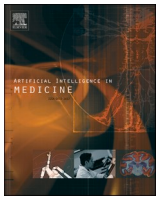
Artificial Intelligence In Medicine 132 (2022) 102379

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Research paper

Automatic diagnosis of arrhythmia with electrocardiogram using multiple instance learning: From rhythm annotation to heartbeat prediction

Xuan Zhang [a,1], Hui Wu [b,1], Ting Chen [a,*], Guangyu Wang [c,*]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[b] Department of Cardiology, Yichang central people's hospital, China three Gorges University, Yichang 443003, China
[c] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Electrocardiogram
Multiple instance learning
Deep learning
Heartbeat classification

ABSTRACT

The electrocardiogram (ECG) is a commonly used technique for detecting arrhythmias and many other cardiac diseases. Automatic ECG diagnosis has seen tremendous success in recent years, owing to the rapid development of the deep learning (DL) approach. Existing works on automatic ECG diagnosis can be divided roughly into two categories: prediction at the rhythm level from an ECG record, and prediction at the heartbeat level, although their relationship was seldom studied previously. In this paper, we address the following question: can we train an abnormal heartbeat detection model using solely data annotated at the rhythm level? We first used multiple instance learning (MIL) to model the relationship between an ECG record (whose label is given at the rhythm level and is provided as an input) and the heartbeats in the ECG (whose labels are to be predicted). Then, we sequentially trained two models, a rhythm model for detecting abnormal heartbeats in an ECG record labeled as arrhythmia, and a heartbeat model for classifying heartbeats as normal or various types of arrhythmias. We trained and tested our models using 61,853 ECG records with rhythm annotations. The experimental results demonstrate that the heartbeat model achieves a macro-average F1 score of 0.807 in classifying four types of arrhythmias as well as normal heartbeats. Our model significantly outperforms the model directly trained with 15,385 ECG heartbeats with heartbeat annotations, demonstrating the viability of our strategy for training a high-performing heartbeat-level automatic diagnostic model using only rhythm annotation.

1. Introduction

The term arrhythmia refers to irregular heartbeats [1]. Electrocardiograms (ECGs) are widely used in clinical practice to diagnose arrhythmias. Each heartbeat produces a period of the ECG waveform, which consists mainly of a P wave, a QRS complex, and a T wave [2]. Arrhythmia is typically associated with either an irregular heart rate or certain morphologic alterations that can be observed in specific heartbeats on an ECG (e.g., a heartbeat with a larger QRS complex and without P wave can be diagnosed as premature ventricular contraction [3]). Since arrhythmia associated with an irregular heart rate could be easily diagnosed using R-R intervals, this paper focuses on the more challenging diagnosis of arrhythmia associated with morphologic alterations. A heartbeat-level diagnosis would provide better interpretability in clinical applications. Nevertheless, cardiologists are unable to execute this duty due to the time commitment: a standard 20-second ECG check comprises an average of 20–30 heartbeats, whereas a 24-hour ambulatory ECG may contain approximately 100,000 heartbeats that must be examined. Thus, it is critical to developing an automated method for diagnosing arrhythmias at the heartbeat level.

Recent advances in machine learning, especially in deep learning, have resulted in great success in a variety of computer-aided diagnosis applications [4–6]. One of the most attractive aspects of deep learning is that it does not require manually developed features for specific inputs and tasks, which is critical in the medical field where defining a precise process for feature extraction is typically challenging. The primary disadvantage of deep learning is that it requires a large amount of annotated data to perform well [7]. To train a deep learning model for detecting arrhythmias of heartbeats, we must first collect a large number of ECG recordings labeled at the heartbeat level [8]. However, this

* Corresponding authors.
E-mail addresses: x-zhang18@mails.tsinghua.edu.cn (X. Zhang), wuhui@ctgu.edu.cn (H. Wu), tingchen@tsinghua.edu.cn (T. Chen), guangyu.wang@bupt.edu.cn (G. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.artmed.2022.102379
Received 12 November 2021; Received in revised form 30 June 2022; Accepted 18 August 2022
Available online 22 August 2022

annotation requirement is difficult to meet in practice due to its time and human resource requirement. On the other hand, a rhythm diagnosis of an ECG record is relatively easier to obtain, for example, in a physical examination center that performs routine ECG checks (with permission). Fig. 1 illustrates the labels at the heartbeat and rhythm levels. Generally, if each heartbeat can be accurately diagnosed, the rhythm label can be inferred easily. However, the converse is much more challenging. The purpose of this study is to build a heartbeat prediction model using the ECG records with rhythm annotations.

This study proposed a novel strategy for developing an arrhythmia diagnosis model for heartbeats using solely rhythm-level data annotations. To begin, we define the computational problem using the framework of multiple instance learning (MIL). MIL aims to classify a bag (rhythm) consisting of a number of instances (heartbeats), with only bag annotations given during training [9]. To solve this problem, we propose a two-stage algorithm illustrated in Fig. 2. First, we train an Attention U-net [10] using ECG records and associated rhythm annotations. This model calculates a saliency map based on the attention weights for heartbeats in an ECG record labeled as arrhythmia, which indicates the locations of abnormal heartbeats. Using the saliency maps, we then construct a pseudo-heartbeat-level training set to train a 6-layer convolutional neural network as a heartbeat arrhythmia classifier. To assess our approach in a real-world setting, we used 68,716 ECG records obtained from Yichang Central People's Hospital in Hubei Province, China. These records were from neonate patients. The experimental results show that our strategy is viable for establishing a high-performing heartbeat automated diagnosis model with rhythm annotations.

This study makes the following contributions: (1) To our best knowledge, few have attempted to train a heartbeat arrhythmia detector using only rhythm annotations. Thus, the setting of the problem is novel. (2) To address this issue, we proposed a novel two-stage algorithm based on the MIL framework. (3) Our algorithm was validated using real-world data. As demonstrated by experimental results, our method is effective.

The remainder of the paper is organized as follows. Section 2 introduces the related works. Section 3 elaborates on our proposed method. Section 4 describes our dataset, experimental design, and results. Section 5 discusses the value and weakness of our work. Finally, Section 6 concludes this study.

2. Related works

2.1. Deep learning in ECG analysis

Numerous previous works have investigated methods for automatic arrhythmia diagnosis using the raw ECG signal and deep learning technique. As noted in the abstract, these studies fall into two broad categories: predictions at the heartbeat level and predictions at the rhythm level. For predictions at the heartbeat level, suggested architectures include 1-dimensional convolutional neural networks [8,11,12], 2-dimensional convolutional neural networks [13–15], recurrent neural networks [16–18], restricted Boltzmann machines [19,20], and a combination of the aforementioned architectures [21,22]. In general, the models designed for heartbeat classification have a simple architecture with a small number of parameters. Each of these works employed the same public dataset for training and testing, the MIT-BIH arrhythmia database [23], which contains only 48 recordings from 47 patients. The absence of a heartbeat-level annotated database confirms what we stated in the introduction: acquiring such annotations is difficult.

Recent studies have proposed a number of rhythm arrhythmia diagnosis models for longer ECG signals (usually >10 s), owing to the powerful representation capabilities of deep neural networks and the growing volume of digitally stored ECG records. In comparison to heartbeat classification models, rhythm classification models usually adopt a more complex architecture, such as a deep residual network [7,24], a squeeze-and-excitation residual network [25,26], a combined CNN and RNN [27,28], and an attention mechanism [29,30]. These models were trained on a large number of ECG records annotated at the rhythm level, and were used to classify ECGs at the same rhythm level. It is worth mentioning that Hannun et al. [7] developed a deep residual network to detect arrhythmia every 1.28 s in 30-second ECGs. However, the model was trained on 91,232 ECGs with approximately heartbeat-level annotations, which came at a very high cost. To our best knowledge, few works have attempted to create a heartbeat-level prediction model using rhythm-level annotated ECGs.

2.2. Multiple instance learning in ECG

We briefly introduced MIL in the introduction section. The basic assumption of MIL is that a bag belongs to a given class if and only if it includes at least one instance of that class [9]. For arrhythmia diagnosis, an ECG record (a bag) is diagnosed as having a certain type of arrhythmia (rhythm-level) if and only if at least one of the heartbeats in the ECG (one instance) is diagnosed as having that particular type of arrhythmia (heartbeat-level). Therefore, it is natural to adopt MIL framework to the ECG diagnosis problem. In contrast to the classic MIL, the final goal of our task is to classify every instance rather than the bag of instances [31].

There are two main MIL approaches [32]: (a) the instance-classification approach, which first classifies each instance independently and then integrates multiple instances into a bag-level prediction; and (b) the instance-embedding approach, which embeds each instance into a vector in a low-dimension space, aggregates the embedded vectors into a single vector, and finally uses a bag-level classifier on the integrated vector.

A few previous works have adopted the MIL technique to the analysis of ECGs. The concept of a 'bag' corresponds to an ECG record, and the concept of an 'instance' corresponds to a heartbeat in an ECG [33,34] or a short segment (not necessary a full heartbeat) in an ECG [35]. All of these works employed the previously described instance-classification approach, and integrated instance-level classification results using max-pooling [35], fixed-percentage-pooling [34], or topic-model-based

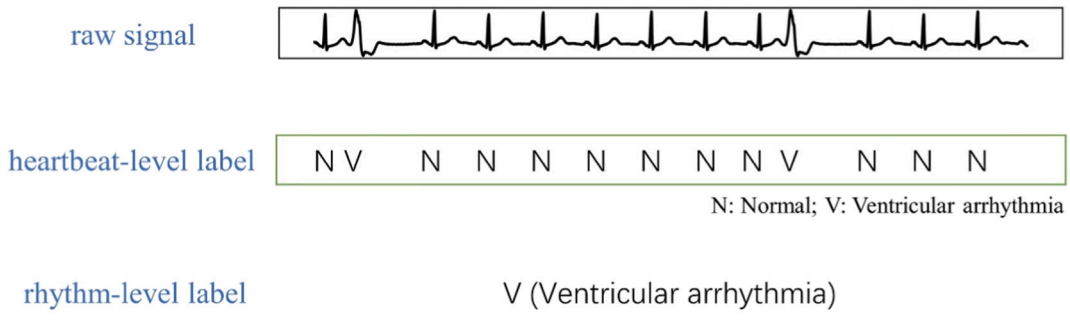


Fig. 1. The illustration of the concept 'labels at the heartbeat level' and 'labels at the rhythm level'. Obviously, obtaining the rhythm-level label is easier.

