the ROC Curve is Receiver Operating Characteristic Curve. This curve is obtained with TPR(true positive rate) as the ordinate and FPR(false positive rate) as the abscissa.

P∗ N represents all the positive and negative sample pairs. The positive and negative sample pairs represented by $\mathrm{P_p} * \mathrm{N_p}, \mathrm{N_{p,p+\Delta p}}$ are the sample pairs composed of the positive sample with the predicted probability of [p,1] and the negative sample number with the predicted probability of $[\mathrm{p},\mathrm{p}+\Delta\mathrm{p}]$. When $[\Delta\mathrm{p}]$ is small enough, it can be understood as the sample pair composed of the negative sample number of $[\mathrm{p},\mathrm{p}+\Delta\mathrm{p}]$ and the prediction probability, and the positive sample predicted value is higher than the sample pair composed of the negative sample predicted value. When the negative sample is divided into several segments according to probability, the positive and negative sample pairs formed by $\mathrm{P_p} * \mathrm{N_p}, \mathrm{N_{p,p+\Delta p}}$ are integrated, and all sample pairs whose positive sample predicted values are higher than the negative sample predicted values are obtained.

(4) Coverage

$$\mathrm{Coverage} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\max_{\mathrm{y}\in\mathrm{Yi}}\mathrm{rank_f}\left(\mathrm{x_{i,}y}\right) - 1\tag{5}$$

Where $\mathrm{rank_f}(\cdot,\cdot)$ is the sorting function corresponding to the real-valued function $\mathrm{f}(\cdot,\cdot)$. This index probes the search depth essential for encompassing all pertinent labels within the sample's category label sequence. Lower values indicate superior performance.

(5) Precision

$$\mathrm{Precision(Pre)} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\tag{6}$$

Precision is used to evaluate the accuracy rate of the detector based on the detection success. For all positive cases (TP + FP) judged by the model, the proportion of real cases (TP) among them.

(6) Recall

$$\mathrm{Recall(Rec)} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\tag{7}$$

Recall is used to evaluate the detection coverage of the detector on all targets to be detected. For all positive cases (TP + FN) in the dataset, the proportion of positive cases (TP) is correctly judged by the model to all positive cases in the dataset.

Table 3
Results of each metrics under 10-fold cross validation.

Metrics Fold	HammingLoss	RankingLoss	MicroF1	MacroF1	MacroAUC	MicroAUC	Average Precision	AverageRecall	Coverage
1	0.0115	0.0126	0.9842	0.9822	0.9858	0.9874	0.9853	0.9851	3.2764
2	0.0117	0.0119	0.9840	0.9817	0.9859	0.9869	0.9854	0.9845	3.2816
3	0.0107	0.0107	0.9853	0.9833	0.9868	0.9882	0.9875	0.9851	3.2534
4	0.0121	0.0116	0.9833	0.9811	0.9855	0.9865	0.9859	0.9828	3.2781
5	0.0110	0.0109	0.9848	0.9830	0.9867	0.9873	0.9883	0.9831	3.2666
6	0.0108	0.0119	0.9851	0.9828	0.9864	0.9880	0.9872	0.9849	3.2724
7	0.0115	0.0112	0.9842	0.9824	0.9860	0.9876	0.9873	0.9830	3.2660
8	0.0116	0.0132	0.9840	0.9819	0.9852	0.9875	0.9862	0.9839	3.2928
9	0.0108	0.0116	0.9852	0.9831	0.9865	0.9881	0.9876	0.9843	3.2819
10	0.0103	0.0118	0.9858	0.9833	0.9868	0.9880	0.9892	0.9837	3.2924
Average	0.0112	0.0117	0.9846	0.9825	0.9862	0.9876	0.9870	0.9840	3.2762
Standard deviation	0.0005	0.0007	0.0007	0.0007	0.0005	0.0005	0.0012	0.0008	0.0115

(7) F1-Score

$$F1 = 2 \bullet \frac{\text{Pre} \bullet \text{Rec}}{\text{Pre} + \text{Rec}} \tag{8}$$

True positive (TP) indicates normal events classified as normal, true negative (TN) indicates abnormal events deemed abnormal, false positive (FP) represents abnormal events misidentified as normal, and false negative (FN) represents normal events wrongly marked as abnormal.