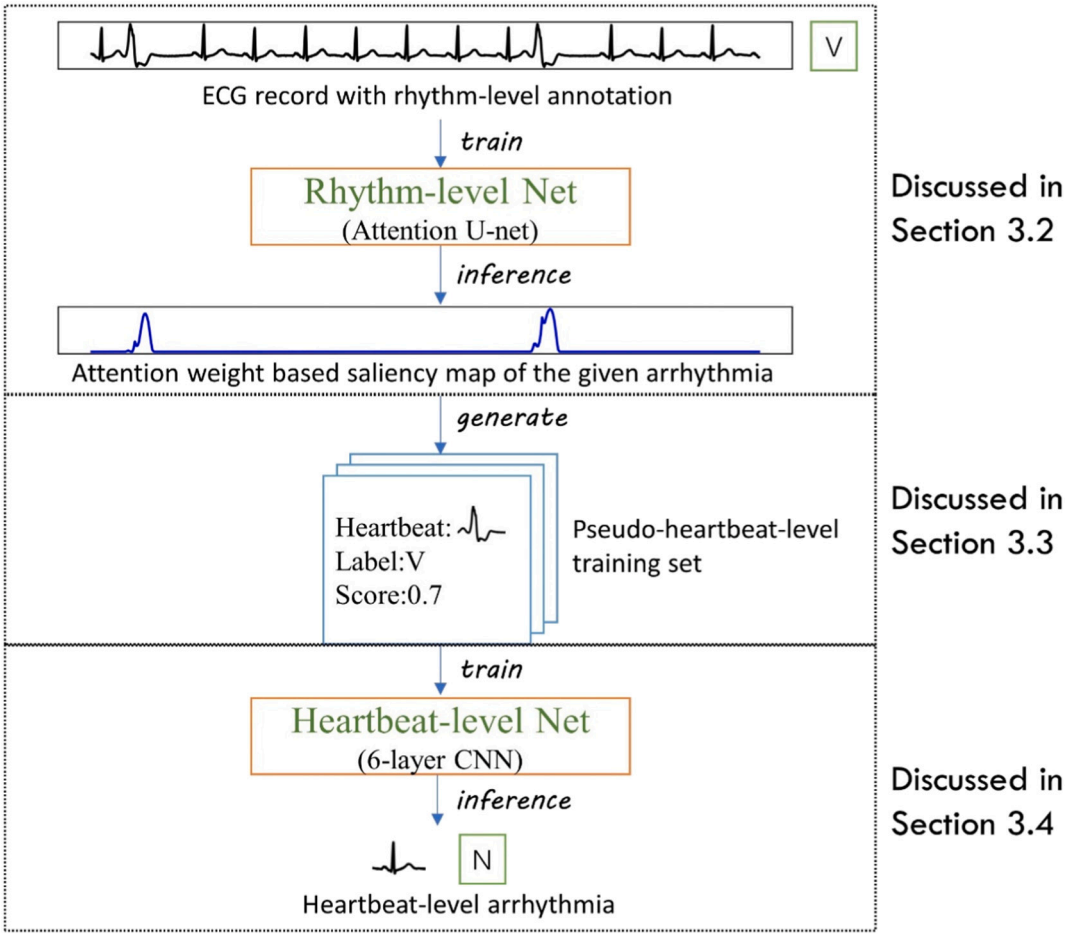


Fig. 2. The workflow of our proposed method.

pooling [33]. The purpose of these works is to detect abnormal heartbeats or ECG segments. However, they did not utilize these results to train a heartbeat classifier.

Our work follows the instance-embedding approach, which generally performs better than the instance-classification approach [32]. Specifically, we follow the attention-based MIL [36], which used the attention weights to gain instance-level interpretability. In our previous work, we presented an architecture named Attention U-net [10] for weakly-supervised localization on chest X-ray images. Additionally, we discovered that the Attention U-net is effective on one-dimensional ECG signals. As a result, we will continue to use this architecture for this task. The details will be discussed in the method section.

3. Method

3.1. Problem formulation

The goal of our research is to construct a heartbeat arrhythmia classifier using just ECG records annotated at rhythm-level. The following four types of arrhythmias are considered in this work: supraventricular arrhythmia, ventricular arrhythmia, right bundle branch block, and left bundle branch block.

Let $T = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be the training set, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ is the i-th ECG record consisting of m heartbeats. For each heartbeat x_{ij}, a hidden label $z_{ij} \in \{0, 1, 2, 3, 4\}$ describes whether the heartbeat is normal ($z_{ij} = 0$) or exhibits one of the arrhythmias ($z_{ij} = 1$, 2, 3,or 4). The given label $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}) \in \{0, 1\}^4$ is a four-dimension binary vector, and $Y_{ik} = 1 (k = 1, 2, 3, 4)$ if and only if $z_{ij} = k$ for some $j \in \{1, 2, \ldots, m\}$. In other words, diagnosing an ECG record with a certain type of arrhythmia implies that at least one of the heartbeats in the ECG exhibits that arrhythmia, which is consistent with the MIL assumption. The goal is to train a model to take an input heartbeat x and correctly predict an output $z \in \{0, 1, 2, 3, 4\}$. It should be noted that classifying an ECG record is a four-class multi-label classification task: for each of the four types of arrhythmias, determine whether the ECG has it (=1) or not (=0), because an ECG record may be diagnosed with two or more types of arrhythmias. Classifying a heartbeat, on the other hand, is a five-class multi-class classification task: determine whether it is normal ($z = 0$) or belongs to one of the four types of the arrhythmias ($z = 1$, 2, 3, 4). The main challenge is that the heartbeat label z is hidden in the training set.

3.2. Rhythm classifier

Fig. 3 describes the network architecture adopted in this study, Attention U-net [10]. It should be noted that in Fig. 3, each convolutional layer is followed by a batch normalization layer [37]. At the training step, ECGs and associated rhythm annotations are used as training data. At the inference step, the model can infer not only the arrhythmia category for a given input ECG, but also a saliency map based on the attention weights of the category indicating the locations of abnormal heartbeats. If an ECG is associated with multiple arrhythmias, the model will output a separate saliency map for each type of arrhythmia. The saliency maps will be used to develop a heartbeat-level diagnosis in our pipeline. We will explain the details in the following.

3.2.1. Inputs

Following the MIL framework, a straightforward approach is to first

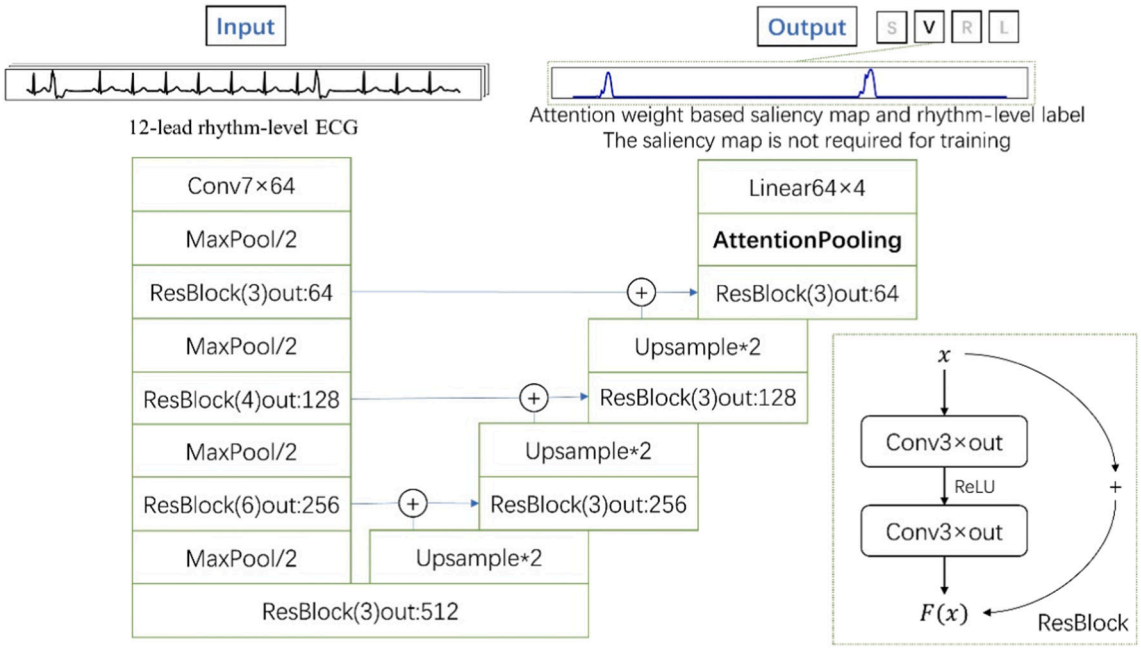


Fig. 3. The architecture of the Attention U-net. The '⊕' means concatenation of the channels together. The 'AttentionPooling' is the attention pooling layer described in the paper.

segment the ECG into multiple heartbeats, then extract heartbeat features using neural networks, and finally aggregate the features for final classification [33,34]. However, some global information, e.g., average heart rate and heart rate variability (HRV), may be lost in this approach. Therefore, our rhythm classifier accepts the whole ECG signal as input rather than individual segmented heartbeats. Besides, we consider 12-lead ECGs in this study where the 12-lead ECG signal is modeled as a one-dimensional time series with 12 channels.

3.2.2. Loss function

This is a multi-label classification task due to the possibility that an ECG may contain multiple heartbeats diagnosed with multiple disease categories. To address this problem, our model divides the task into multiple independent binary classification sub-tasks. Considering each sample with the true label $\boldsymbol{y} = (y_1, y_2, y_3, y_4)$ for four types of arrhythmias and the predicted probability $\boldsymbol{p} = (p_1, p_2, p_3, p_4)$, we adopt the cross-entropy loss function which is defined as:

$$l = -\sum_i (y_i log(p_i) + (1 - y_i) log(1 - p_i))$$

3.2.3. U-structure

The U-structure, inspired by the work in [38], is a classical architecture in the field of medical image semantic segmentation. It consists of three major components, an encoder, a decoder, and horizontal connections between the encoder and the decoder. The encoder is a convolutional neural network (in our architecture, the ResNet34 [39]) that encodes an input ECG into low-dimensional features. The decoder is a convolutional neural network with up-sampling layers that decode the low-dimensional features to high-dimensional representations for precise localization. The horizontal connections are special operations that concatenate feature maps of the encoder to the decoder at the same level, hence improving the accuracy of feature localization. We have shown [10] that the U-structure could not only slightly increase the classification performance, but also greatly increase the localization performance.

3.2.4. Instance

To make a connection between the ECG diagnosis problem and MIL, we previously referred to the 'heartbeat' as the 'instance'. However, the precise definition of the instance in our method is more abstract. We defined an instance as the 'a sequence of consecutive data points consisting of center data points and their neighborhood' in the origin signal. To illustrate the concept, we present a toy CNN model illustrated in Fig. 4, which consists of a 1-d CNN with two convolutional layers of kernel size 3 and one pooling layer of stride 2. The right panel of Fig. 4 illustrates the definition of an instance. The left panel of Fig. 4 illustrates two overlapping instances where (1) their center data points do not overlap and (2) the size of the neighborhood depends on the network architecture. In the following, we choose the center data points to represent an instance.

After the U-structure in our model, we obtain a 64-dimensional feature map of length L, denoted by $A^{64\times L} = (h_1, h_2, \ldots, h_L)$, where L is the number of instances, equal to half of the input length, and h_i is a 64-dimensional column vector. The h_1 vector, corresponding to the first instance, can be understood as the extracted features from the first two data points and their neighborhood in the original ECG signal. Similarly, the rest of the $L-1$ vectors are defined. Therefore, we have obtained the feature representations for all L instances.

Compared to the definition of 'heartbeat', our defined instance has the following benefits. First, the instances are fine-grained, permitting more precise interpretation at the instance-level. Second, an instance is not limited to a single heartbeat, but can include features from the context data points.

3.2.5. Attention pooling

In Section 3.2.4, we have obtained the feature representations of all L instances $A^{64\times L} = (h_1, h_2, \ldots, h_L)$. The issue is how to aggregate them into a single feature representation.

The attention pooling [36] suggests to use a weighted-sum form as

$$\boldsymbol{h}_c = \frac{1}{L}\sum_{i=1}^{L} w_{ci}\boldsymbol{h}_i,$$

where $c \in \{1, 2, 3, 4\}$ indicates the four classes of arrhythmia and w_{ci} is the weight of instance $\boldsymbol{h}_i$ and class c. The w_{ci} is computed automatically by the attention mechanism described in the following. For a given class c, we first calculate a score t_{ci} for each $\boldsymbol{h}_i$ such that

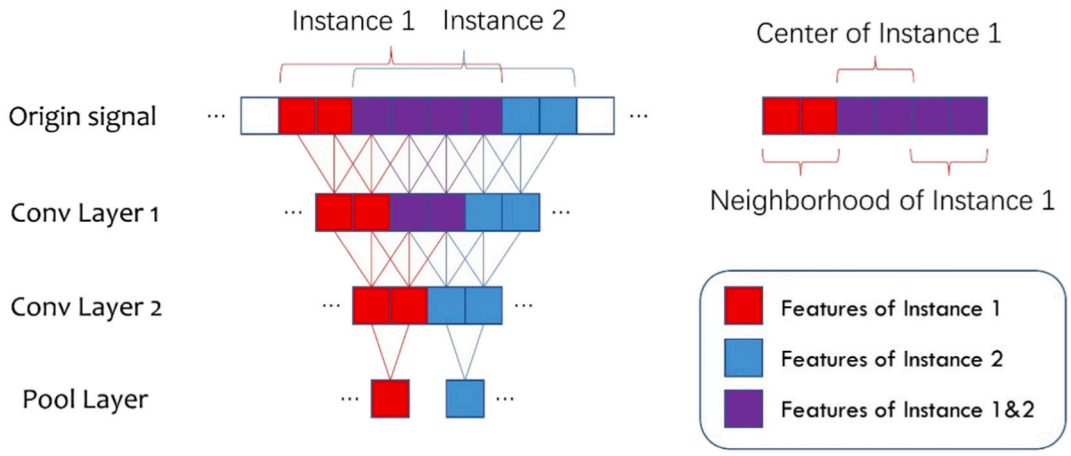


Fig. 4. A toy 1-d CNN model with two overlapping instances as input. Each instance consists of two central data points and four neighborhood data points.

$$t_{ci} = \boldsymbol{W}_{2c} tanh(\boldsymbol{W}_{1c}\boldsymbol{h}_i),$$

where $\boldsymbol{W}_{1c} \in \mathbb{R}^{64\times32}$, $\boldsymbol{W}_{2c} \in \mathbb{R}^{32\times1}$ are trainable parameters. We then use a softmax function to normalize t_{ci} to w_{ci} where $w_{ci} = \exp(t_{ci})/\sum_j \exp(t_{cj})$.

After the attention pooling, a linear layer $\boldsymbol{v_c}$ is used to map $\boldsymbol{h}_c$ to a real number, which is then converted to the final predicted probability p_c using a sigmoid function

$$p_c = (1 + exp(-\boldsymbol{v}_c \bullet \boldsymbol{h}_c))^{-1}$$

3.2.6. Saliency map based on the attention weight

Although the attention weights generated previously can be used directly as the saliency map, we combined our model with the class activation mapping (CAM) technique [40], which is widely used in CNN-based models for weakly-supervised object localization. To generate the saliency map for category c using the CAM-combined approach, we consider a 64-dimenstional parameter vector in the linear layer $\boldsymbol{v_c} = (v_{c1}, v_{c2}, \ldots, v_{c64})$ and the weighted feature maps $\boldsymbol{B}_c^{64\times L} = (w_{c1}\boldsymbol{h}_1, w_{c2}\boldsymbol{h}_2, \ldots, w_{cL}\boldsymbol{h}_L) = (\boldsymbol{b}_{c1}, \boldsymbol{b}_{c2}, \ldots, \boldsymbol{b}_{c64})^T$, where $\boldsymbol{b}_{cj}$ is the j-th feature map. The saliency map is defined as

$$\boldsymbol{L}_c = \mathrm{ReLU}(v_{c1}\boldsymbol{b}_{c1} + v_{c2}\boldsymbol{b}_{c2} + \ldots + v_{c64}\boldsymbol{b}_{c64}),$$

where the rectified linear unit function (ReLU) is defined as $\mathrm{ReLU}(x) = x\mathbf{1}_{[0,+\infty)}(x)$. The length of the saliency map is equal to L. If necessary, an upsampling algorithm may be applied to improve the fit of the input ECG.

3.3. Heartbeat scores

Given the rhythm labels for the training set, we can use the Attention U-net to obtain the saliency map $\boldsymbol{L}_c$ for each ECG record with arrhythmia. Due to the use of the ReLU function, all the values of the saliency map are greater than or equal to 0. We assume that some saliency map values are larger than 0 (otherwise, it implies that the model predicts with low confidence, and the sample will be dropped). We start by normalizing the saliency map, ensuring that all the values are summed to one. The next preprocessing step is to segment the ECG record into heartbeats. We identify all R-peaks using the approach given by Ivaylo et al. [41]. The starting point for the i-th heartbeat with R-peak location r_i is then $r_{i-1} + (r_i - r_{i-1}) \times 2/3$ while the ending point is $r_i + (r_{i+1} - r_i) \times 2/3$. It should be noted that the first and last heartbeats are dropped since they are usually incomplete.

After these preparations, we can calculate the sum of saliency map weights for all the heartbeats, excluding the first and last. The weight density of the i-th heartbeat with respect to class c, d_{ci}, is defined as the ratio of its weight sum over its length of signal, which is equal to the ending point location minus the starting point location. Intuitively, a heartbeat with a higher weight density is more likely to be abnormal. However, relying solely on the density, it is difficult to determine which heartbeats are abnormal (examples are provided in the Experiments section). Hence, we proposed a soft strategy in which we consider all heartbeats in an ECG labeled with arrhythmia class c to be arrhythmia class c but use *score* to quantify the confidence of the label. Let $D_c = (d_{c1}, d_{c2}, \ldots, d_{cm})$ be the arrhythmia class c's density for m heartbeats in an ECG, and $s_{ci} = d_{ci}/\underset{j}{max} d_{cj}$ be the *score* for the i-th heartbeat. The score has the following properties: (1) it is between 0 and 1, and (2) at least one heartbeat in each ECG record has a score of 1. We will use the score to train a heartbeat classifier.