To better demonstrate the test results, we counted the test results of each fold and average in the 10-fold cross-validation, as shown in Table 3.

At the same time, to show the test situation of each label, and to understand which labels are better classified and which labels are average, we took out each category of labels separately for analysis, in which the ROC curve of each category are shown in Fig. 8, and the confusion matrix of each category is shown in Fig. 9.

As can be seen from Figs. 8 and 9, when the effectiveness of each category of tag classification is evaluated separately, the two categories of tags with labels "1″ and "5″ are relatively weak among the 11 categories of tags, while the other categories of tags are relatively better.

In addition, we also conducted comparative experiments to compare the variation trends of each metric under different thresholds to reflect the robustness of the model, as shown in Table 4 and Fig. 10.

If all metrics are evaluated only at the threshold of 0.5, it is not enough to fully demonstrate the excellent performance of this model. Therefore, we calculate metrics at different thresholds. When the threshold increases from 0.5 to 0.8, all metrics tend to be basically stable, and the trend of performance decline is not obvious. When the threshold increases from 0.8 to 0.95, there is a relatively significant downward trend in all indicators. When the threshold value increases from 0.95 to 0.99, the indicators have a clear downward trend only. From the comparison experiment, it can be seen that the probability values of each label obtained from the model test are more accurate, which also reflects the scientific and robust nature of the model.

Finally, we compared the research contents and results with those of other researchers, as shown in Table 5. In general, more categories and labels mean a more difficult task. Taken together, with the number of categories and labels greater than or equal to those of previous studies, the method used in our study still achieves superior results.

6. Discussion

Our research demonstrates that the SAR model we developed, which integrates attention and ResNet mechanisms, surpasses previous studies in the accurate classification of multi-category and multi-label arrhythmia signals. This can assist cardiologists in making scientifically sound and automatic diagnoses of arrhythmia diseases. The model is capable of learning and predicting arrhythmic ECG signals across 11 categories and up to 5 co-existing labels in the dataset. The testing results were promising, with an average HammingLoss of 1.12 %, average RankingLoss of 1.17 %, average Micro F1-score of 98.46 %, average Micro AUC of 98.76 %, and average Coverage of 3.2762. These results outperform those of previous research, demonstrating the scientific validity and robustness of our model. The SAR model leverages the power of self-attention structures, ResNet Blocks, and other deep learning techniques to focus on and extract key features from the ECG data while ignoring less impactful features. This approach allows for efficient and effective arrhythmia classification.

Our research builds upon previous studies in arrhythmia detection and classification. For instance, Zhu et al. (2020) developed a CNN-based model for arrhythmia classification, achieving an AUC of 98.3 % and F1-score of 84.5 % [39]. Ge et al. (2021) proposed a Multi-Label Correlation guided feature fusion network for abnormal ECG diagnosis, achieving a precision of 81.6 % and F1-score of

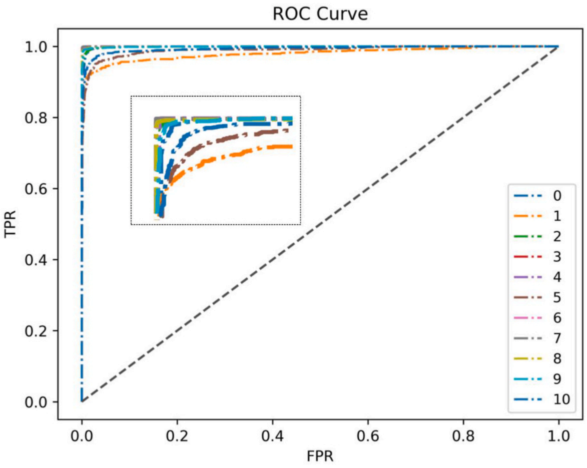


Fig. 8. ROC curve for each label.