3.4. Heartbeat classifier

As discussed in Section 3.1, the primary challenge arises from the fact that the hidden label for each heartbeat in the training set is unknown. Now, we use all the data to construct a pseudo-heartbeat-level training set. Each element in the training set includes a heartbeat waveform, a pseudo label, and a score. There are two kinds of ECG records. (1) For ECG records with arrhythmia, each heartbeat is assigned a score (described in Section 3.3). We incorporate all of them into the training set. It should be noted that if an ECG record is labeled with more than one types of arrhythmias, such as two (e.g., both supraventricular and ventricular arrhythmias are observed), the heartbeats will be used twice in the training set with distinct pseudo labels and scores. (2) For normal ECG records, all heartbeats should be normal. Therefore, we could add them directly to the training set with a score of 1. Since the overall number of normal heartbeats is much larger than that of abnormal heartbeats, we sample a subset of normal heartbeats to improve the computational efficiency.

We constructed a basic six-layer CNN for heartbeat arrhythmia classification, as illustrated in Fig. 5. In the early stage of the experiment, we also implemented two models [12,15] for heartbeat classification and found that our designed architecture outperformed both. Each convolutional layer in Fig. 5 is followed by a layer-normalization layer [42], which performed better in heartbeat network than the batch-normalization layer.

The objective is to classify a heartbeat into one of the five categories: normal, supraventricular arrhythmia, ventricular arrhythmia, right bundle branch block, and left bundle branch block. Let $(p_0, p_1, p_2, p_3, p_4)$ denote the model's output probability vector corresponding to the five classes above, and $p_0 + \ldots + p_4 = 1$. The cross-entropy loss function for a sample is

$$l_0 = -log(p_i),$$

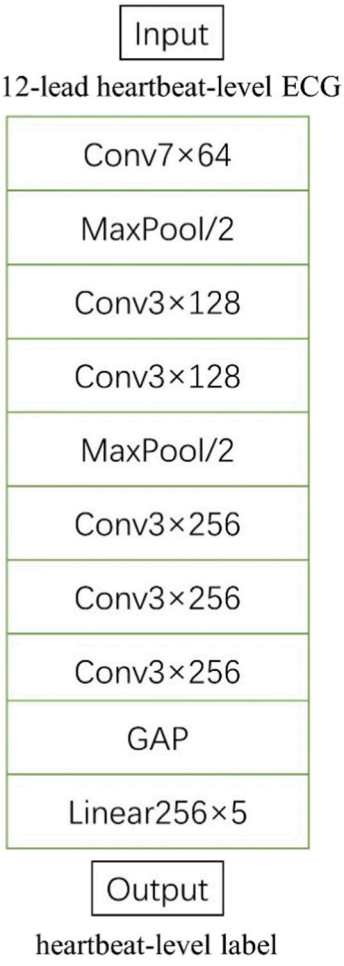


Fig. 5. The architecture of the heartbeat network. The 'GAP' stands for the global average pooling layer.

where $i \in \{0,1,2,3,4\}$ is the ground truth label. However, we cannot guarantee the accuracy of the labels in our training set. Therefore, we improved the loss function to

$$l = -(score)^{\alpha} log(p_i),$$

where α is a hyperparameter greater than or equal to zero. The added exponential term is inspired by the work in [43,44]. The hyperparameter α balances the opportunity to learn from new data against the risk of learning from noisy data. When α is set to zero, the loss function reduces to the cross-entropy, l_0, so the model treats all heartbeats extracted from a certain type of arrhythmia record as that arrhythmia. When α is set to infinity, only samples with a score of 1 are included in the training process, so only the heartbeat with the highest density from an arrhythmia ECG record is used for training. We will discuss the impact of α in the Section 4.3.5.

4. Experiments

4.1. The dataset

ECG records were obtained from Yichang Central People's Hospital in Hubei Province, China. Between January 2016 and June 2018, a total of 68,716 ECG records of 24 s in length were acquired and diagnosed. The records are from neonatal patients. Although many reference variables in the ECG are different between neonates and adults (e.g., heartrate, electrical axis, QRS duration), the diagnostic principles for neonatal and adult ECGs are similar. Therefore, the experimental results should be indicative of how well our approach works on other ECG datasets. Each ECG was obtained from a distinct outpatient and was recorded at a frequency of 1000 Hz. Each record is diagnosed as either normal or having at least one of the four types of arrhythmias. We randomly selected 90 % of the data (61,853 records) as the training set. The remaining 6863 records were used as the validation set of the Attention U-net. Additionally, an experienced cardiologist annotated 633 records with arrhythmia in the validation set on a heartbeat-by-heartbeat basis. These annotated heartbeats serve as the validation and testing sets of the final model: on the patient-level, 63 records (1623 heartbeats) were randomly selected for validation, and the remaining 570 records (15,385 heartbeats) were used for testing. Table 1 summarizes the dataset's detailed statistics. Specifically, supraventricular arrhythmia includes atrial premature beat, junctional premature beat, atrial escape beat, junctional escape beat, and atrial fibrillation. Ventricular arrhythmia includes premature ventricular contraction and ventricular escape beat. Left bundle branch block also includes left anterior fascicular block and left posterior fascicular block.

Although there are publicly assessable ECG databases with heartbeat annotations, we chose not to use them mainly due to the small size of the records. For example, the MIT-BIH arrhythmia database [23] contains only 48 records from 47 patients. Although each record contains a large number of heartbeats, it is rather difficult for the model to learn appropriate features from such a small sample size. Even if we could train a model from this dataset, it may not be very generalizable. In comparison, we have obtained a much larger set of ECG records to develop and validate our method.

4.2. Experimental setup

The origin ECG signals were sampled at the frequency of 1000 Hz. To save the computational cost, we down-sampled the signals to 200 Hz to reduce the data size while preserving critical information for arrhythmia diagnosis. Except for the down-sampling operation, we did not perform any other preprocessing on the raw data.

For the Attention U-net, we applied the initialization method proposed by He et al. [45] to initialize the parameters. The batch size was set to 64, and the learning rate was set to 0.0003. We did not apply any L2-norm regularization since we have already used normalization layers [46]. The Adam optimizer [47] was used to update the parameters at the training, with the default hyperparameter setting ($\beta_1 = 0.9, \beta_2 = 0.999$).

For the heartbeat network, we applied the initialization method proposed by He et al. [45] to initialize the parameters. The batch size was set to 128, and the learning rate was set to 0.0003. Similar to the Attention U-net, no L2-norm regularization was used. We evaluated the performance on the validation set and applied early stopping, which is also an efficient approach to preventing overfitting. We sampled only one heartbeat from each normal record in the origin training set to establish a balance between normal and abnormal heartbeats. Since the normal heartbeats within one record are morphologically similar, this sampling strategy could preserve the data diversity of normal

Table 1
The statistics of the dataset.

	Training (rhythm)	Validation (rhythm)	Validation (heartbeat)	Testing (heartbeat)
Normal (N)	56,031	6250	622	7349
Supraventricular Arrhythmia (S)	1748	198	114	847
Ventricular Arrhythmia (V)	1505	160	29	543
Right Bundle Branch Block (R)	2177	218	597	4785
Left Bundle Branch Block (L)	744	83	261	1861
Total	61,853	6863	1623	15,385

heartbeats. Besides, class weights are introduced to address the issue of class imbalance. The weight of class c is determined by

$$\frac{const}{\sum_{heartbeat\ i\ labeled\ as\ c} score_i}$$

where *const* is a constant. The Adam optimizer [47] was used in this model with the default hyperparameter setting ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Since the CNN model cannot accept varying lengths in a minibatch during the training phase, we sampled each heartbeat with 0.83 s equally around the R-peak (0.27 s before and 0.56 s after the R-peak). In the testing phase, we segmented the heartbeats using the method described in Section 3.3.

We chose the best hyperparameters using a grid search. For each model, we set the learning rate to {0.001, 0.0003, 0.0001} and the batch size to {32, 64, 128}. Besides, we applied a group normalization (GN) technique [48] to the heartbeat network by setting the hyperparameter 'groups' to {1, 32, #channels}. We found that setting 'groups' to 1 yielded the best performance, although in this situation the GN technique degenerated into the layer normalization (LN) technique [48]. Therefore, we described our model as 'using layer-normalization layers' in Section 3.4.

We evaluated the classification performance of the Attention U-net using the macro-averaged area under the curve (AUC) of the receiver operating characteristic (ROC) curve and of the precision-recall (PR) curve. The ROC curve is a function of true positive rate (TPR) and false positive rate (FPR), and the PR curve is a function of precision and recall. The macro-averaged AUC value is calculated by averaging all of the AUC values for each class. The AUC of ROC is a widely used metric for binary classification. The AUC of PR is a more informative metric than the AUC of ROC on imbalanced dataset [49].

For the heartbeat net, we simply applied an argmax rule to the prediction vector to infer the label, and thus no thresholding step was required. We used the confusion matrix, sensitivity, accuracy, and F1 score as metrics for the heartbeat net.

The model is implemented with PyTorch [50], a deep learning framework based on Python.

4.3. Results

4.3.1. Classification ability of the rhythm net

The Attention U-net's classification results are listed in Table 2. The macro-averaged AUC of the ROC curve is 0.989, while the macro-averaged AUC of the PR curve is 0.878. These results indicate that our rhythm model has learned critical features associated with arrhythmia detection.