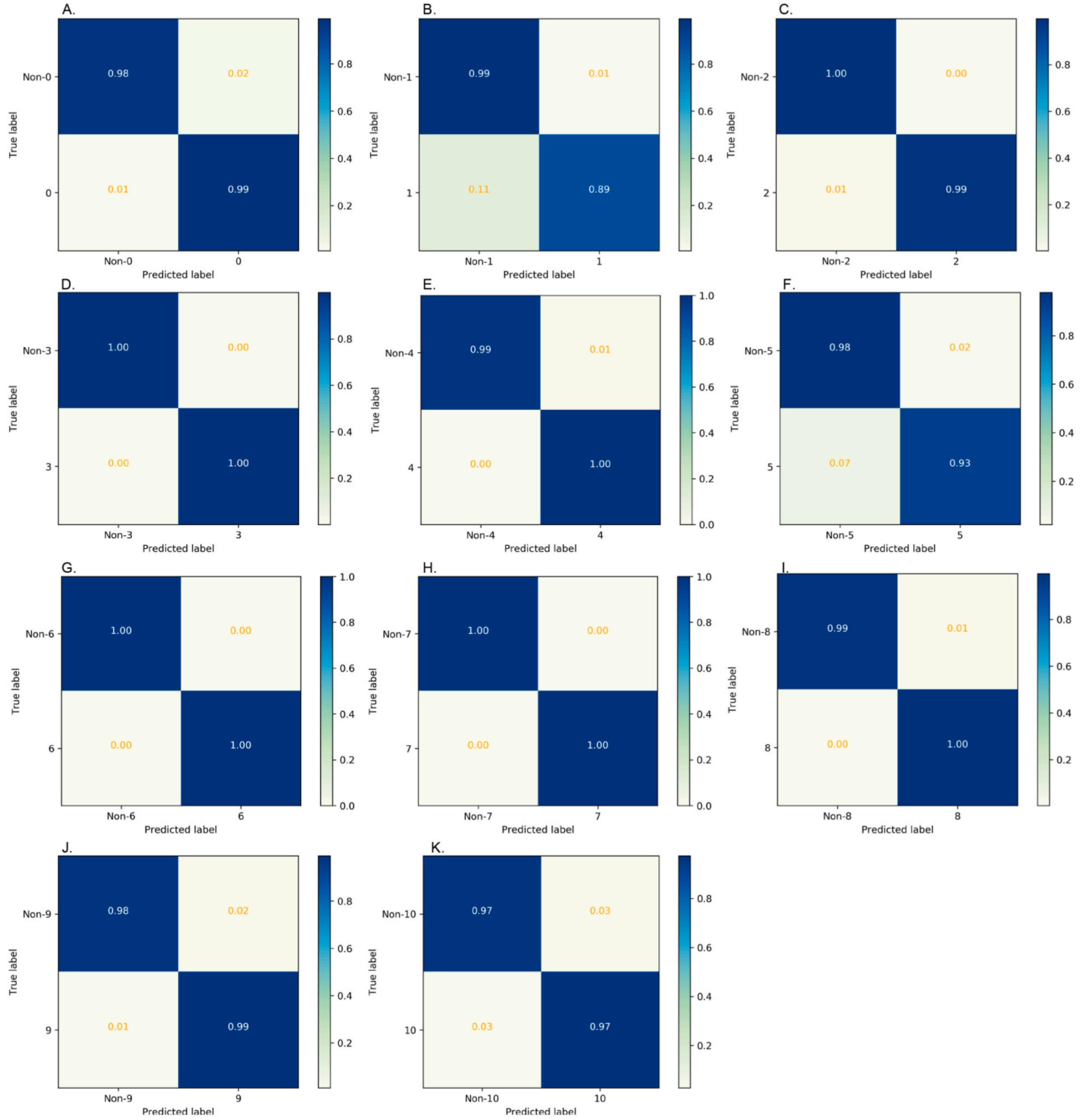


Fig. 9. Confusion matrix for each label. A–K: lable 0–10.

82.7 % [3]. Our model outperforms these previous models, suggesting that our approach is more effective for arrhythmia classification.

Our findings reveal certain limitations in the performance of the SAR model, particularly in identifying data with labels 1 and 5. The ROC curve and confusion matrix analysis indicate suboptimal results for these specific labels, prompting further investigation into the underlying causes, as shown in Figs. 8 and 9. To gain insights, we conducted a detailed analysis by comparing the features and labels of the data specifically associated with labels 1 and 5. Given these circumstances, we turned our attention to the possibility of incorrect labeling in the dataset. To address this concern, we plan to collaborate with ECG specialists by visiting hospitals and seeking their expert opinion. Their insights and expertise will help determine whether this labeling issue is indeed present and, if confirmed, allow us to rectify it accordingly.