4.3.2. Localization ability of the rhythm net

In Fig. 6, we illustrate four cases to demonstrate the localization capability of the rhythm model. The first two cases were diagnosed with supraventricular (S) and ventricular (V) arrhythmias, indicating that they were relatively difficult and complex cases. In the first case (Fig. 6a), there are only three abnormal heartbeats, and the saliency map successfully locates all abnormal heartbeats. However, in the second case (Fig. 6b) where there are multiple abnormal heartbeats, the localization results are not perfect: for supraventricular arrhythmia, the model made a mistake on a ventricular heartbeat despite the small weight, and for ventricular arrhythmia, it missed many abnormal heartbeats. The third case (Fig. 6c) was diagnosed with right bundle branch block (R) and the last case (Fig. 6d) was diagnosed with left bundle branch block (L). Compared to the first two arrhythmias, the R and L arrhythmias show a higher frequency of abnormal heartbeats. Similar to the second case, the saliency maps for R and L missed many abnormal heartbeats. Since the saliency maps were derived from the rhythm net, the model was driven to identify ONE abnormal heartbeat from each abnormal record, rather than ALL abnormal heartbeats. This is why saliency maps alone are incapable of detecting all abnormal heartbeats. Nevertheless, the saliency map can still be used to confidently identify a large number of abnormal heartbeats.

Table 2
The validation results of the rhythm network.

	AUC of ROC curve	AUC of PR curve
Supraventricular Arrhythmia	0.979	0.790
Ventricular Arrhythmia	0.986	0.888
Right Bundle Branch Block	0.997	0.967
Left Bundle Branch Block	0.995	0.865
Average	0.989	0.878

4.3.3. Distributions of different arrhythmia's scores

We use saliency maps based on the attention weights to calculate scores for all heartbeats in the abnormal ECG records. Fig. 7a depicts the score distribution for each category of arrhythmia. The curves for supraventricular and ventricular arrhythmias are steeper because of the lower frequency of abnormal heartbeats in these patients (see Fig. 6a and Fig. 6b for examples). The majority of heartbeats in these two types of arrhythmias are normal and are likely assigned a score near zero, whereas a few abnormal heartbeats are likely assigned a score near one. In comparison, the curves for left and right bundle branch block arrhythmias are smoother, because of the higher frequency of abnormal heartbeats in these patients (see Fig. 6c and Fig. 6d for examples). The majority of heartbeats in these two types of arrhythmias are abnormal and are likely assigned a positive score. Ideally, all of these positive scores should be one, but this cannot be guaranteed due to the MIL characteristics. Therefore, the score distributions for these four types of arrhythmias are consistent with their clinical characteristics, particularly the frequency of abnormal heartbeats.

We conducted an experiment to estimate the fraction of abnormal heartbeats that our model missed, also known as recall or sensitivity. Since we do not have heartbeat annotations for all ECG records, we use the rhythm validation dataset that has the ground-truth heartbeat labels to estimate the recall values at various score thresholds. Specifically, we trained the model using all training and validation data, generated scores for all heartbeats in the validation dataset, and calculated recall values using the ground-truth heartbeat labels. Fig. 7b illustrates the fractions of recalled abnormal samples with respect to various score thresholds for each type of arrhythmia. Since the frequencies of R and L heartbeats are higher, there is no surprise that their recall rates are low. We also notice that the recalled rates of S samples are significantly lower than that of V samples, which is consistent with the classification performance difference between S and V in Table 2.

4.3.4. Classification results of the heartbeat net

We begin by setting the hyperparameter α to 1.0 when training the heartbeat network. The pseudo-heartbeat training set has 55,632 N beats, 2934 S beats, 3483 V beats, 20,443 R beats, and 3729 L beats (for example, one S beat with score 0.5 is regarded as 0.5 beat).

Table 3 compares our MIL model with a baseline model (baseline), which is a pure heartbeat network identical to our proposed one in Fig. 5. The baseline model was trained on the heartbeat data with ground-truth heartbeat labels, using identical hyperparameters as those proposed in our model. We applied a patient-level five-fold cross-validation strategy to estimate the performance, with each of the five folds consisting of a disjointed set of patients (or records).

As demonstrated in Table 3, our MIL model outperforms the baseline model across nearly all types of heartbeats and all metrics, except in the sensitivity of Normal Arrhythmia and Right Bundle Branch Block, and accuracy of Supraventricular Arrhythmia. Fig. 8 illustrates the confusion matrixes for both models. None of the models was able to accurately distinguish supraventricular from normal heartbeats, which is a challenging task in clinical practice as well as research [7]. The fundamental reason is that the morphological difference between supraventricular

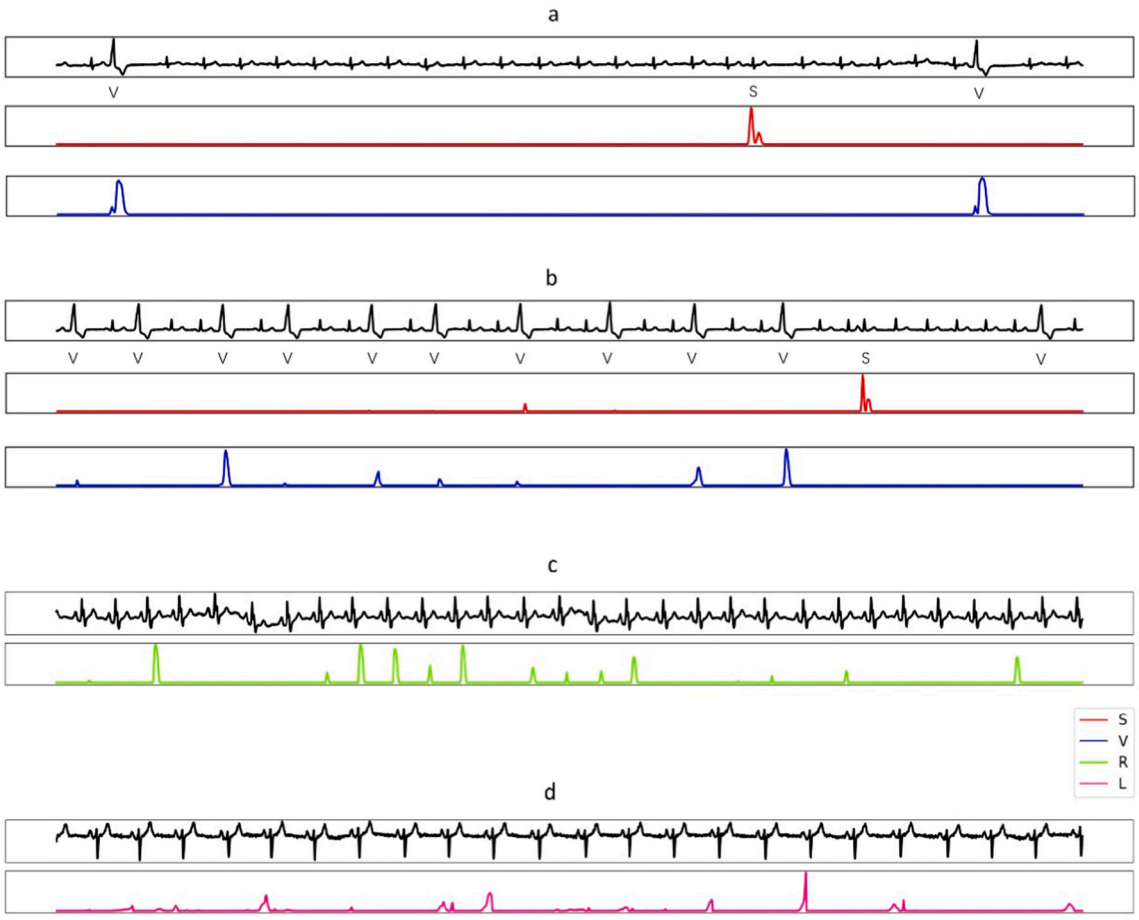


Fig. 6. Four examples of the generated saliency maps. For each example, the raw lead-V2 ECG signal is plotted in black line. The ECG records for (a), (b), (c), and (d) were diagnosed with S&V, S&V, R and L, respectively. The abnormal heartbeats of S and V are displayed in (a) and (b). All the heartbeats are abnormal (R or L) in (c) and (d).

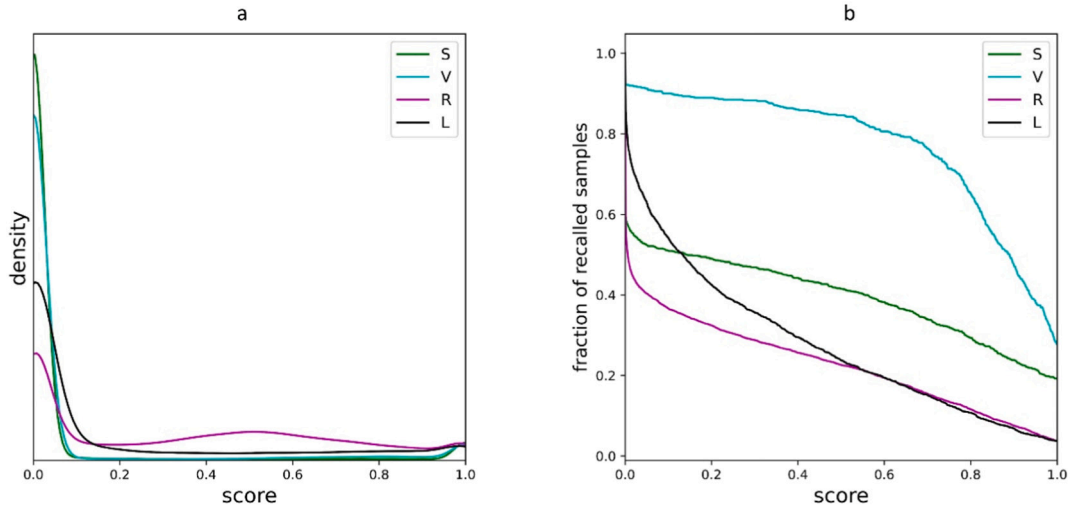


Fig. 7. (a) The score distribution for each category of arrhythmia in the pseudo heartbeat-level training set. (b) The fraction of recalled abnormal samples for each category of arrhythmia at various score thresholds.

and normal heartbeats is small.

4.3.5. *Impact of the hyperparameter α*

Our model puts weight $score^{\alpha}$ to each selected heartbeat in the training set. We test various values of α and observe the resulting performance measured by the F1 scores. We set α = 0.0,0.5,0.67,1.0,1.5,2.0,1000.0 respectively. The results are illustrated in Fig. 9. The worst case is when α = 0, which corresponds to the situation where the score is not used and the prior information of the saliency map is not utilized. While setting α to 1.0 resulted in the best

Table 3
Comparison of the heartbeat net trained on the rhythm dataset ('MIL') and on the ground-truth heartbeat labels using a 5-fold cross validation strategy ('baseline'). [a, b] is the 95 % CI.

	Sensitivity (MIL)	Sensitivity (baseline)	Accuracy (MIL)	Accuracy (baseline)	F1 score (MIL)	F1 score (baseline)
Normal	0.872[0.864,0.879]	**0.897**[0.890,0.904]	**0.890**[0.885,0.895]	0.872[0.867,0.877]	**0.883**[0.877,0.889]	0.870[0.865,0.876]
Supraventricular Arrhythmia	**0.522**[0.488,0.556]	0.275[0.245,0.306]	0.935[0.931,0.939]	**0.942**[0.939,0.946]	**0.469**[0.441,0.496]	0.344[0.312,0.377]
Ventricular Arrhythmia	**0.875**[0.847,0.902]	0.849[0.818,0.878]	**0.993**[0.991,0.994]	0.983[0.980,0.985]	**0.894**[0.873,0.912]	0.775[0.748,0.800]
Right Bundle Branch Block	0.918[0.910,0.925]	**0.919**[0.911,0.926]	**0.957**[0.954,0.961]	0.947[0.943,0.950]	**0.931**[0.925,0.936]	0.915[0.909,0.921]
Left Bundle Branch Block	**0.892**[0.878,0.906]	0.737[0.716,0.757]	**0.964**[0.961,0.967]	0.953[0.950,0.956]	**0.857**[0.845,0.869]	0.792[0.777,0.806]
Average	**0.816**[0.806,0.825]	0.735[0.726,0.745]	**0.948**[0.946,0.950]	0.939[0.937,0.942]	**0.807**[0.798,0.815]	0.739[0.729,0.749]

The bold number indicates that the result of the corresponding method (MIL or baseline) is better than the other method under certain category and evaluation metric.

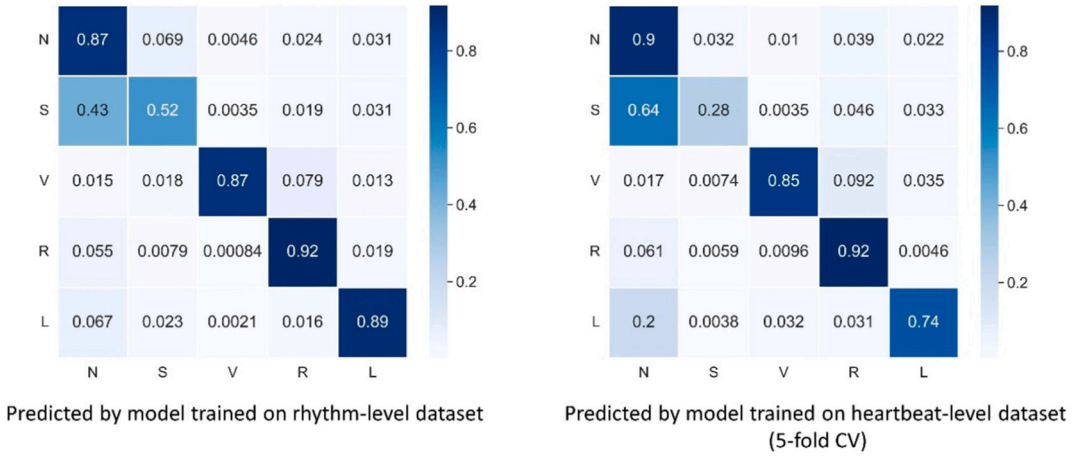


Fig. 8. The confusion matrices for the MIL model (Left) and the baseline model (right).

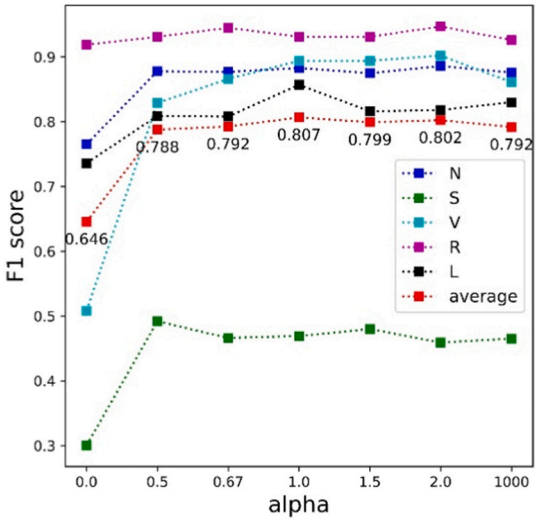


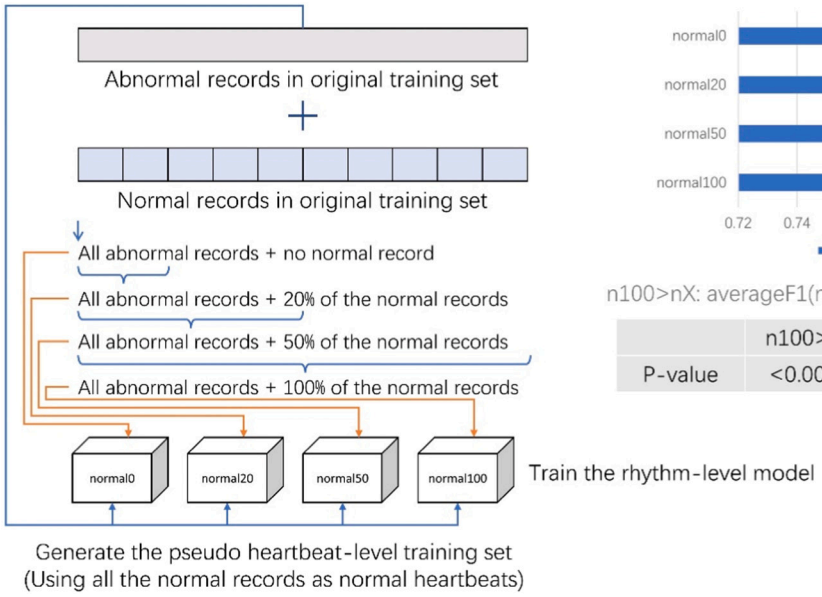
Fig. 9. The F1 scores at various values of the hyperparameter alpha for each category.

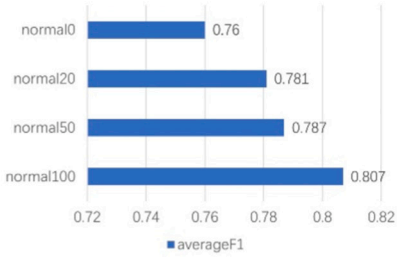
performance, we found that the performance did not vary significantly when $\alpha \geq 0.5$. This indicates that our method is not sensitive to the hyperparameter α. The performance remains acceptable when $\alpha = 1000$, of which the performance is similar to that of $\alpha = \infty$. The results demonstrate that the weight $score^{\alpha}$ controls the effect of noisy labels that usually have low scores ($score^{\alpha}$ is small for scores close to 0), and contributes to the high performance. We can also observe that the F1 scores for the right bundle branch block and left bundle branch block vary only slightly with different settings of α. Again, since these two types of arrhythmias occur frequently, even the baseline ($\alpha = 0$) method can achieve acceptable performance. Our method, on the other hand, significantly improves the classification performance of supraventricular and ventricular arrhythmias.

4.3.6. *Impact of the number of training samples*

We designed two experiments to elucidate the rhythm network. First, given that the rhythm network is used to generate saliency maps for irregular records, a natural question is whether normal records are required to train the network. To address this question, we trained three additional rhythm networks using all the abnormal records but each with a different fraction of normal records, 0 %, 20 %, and 50 %, from the original training set. These trained networks are named as 'normal0', 'normal20', and 'normal50', respectively. Besides, we named our origin rhythm model which is trained with all normal records as 'normal100'. Then, we constructed three pseudo-heartbeat-level training sets using the three rhythm networks respectively and trained the heartbeat models. Fig. 10 (left) depicts the experimental setting, while Fig. 10 (right) illustrates the resulting averaged F1 scores in the testing set. We observe that the performance improves as the number of normal records increases, indicating that the normal records can aid the model in learning the discriminative features of the abnormal records.