The composition of the dataset, specifically the limitations imposed by the MIT-BIH database on the number of labels assigned to each sample, was also a factor in our next study. We included data samples with a maximum of 5 labels for training, validation, and testing. However, to validate the SAR model in a broader context, we plan to conduct experiments on a larger cohort and with real-

Table 4
Evaluation metrics under different thresholds.

Metrics Threshold	HammingLoss	RankingLoss	MicroF1	MacroF1	MacroAUC	MicroAUC	AveragePrecision	AverageRecall	Coverage
0.5	0.0112	0.0117	0.9846	0.9825	0.9862	0.9876	0.9870	0.9840	3.2762
0.6	0.0111	0.0118	0.9846	0.9827	0.9870	0.9872	0.9882	0.9829	3.2845
0.7	0.0111	0.0121	0.9846	0.9827	0.9874	0.9869	0.9893	0.9818	3.3159
0.8	0.0113	0.0123	0.9843	0.9825	0.9862	0.9864	0.9904	0.9802	3.3290
0.9	0.0119	0.0133	0.9834	0.9816	0.9853	0.9852	0.9918	0.9769	3.3601
0.95	0.0127	0.0143	0.9823	0.9805	0.9842	0.9839	0.9932	0.9735	3.3957
0.99	0.0162	0.0191	0.9772	0.9754	0.9791	0.9786	0.9954	0.9618	3.5345
Average	0.0122	0.0135	0.9830	0.9811	0.9851	0.9851	0.9908	0.9773	3.3566
Standard deviation	0.0017	0.0024	0.0025	0.0025	0.0026	0.0029	0.0027	0.0072	0.0822

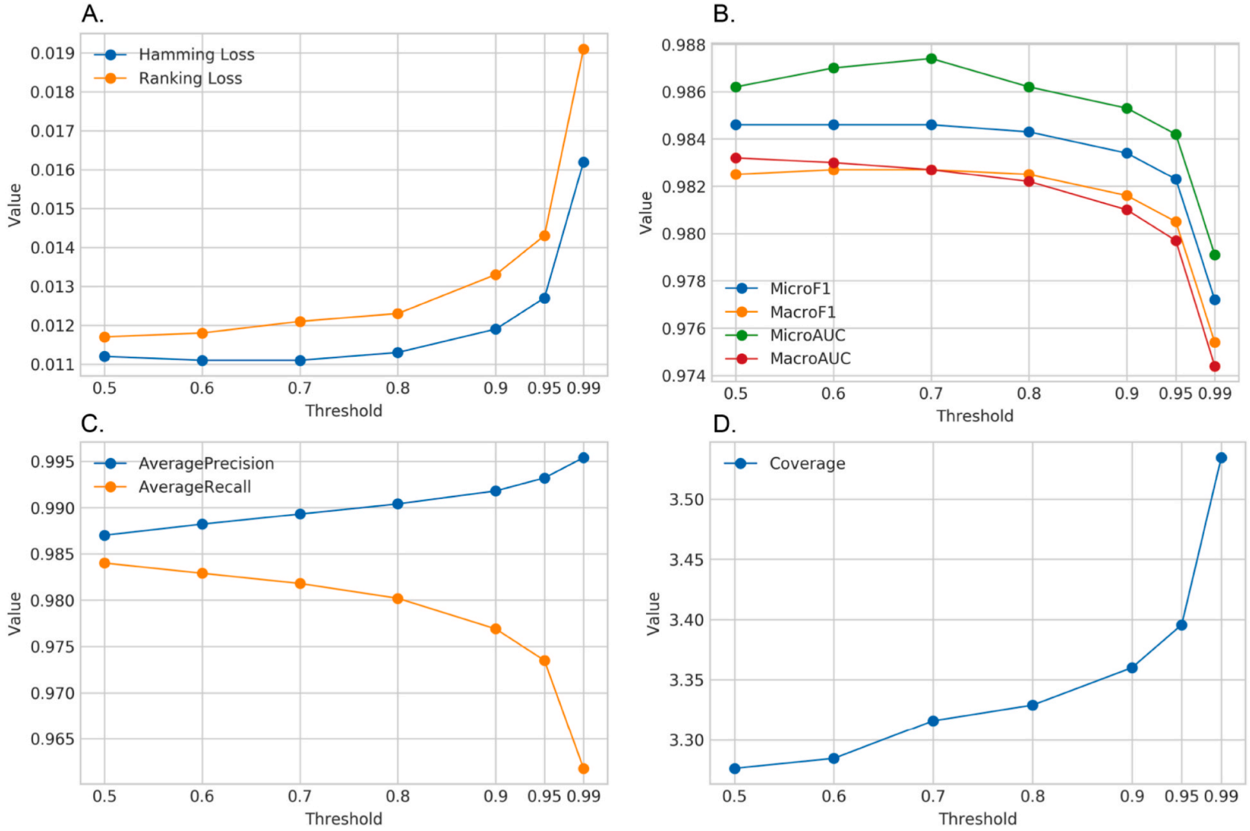


Fig. 10. Line chart of each evaluation metrics under different thresholds. A: Hamming and ranking loss; B: Micro F1 and AUC, Macro F1 and AUC; C: Average precision and recall; D: Coverage.

Table 5
Comparison with other researches.

Authors	Year	Categories and labels	Model	Results
Zhu et al. [39]	2020	8 categories 5 labels	CNN	AUC = 98.3 %, Sen = 86.7 %, Spe = 99.5 %, F1 = 84.5 %
Ge et al. [3]	2021	4 categories 4 labels	MLC-CNN	Pre = 81.6 %, Sen = 84.5 %, F1 = 82.7 %

Ours	**2022**	**11categories 5 labels**	**SAR**	**Hamming loss=1.12 %,
				Ranking Loss=1.17 %,
				MicroF1=98.46 %, MacroF1=98.25 %,
				MicroAUC=98.76 %, MacroAUC=98.62 %, AveragePrecision=98.70 %,
				AverageRecall=98.40 %, coverage=3.2762**

world data.

By validating the SAR model in a clinical setting and refining its performance based on real data, we aim to make a valuable contribution to the field, ultimately improving the accuracy and efficiency of arrhythmia diagnosis. Our future work will involve testing the model using actual hospital data and optimizing the model based on the test results. This will help us to further validate the effectiveness of our model and make practical contributions to clinical practice.

7. Conclusions

Automatic diagnosis of arrhythmia has important research value, which can reduce and effectively improve detection efficiency and reduce the workload of clinicians.

In this research, we innovatively built SAR, a model that can learn and predict arrhythmic ECG signals for 11 categories and up to 5 co-existing labels in the dataset, through testing, in which the average HammingLoss is 1.12 %, average RankingLoss is 1.17 %, average Micro F1-score is 98.46 %, average Micro AUC is 98.76 %, and average Coverage is 3.2762. The results are better than previous research results, which shows the scientific validity and robustness of the model. In the next step, we will continue to promote methodological research, test the model using actual hospital data, and optimize the model based on the test results, in an effort to make practical contributions to clinical practice.

Data availability statement

Data associated with this study has been deposited at MIT-BIH Arrhythmia. Database: https://www.physionet.org/content/mitdb/1.0.0/

Funding

This research was funded by.

1. The National Natural Science Foundation of China, grant number 71764035;
2. The National Natural Science Foundation of China, grant number 71864021;
3. Peking Union Medical College Group Medical Discipline Construction Project, NO: WH10022021145;
4. Guilin talent mini-highland scientific research project, (Municipal Committee Talent Office of Guilin City [2020] No. 3–05).

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Liuyang Yang: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yaqing Zheng:** Conceptualization, Data curation, Formal analysis, Writing – original draft. **Zhimin Liu:** Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **Rui Tang:** Conceptualization, Methodology, Supervision. **Libing Ma:** Conceptualization, Funding acquisition, Supervision. **Yu Chen:** Conceptualization, Funding acquisition, Supervision. **Ting Zhang:** Data curation, Funding acquisition, Resources, Writing – review & editing. **Wei Li:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Our deepest gratitude goes to the anonymous reviewers and editors for their careful review and comments and thoughtful suggestions that have helped improve this paper substantially.

References

[1] Y.R. Li, K. Li, Inter-patient arrhythmia classification with improved deep residual convolutional neural network, Comput. Methods Progr. Biomed. 214 (2022), 106582, https://doi.org/10.1016/j.cmpb.2021.106582.

[2] N. Jannah, S. Hadjiloucas, J. Al-Malki, Arrhythmia detection using multi-lead ECG spectra and complex support vector machine classifiers, Proc. Comput. Sci. 194 (2021) 69–79, https://doi.org/10.1016/j.procs.2021.10.060.