Second, we ask whether the number of training records has an effect on the generation of reliable saliency maps. We trained the rhythm network using 10 % (set A), 50 % (set B), and 100 % (set C) of the records in the original training set, respectively, ensuring that $A \subseteq B \subseteq C$. These trained networks are named as 'part10', 'part50', and 'part100', respectively. Then, we constructed pseudo-heartbeat-level training sets for all the three models using the first set (set A), and trained the heartbeat models. Fig. 11 (left) depicts the experimental setting, while Fig. 11 (right) illustrates the resulting averaged F1 scores in the testing set. We used the same 10 % of the records to train the heartbeat model. The only difference is the number of records used to train the rhythm network. As can be seen, the number of training records contributes positively to overall performance improvement. Since the saliency maps bridge the networks at the rhythm-level and heartbeat-level, we can

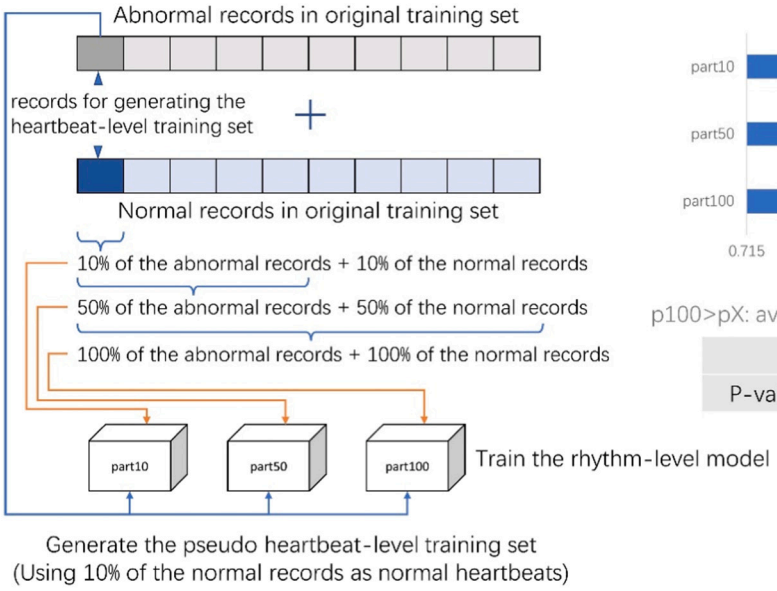


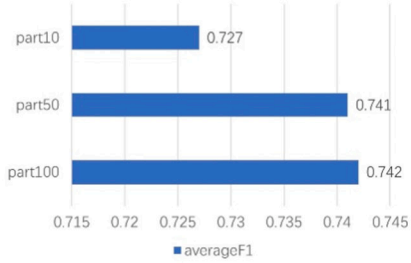


n100>nX: averageF1(normal100)>averageF1(normalX)

	n100>n0	n100>n20	n100>n50
P-value	<0.0001	<0.0001	<0.0001

Fig. 10. Left: Various fractions (0 %, 20 %, 50 %, and 100 %) of the normal records in the original rhythm-level training set were used to train the Attention U-net and generate the heartbeat-level dataset. Right: The averaged F1 scores for these fractions on the heartbeat-level test set, and the statistical significance for the differences of the results.





p100>pX: averageF1(part100)>averageF1(partX)

	p100>p10	p100>p50
P-value	<0.0001	0.3971

Fig. 11. Left: Various fractions (10 %, 50 %, and 100 %) of the records in the original rhythm-level training set were used to train the Attention U-net. Only 10 % of the records were used to generate the heartbeat-level dataset for testing. Right: The averaged F1 scores for all fractions on the heartbeat-level test set, and the statistical significance for the differences of the results.

conclude that increasing the number of training records does contribute positively to the generation of more reliable saliency maps. We also notice that the improvement from 'part50' to 'part100' is not statistically significant, indicating that the ability of generating reliable saliency maps is close to the upper bound when using 50 % data for training.

5. Discussion

With the rapid development of medical informatics, researchers have recently gained access to an increasing volume of raw ECG signals, which has prompted multiple studies concentrating on automatically diagnosing arrhythmia (see Section 2). Nonetheless, most of these works are limited to rhythm-level analysis since there are few heartbeat labels for raw ECG signals. Compared to the rhythm model, the heartbeat model has at least two advantages. (1) Because the number of model parameters is significantly less, it is easier to perform real-time analysis, which is necessary for clinical applications such as generating ECG checking reports and analyzing ECG data from wearable devices such as the Holter monitor. (2) The heartbeat classification report is easier to comprehend for human experts. Although our work focuses on utilizing the ECG records with rhythm labels, our model incorporates a heartbeat model to achieve the benefits mentioned above.

We collected 68,716 records from Yichang Central People's Hospital in Hubei Province, China, to train and evaluate our proposed method. These de-identified data (i.e., patient-related information is removed)

were diagnosed at the rhythm level by cardiologists. The setting of the dataset is comparable to a variety of real-world events, which adds credibility to our experimental results. The dataset is sufficiently large to train a robust model.

To compare our heartbeat model to the model trained on the normal heartbeat dataset (i.e., data with ground-truth heartbeat annotations), we utilized the testing set to train a model using a patient-level-five-fold cross-validation strategy. To train the model, we used 61,853 records, including 56,031 normal records and 5822 abnormal records. On the other hand, the baseline model is trained using heartbeats from 633 abnormal records. This means that by using about nine times the number of abnormal records and a number of normal records with only rhythm labels, our model can outperform the baseline model trained on heartbeat labels. Although this is not a direct comparison, we believe it is meaningful since, in the real world, it is difficult to obtain a large number of heartbeat annotations, whereas it is much easier to obtain a large number of ECGs with rhythm annotations.

The hyperparameter α is introduced to strike a balance between the number of training samples and the issue of noisy labels. In this study, we analyzed the effect of different values of α. In our experiment, setting $\alpha = 1$ produced the best performance, which is equivalent to using the origin scores as the weights of the loss function. Although it is unknown whether this setting is optimal for other datasets, we found that the performance is relatively insensitive to the value α, for example, that the average F1 score varies only slightly when $\alpha \geq 0.5$. Therefore, there is no need to exert excessive tuning effort when applying our method.

We demonstrated that by increasing the number of normal ECGs or all records, we can generate more reliable saliency maps to train a more accurate heartbeat classification model. As mentioned previously in the introduction section, deep-learning-based approaches typically require a large amount of labeled data. Our approach is no different. To improve the overall performance, our results suggest that a low-cost strategy is to collect more normal records into the training set, since abnormal records are usually much harder to obtain. Nonetheless, it is not clear that our approach is suitable for detecting rare types of arrhythmias since obtaining these records is extremely difficult.

An alternative approach to developing a heartbeat classifier is a one-stage algorithm that is similar to the model proposed in [34] with a few modifications. In this algorithm, we first segment a record into multiple heartbeats, then classify them using a heartbeat classification network, and finally aggregate the heartbeat classification probabilities into one rhythm classification probability using the instance-level MIL pooling function. In [34], the rhythm classification probability is defined as the average of the highest 20 % heartbeat classification probabilities. After training the model on the rhythm ECG data, we can use the heartbeat classification network for heartbeat diagnosis. We implemented this approach and trained and tested it on our annotated heartbeat data, resulting in an average F1 score of 0.765 (95 % CI: [0.755, 0.774]) for heartbeat classification, which is outperformed by our proposed two-stage algorithm (p-value$<$0.0001). The primary difference between the two approaches is that our model used the entire ECG record as input, rather than segmented heartbeats, leveraging global and context information to more precisely locate abnormal heartbeats.

One of the major limitations of our work is that we only consider four distinct types of arrhythmias. Other arrhythmias, such as atrioventricular block, are excluded because of the scarcity of these records in the hospital. Additional data collection will be necessary to train our method for such tasks. Another shortcoming is that we used only the saliency maps as the prior to identify the abnormal heartbeats. Indeed, clinical notes can occasionally provide useful information, such as an ECG record diagnosed as premature ventricular bigeminy, a condition in which normal and ventricular heartbeats alternate throughout a period. In our experiment, we simply categorized this ECG as ventricular arrhythmia and ignored the information above that would be useful in detecting abnormal heartbeats. We may examine how to utilize this information to improve the detection of abnormal heartbeats in the future. Finally, the data we collected are from a single hospital, with the majority of the records pertaining to neonate patients. Therefore, the performance of external validation using data from diverse hospitals and different age distributions cannot be guaranteed. For example, when we directly applied our model to the MIT-BIH arrhythmia dataset for testing, we obtained an average F1 score of 0.249, which is significantly lower than the performance of our internal validation. Our model's poor performance on MIT-BIH is mostly explained by the significant difference between neonatal and adult ECGs. While neonatal, pediatric, and adult ECG diagnoses are all based on the same principles, there are significant variations due to physiological and anatomical changes during childhood development, with the most dramatic changes occurring during the first year of life. Infants and children have distinct ECG amplitudes, intervals, and waveforms, and there are marked age variations throughout development. More specifically, because a child's heart is smaller than an adult's, it has fewer myocardial cells to depolarize and repolarize. This explains why all ECG intervals in infants are much shorter (PR intervals, QRS duration, QT intervals, etc.). As the heart grows by age, these intervals increase as well. Other possible explanations include the fact that the MIT-MIH data were collected several decades ago using an old electrocardiograph and acquisition protocol, while our data were collected recently, and that the annotation standards for the MIT-BIH data and our data may differ slightly. To address the issue further, data from multi-center, multi-age, and multi-national groups should be collected.

6. Conclusion and future work

We proposed a method for training a heartbeat arrhythmia classifier solely using rhythm annotations. We trained an Attention U-net as a rhythm classifier and then generated saliency maps based on the attention weight as the abnormal training samples. Then, for each heartbeat in abnormal records, we calculate a score and use it as a weight in training the heartbeat classifier. We collected 68,716 ECG records to evaluate our method. Our model achieved a macro-averaged F1 score of 0.807 in classifying normal, supraventricular arrhythmia, ventricular arrhythmia, right bundle branch block, and left bundle branch block. We compared our model to the baseline model trained on 15,385 labeled heartbeats, and we discovered that our model significantly outperformed the baseline model. We also discovered that our method was robust to hyperparameter selection, indicating its stability. Additionally, we discussed the effect of the number of training records on the rhythm network and we discovered that increasing the amount of data helps to improve the performance of the heartbeat model. All of these results demonstrate the reliability of our method for training heartbeat arrhythmia classifiers using only rhythm annotations. We expect that as more ECG records with rhythm annotations become available, we are able to use our computational framework to develop a more precise heartbeat diagnosis model.