[3] Z. Ge, et al., Multi-label correlation guided feature fusion network for abnormal ECG diagnosis, Knowl. Base Syst. 233 (2021), 107508, https://doi.org/10.1016/j.procs.2021.10.060.

[4] K.J. Ch, et al., Arrhythmia detection based on multi-scale fusion of hybrid deep models from single lead ECG recordings: a multicenter dataset study, Biomed. Signal Process Control 77 (2022), 103753, https://doi.org/10.1016/j.bspc.2022.103753.

[5] A. Chen, et al., Multi-information fusion neural networks for arrhythmia automatic detection, Comput. Methods Progr. Biomed. 193 (2020), 105479, https://doi.org/10.1016/j.cmpb.2020.105479.

[6] K. Ma, C.A. Zhan, F. Yang, Multi-classification of arrhythmias using ResNet with CBAM on CWGAN-GP augmented ECG gramian angular summation field, Biomed. Signal Process Control 77 (2022), 103684, https://doi.org/10.1016/j.bspc.2022.103684.

[7] Z. Sun, C. Wang, Y. Zhao, et al., Multi-label ECG signal classification based on ensemble classifier, IEEE Access 8 (2020) 117986–117996, https://doi.org/10.1109/ACCESS.2020.3004908.

[8] H. Liu, Z. Zhao, Q. She, Self-supervised ECG pre-training, Biomed. Signal Process Control 70 (2021), 103010, https://doi.org/10.1016/j.bspc.2021.103010.

[9] B.-J. Singstad, E.M. Muten, P.H. Brekke, Multi-Label ECG Classification Using Convolutional Neural Networks in a Classifier Chain, 2021 Computing in Cardiology, CinC, 2021, pp. 1–4, https://doi.org/10.23919/CinC53138.2021.9.

[10] J. Cai, W. Sun, J. Guan, et al., Multi-ECGNet for ECG arrythmia multi-label classification, IEEE Access 8 (2020) 110848–110858, https://doi.org/10.1109/ACCESS.2020.3001284.

[11] J. Yoo, et al., K-labelsets method for multi-label ECG signal classification based on SE-ResNet, Appl. Sci. 11 (16) (2021) 7758, https://doi.org/10.3390/app11167758.

[12] P. Rong, T. Luo, J. Li, et al., Multi-label disease diagnosis based on unbalanced ECG data, in: 2020 IEEE 9th Data Driven Control and Learning Systems Conference, DDCLS, 2020, pp. 253–259, https://doi.org/10.1109/DDCLS49620.2020.9275099.

[13] B. Puszkarski, K. Hryniów, G. Sarwas, Comparison of neural basis expansion analysis for interpretable time series (N-BEATS) and recurrent neural networks for heart dysfunction classification, Physiol. Meas. 43 (6) (2022) 1–4, https://doi.org/10.1088/1361-6579/ac6e55.

[14] V.A. Ardeti, et al., An overview on state-of-the-art electrocardiogram signal processing methods: traditional to AI-based approaches, Heliyon 217 (2023), 119561, https://doi.org/10.1016/j.eswa.2023.119561.

[15] Luo, Yang, Cai, et al., Multi-classification of arrhythmias using a HCRNet on imbalanced ECG datasets, Comput. Methods Progr. Biomed. 208 (2021), 106258, https://doi.org/10.1016/j.cmpb.2021.106258.

[16] H.V. Denysyuk, et al., Algorithms for automated diagnosis of cardiovascular diseases based on ECG data: a comprehensive systematic review, Heliyon 9 (2) (2023), e13601, https://doi.org/10.1016/j.heliyon.2023.e13601.

[17] Z. Liu, et al., Accurate detection of arrhythmias on raw electrocardiogram images: an aggregation attention multi-label model for diagnostic assistance, Med. Eng. Phys. 114 (2023), 103964, https://doi.org/10.1016/j.medengphy.2023.103964.

[18] N. Du, et al., FM-ECG: a fine-grained multi-label framework for ECG image classification, Inf. Sci. 549 (2021) 164–177, https://doi.org/10.1016/j.ins.2020.10.014.

[19] G. Nalbantov, S. Ivanov, J. Van Prehn, Multi-Class Classification of Pathologies Found on Short ECG Signals, 2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.071.

[20] S. Yang, H. Xiang, Q. Kong, et al., Multi-label Classification of Electrocardiogram with Modified Residual Networks,2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.007.

[21] Y. Feng, E. Vigmond, Deep Multi-Label Multi-Instance Classification on 12-Lead ECG, 2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.095.

[22] N. Osnabrugge, et al., Multi-Label Classification on 12, 6, 4, 3 and 2 Lead Electrocardiography Signals Using Convolutional Recurrent Neural Networks,2021 Computing in Cardiology, CinC, 2021, pp. 1–4, https://doi.org/10.23919/CinC53138.2021.9662725.

[23] Z. Zhu, et al., Classification of Cardiac Abnormalities from ECG Signals Using SE-ResNet, 2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.281.

[24] D. Wang, et al., Automatic detection of arrhythmia based on multi-resolution representation of ECG signal, Sensors 20 (6) (2020) 1579, https://doi.org/10.3390/s20061579.

[25] A.W. Wong, A. Salimi, A. Hindle, et al., Multilabel 12-Lead Electrocardiogram Classification Using Beat to Sequence Autoencoders, ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 1270–1274, https://doi.org/10.1109/ICASSP39728.2021.9414934.

[26] R. Pardasani, N. Awasthi, Classification of 12 Lead ECG Signal Using 1D-Convolutional Neural Network with Class Dependent Threshold,2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.277.

[27] D. Borra, A. Andaló, S. Severi, et al., On the Application of Convolutional Neural Networks for 12-lead ECG Multi-Label Classification Using Datasets from Multiple Centers,2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.349.

[28] D. Li, et al., Automatic classification system of arrhythmias using 12-lead ECGs with a deep neural network based on an attention mechanism, Symmetry (Basel) 12 (1827) (2020) 1827, https://doi.org/10.3390/sym12111827.

[29] Y. Liu, et al., Multi-Label Classification of Multi-Lead ECG Based on Deep 1D Convolutional Neural Networks with Residual and Attention Mechanism,2021 Computing in Cardiology, CinC, 2021, pp. 1–4, https://doi.org/10.23919/CinC53138.2021.9662873.

[30] H. Wang, et al., A weighted graph attention network based method for multi-label classification of electrocardiogram abnormalities, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2020, pp. 418–421, https://doi.org/10.1109/EMBC44109.2020.9175981.

[31] Z. Li, H. Zhang, Automatic detection for multi-labeled cardiac arrhythmia based on frame blocking preprocessing and residual networks, Frontiers in Cardiovascular Medicine 8 (2021), https://doi.org/10.3389/fcvm.2021.616585.

[32] Y. Liu, et al., Multi-Label Classification of 12-lead ECGs by Using Residual CNN and Class-Wise Attention, 2020 Computing in Cardiology, 2020, pp. 1–4, https://doi.org/10.22489/CinC.2020.285.

[33] C. Li, et al., DeepECG: image-based electrocardiogram interpretation with deep convolutional neural networks, Biomed. Signal Process Control 69 (2021), 102824, https://doi.org/10.1016/j.bspc.2021.102824.

[34] V. Sangha, et al., Automated multilabel diagnosis on electrocardiographic images and signals, Nat. Commun. 13 (1) (2022), https://doi.org/10.1038/s41467-022-29153-3.

[35] Moody Gb, Mark Rg, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. 20 (3) (2001) 45–50, https://doi.org/10.1109/51.932724 (PMID: 11446209).

[36] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation [Online] 101 (23) (2000), e215–e220, https://doi.org/10.1161/01.CIR.101.23.e215.

[37] N. Jannah, S. Hadjiloucas, J. Al-Malki, Arrhythmia detection using multi-lead ECG spectra and complex support vector machine classifiers, Proc. Comput. Sci. 194 (2021) 69–79, https://doi.org/10.1016/j.procs.2021.10.060.

[38] L. Song, et al., PreCar_Deep：A deep learning framework for prediction of protein carbonylation sites based on Borderline-SMOTE strategy, Chemometr. Intell. Lab. Syst. 218 (2021), 104428, https://doi.org/10.1016/j.chemolab.2021.104428.

[39] H. Zhu, et al., Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study, The Lancet Digital Health 2 (7) (2020) e348–e357, https://doi.org/10.1016/S2589-7500(20)30107-2.