In the future, we want to (1) improve the heartbeat classification performance by exploring more efficient network architectures for both rhythm and heartbeat networks, and (2) conduct clinical trials to assess how well our model performs compared to cardiologists.

Declaration of competing interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (2021YFF1201303 and 2019YFB1404804), National Natural Science Foundation of China (grants 61872218 and 61721003), Guoqiang Institute of Tsinghua University, Tsinghua University Initiative Scientific Research Program, and Beijing National

Research Center for Information Science and Technology (BNRist). The funders had no roles in study design, data collection and analysis, the decision to publish, and preparation of the manuscript.

References

[1] Clinical ECG interpretation. https://ecgwaves.com/course/the-ecg-book/.html; 2022.

[2] Jambukia SH, Dabhi VK, Prajapati HB. Classification of ECG signals using machine learning techniques: a survey. In: 2015 international conference on advances in computer engineering and applications; 2015. p. 714–21. https://doi.org/10.1109/ICACEA.2015.7164783.

[3] Memon MS, Lakhan A, Mohammed MA, Qabulio M, AlTurjman F, Abdulkareem KH, et al. Machine learning-data mining integrated approach for premature ventricular contraction prediction. Neural Comput & Applic 2021:1–17. https://doi.org/10.1007/s00521-021-05820-2.

[4] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med 2019;25(1):70–4. https://doi.org/10.1038/s41591-018-0240-2.

[5] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.

[6] Onan A. Biomedical text categorization based on ensemble pruning and optimized topic modelling. Comput Math Methods Med 2018. https://doi.org/10.1155/2018/2497471.

[7] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–9. https://doi.org/10.1038/s41591-018-0268-3.

[8] Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, San Tan R. A deep convolutional neural network model to classify heartbeats. Comput Biol Med 2017;89:389–96. https://doi.org/10.1016/j.compbiomed.2017.08.022.

[9] Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Adv Neural Inf Proces Syst 1997;10:570–6.

[10] Zhang X, Chen T. Attention U-net for interpretable classification on chest X-ray image. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM); 2020. p. 901–8. https://doi.org/10.1109/BIBM49941.2020.9313354.

[11] Rajkumar A, Ganesan M, Lavanya R. Arrhythmia classification on ECG using deep learning. In: 2019 5th international conference on advanced computing & communication systems (ICACCS); 2019. p. 365–9. https://doi.org/10.1109/ICACCS.2019.8728362.

[12] Li F, Wu J, Jia M, Chen Z, Pu Y. Automated heartbeat classification exploiting convolutional neural network with channel-wise attention. IEEE Access 2019;7:122955–63. https://doi.org/10.1109/ACCESS.2019.2938617.

[13] Izci E, Ozdemir MA, Degirmenci M. Cardiac arrhythmia detection from 2d ecg images by using deep learning technique. In: 2019 medical technologies congress (TIPTEKNO); 2019. p. 1–4. https://doi.org/10.1109/TIPTEKNO.2019.8895011.

[14] Degirmenci M, Ozdemir MA, Izci E, et al. Arrhythmic heartbeat classification using 2d convolutional neural networks. IRBM 2021. https://doi.org/10.1016/j.irbm.2021.04.002.

[15] Ahmad Z, Tabassum A, Guan L, Khan NM. ECG heartbeat classification using multimodal fusion. IEEE Access 2021;9:100615–26. https://doi.org/10.1109/ACCESS.2021.3097614.

[16] Hiriyannaiah S, Siddesh GM, Kiran MHM, et al. A comparative study and analysis of LSTM deep neural networks for heartbeats classification. Health Technol 2021;11(3):663–71. https://doi.org/10.1007/s12553-021-00552-8.

[17] Singh S, Pandey SK, Pawar U, et al. Classification of ECG arrhythmia using recurrent neural networks. Procedia Comput Sci 2018;132:1290–7. https://doi.org/10.1016/j.procs.2018.05.045.

[18] Zhang C, Wang G, Zhao J, et al. Patient-specific ECG classification based on recurrent neural networks and clustering technique. In: 2017 13th IASTED international conference on biomedical engineering (BioMed); 2017. p. 63–7. https://doi.org/10.2316/P.2017.852-029.

[19] Yan Y, Qin X, Wu Y, et al. A restricted Boltzmann machine based two-lead electrocardiography classification. In: 2015 IEEE 12th international conference on wearable and implantable body sensor networks (BSN); 2015. p. 1–9. https://doi.org/10.1109/BSN.2015.7299399.

[20] Pandey SK, Janghel RR, Dev AV, et al. Automated arrhythmia detection from electrocardiogram signal using stacked restricted Boltzmann machine model. SN Appl Sci 2021;3(6):1–10. https://doi.org/10.1007/s42452-021-04621-5.

[21] Essa E, Xie X. An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification. IEEE Access 2021;9:103452–64. https://doi.org/10.1109/ACCESS.2021.3098986.

[22] Xu X, Jeong S, Li J. Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. IEEE Access 2020;8:125380–8. https://doi.org/10.1109/ACCESS.2020.3006707.

[23] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

[24] Kharshid A, Alhichri HS, Ouni R, et al. Classification of short-time single-lead ECG recordings using deep residual CNN. In: 2019 2nd international conference on new trends in computing sciences (ICTCS); 2019. p. 1–6. https://doi.org/10.1109/ICTCS.2019.8923079.

[25] Park J, Kim J, Jung S, et al. ECG-signal multi-classification model based on squeeze-and-excitation residual neural networks. Appl Sci 2020;10(18):6495. https://doi.org/10.3390/app10186495.

[26] Zhang H, Zhao W, Liu S, SE-ECGNet. A multi-scale deep residual network with squeeze-and-excitation module for ECG signal classification. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM); 2020. p. 2685–91. https://doi.org/10.1109/BIBM49941.2020.9313548.

[27] Xiong Z, Nash MP, Cheng E, et al. ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. Physiol Meas 2018;39(9):094006.

[28] Kim K. Arrhythmia classification in multi-channel ECG signals using deep neural networks. Technical report no. UCB/EECS-2018-80. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-80.html; 2018.

[29] Zhang J, Liu A, Gao M, et al. ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. Artif Intell Med 2020;106:101856. https://doi.org/10.1016/j.artmed.2020.101856.

[30] Hong S, Xiao C, Ma T, Mina, et al. Multilevel knowledge-guided attention for modeling electrocardiography signals. In: The twenty-eighth international joint conference on artificial intelligence (IJCAI); 2019. p. 5888–94. https://doi.org/10.24963/ijcai.2019/816.

[31] Carbonneau M-A, Granger E, Gagnon G. Score thresholding for accurate instance classification in multiple instance learning. In: 2016 sixth international conference on image processing theory, tools and applications (IPTA); 2016. p. 1–6. https://doi.org/10.1109/IPTA.2016.7821026.

[32] Wang X, Yan Y, Tang P, et al. Revisiting multiple instance neural networks. Pattern Recogn 2018;74:15–24. https://doi.org/10.1016/j.patcog.2017.08.026.

[33] Sun L, Lu Y, Yang K, et al. ECG analysis using multiple instance learning for myocardial infarction detection. IEEE Trans Biomed Eng 2012;59(12):3348–56. https://doi.org/10.1109/TBME.2012.2213597.

[34] Shanmugam D, Blalock D, Guttag J. Multiple instance learning for ECG risk stratification. In: Machine learning for healthcare conference. PMLR; 2019. p. 124–39.

[35] Novotna P, Vicar T, Ronzhina M, et al. Deep-learning premature contraction localization in 12-Lead ECG from whole signal annotations. In: 2020 computing in cardiology; 2020. p. 1–4. https://doi.org/10.22489/CinC.2020.193.

[36] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR; 2018. p. 2127–36. https://doi.org/10.48550/arXiv.1802.04712.

[37] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR; 2015. p. 448–56. https://doi.org/10.48550/arXiv.1502.03167.

[38] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

[39] He K, Zhang X, Ren S. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.

[40] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2921–9. https://doi.org/10.1109/CVPR.2016.319.

[41] Christov II. Real time electrocardiogram QRS detection using combined adaptive threshold. Biomed Eng Online 2004;3(1):1–9. https://doi.org/10.1186/1475-925X-3-28.

[42] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint; 2016. https://doi.org/10.48550/arXiv.1607.06450.

[43] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–27. https://doi.org/10.1109/TPAMI.2018.2858826.

[44] Ben-Baruch E, Ridnik T, Zamir N. Asymmetric loss for multi-label classification. arXiv preprint; 2020. https://doi.org/10.48550/arXiv.2009.14119.

[45] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1026–34. https://doi.org/10.1109/ICCV.2015.123.

[46] Van Laarhoven T. L2 regularization versus batch and weight normalization. arXiv preprint; 2017. https://doi.org/10.48550/arXiv.1706.05350.

[47] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint; 2014. https://doi.org/10.48550/arXiv.1412.6980.

[48] Wu Y, He K. Group normalization. Int J Comput Vis 2020;128(3):742–55. https://doi.org/10.1007/s11263-019-01198-w.

[49] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.

[50] Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Proces Syst 2019;32. https://doi.org/10.48550/arXiv.1912.01703.