Identification of 27 cardiac abnormalities from multi-lead ECG signals: An ensembled Se-ResNet framework with Sign Loss function

Zhaowei Zhu\textsuperscript{1†}, Xiang Lan\textsuperscript{2†}, Tingting Zhao\textsuperscript{1}, Yangming Guo\textsuperscript{1}, Pipin Kojodjojo\textsuperscript{2}, Zhuoyang Xu\textsuperscript{1}, Zhuo Liu\textsuperscript{1}, Siqi Liu\textsuperscript{3}, Han Wang\textsuperscript{2}, Xingzhi Sun\textsuperscript{1‡}, Mengling Feng\textsuperscript{2,4‡}

\textsuperscript{1} Ping An Technology, Beijing, China
\textsuperscript{2} Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore
\textsuperscript{3} NUS Graduate School for Integrative Sciences & Engineering, Singapore
\textsuperscript{4} Institute of Data Science, National University of Singapore, Singapore

E-mail: zhuzhaowei262@pingan.com.cn, ephlanx@nus.edu.sg

Abstract.

Objective: Cardiovascular disease is a major threat to health and one of the primary causes of death globally, where the cardiac abnormality is the most common type of cardiovascular disease. The early and accurate diagnosis of cardiac abnormalities will allow early treatment and intervention to prevent further progression of the disease. In accordance with the PhysioNet/Computing in Cardiology Challenge 2020, our objective is to develop an algorithm that automatically identifies 27 types of cardiac abnormality from 12-lead ECG recordings.

Approach: Firstly, we ensembled two Se-ResNet and one rule-based model to enhance model performance and efficiency customizing to the heterogeneous characteristics of various cardiac abnormalities. Secondly, we introduce a Sign Loss to tackle the problem of class imbalance, and thus improve the model’s generalizability. Thirdly, in order to mitigate the divergence of data from different sources and help the model train better, a series of multi-source data preprocessing method was proposed and applied.

Main Results: Our proposed approach achieved a challenge validation score of 0.682, and a full test score of 0.514, placed us 3rd out of 40 in the official ranking.

Significance: We proposed an accurate and robust predictive framework that combines both deep neural networks and rule-based models to automatically classify multiple cardiac abnormalities. Our framework is able to identify 27 types of cardiac abnormalities presented in 12-lead ECG data. And our proposed Sign Loss makes the framework able to tackle the imbalanced dataset. Meanwhile, our framework was trained and verified on 6 datasets from different countries, and our proposed multi-source preprocessing methods are able to reduce the discrepancies of ECG datasets from different sources.

\textsuperscript{†}These two authors are jointly first authors.
\textsuperscript{‡}These two authors are jointly corresponding authors.
1. Introduction

Cardiovascular disease is one of the primary causes of death globally, which took an estimated 16.7 million lives each year, according to the World Health Organization Gaziano et al. (2010). An important type of cardiovascular disease is a cardiac abnormality. The early and accurate diagnosis of cardiac abnormalities can prevent sudden cardiac death and other lethal illnesses caused by heart disease, also increase the chances of successful treatments. Artis et al. (1991)

Cardiovascular abnormalities are commonly diagnosed with Electrocardiogram (ECG) examinations Kligfield et al. (2007). To detect cardiac abnormalities, continuous monitoring of the ECG signal is required. The diagnosis process requires well-trained clinicians to carefully examine and identify the inter-beat and intra-beat patterns. The process can be both time-consuming and error-prone Bickerton and Pooler (2019). Hence, a quick and accurate algorithm for automatic ECG pattern classification is always desired.

Some earlier works have been reported to detect abnormal heart conditions automatically Martínez et al. (2004); Minami et al. (1999); Mahmoodabadi et al. (2005); Alexakis et al. (2003). These approaches are mainly based on frequency domain features, time-frequency analysis, transforms such as Wavelet transform and Fourier transform. However, such techniques are not able to capture non-linear features of the ECG signal, thus the performance of these approaches is not good.

More recently, a number of works have demonstrated the ability of non-linear machine learning techniques in the field of ECG analysis. Vafaie et al. (2014) proposed a classifier to predict heart diseases, in which a fuzzy-classifier was constructed and the genetic algorithm was applied to estimate parameters. Chen et al. (2018) designed a gradient boosting algorithm to detect atrial fibrillation, piece-wise linear splines were used for the feature selection. However, these techniques are usually required to find optimal features, or manually generate features by sophisticated feature engineering for the model to learn. This procedure actually requires expert knowledge and several times verification, which is time-consuming.

In recent years, deep learning and neural networks, especially the Convolutional Neural Networks (CNNs) LeCun et al. (1995), have gained promising results in many areas. For cardiac abnormalities detection, Xiong et al. (2018) developed a 21-layer 1D convolutional recurrent neural network to detect atrial fibrillation, trained on single lead ECG data. Sodmann et al. (2018) proposed a CNN model to annotate QRS complexes, P waves, T waves, noise, and interbeat of ECG segments, thus improve detection performance. Warrick and Homsi (2018) designed an ensemble deep learning model for automatic classification of cardiac arrhythmias based on single lead ECGs, which fused the decision of ten classifiers to derive a more powerful ECG classification scheme than a single deep classifier.

The novel contributions of this work include:

1) Most of the previous works focus on one or at most 9 cardiac abnormalities
Identification of 27 cardiac abnormalities from multi-lead ECG signals

Wang et al. (2019); Luo et al. (2019); Huang et al. (2019); Xia et al. (2019). But here we aim to develop a robust model that is able to generalize over and identify twenty-seven different types of cardiac abnormalities. In addition, most of the existing works only deal with single lead ECG signals Andreotti et al. (2017); Billeci et al. (2017); Bin et al. (2017); Warrick and Homsi (2018), while in the clinical practice, 12-lead ECGs are more commonly used for abnormality detection and diagnosis. Therefore, to be more applicable to the actual clinical setting, we develop a model that reads 12-lead ECG signals instead.

2) From our statistical analysis of the dataset, we found a marked class imbalance of the dataset. Where the number of samples is more than 20,000 in some majority cardiac abnormalities, while some minor cardiac abnormalities have less than 500 samples. Consequently, the model is preferred to learn information from majority categories, and perform poorly in minor categories. To mitigate the negative effects caused by class imbalance problem, we propose to solve it with a Sign Loss function.

3) Unlike the aforementioned works that developed on a small dataset from a single source. We have train and verified the performance of our model over 6 different datasets across the world, and propose a series of multi-source ECG data preprocessing method.

2. Methods

The overall framework design is shown in figure 1 and our methods will be elaborated below.

2.1. Datasets

Training dataset. The public challenge data consist of 43,101 12-lead ECG signals from 6 different datasets over the world, namely CPSC, CPSC2, PTB, PTB-XL, Georgia, and INCART. The sampling frequency of the signals varies from 257 Hz to 1000 Hz, and the length of the signals varies from 6 seconds to 30 minutes. There are 111 labeled abnormalities in total, of which 27 are included in the final scoring metrics. From these data, we created our offline training set and validation set. Table 1 shows basic information of the six different datasets.

Offline validation dataset. To monitor our model’s performance, after we processed original training data as described in 2.2, we randomly split 80% as the training set and 20% as our offline validation set. The final sizes of the training set and offline validation set are 30,172 and 7,544 respectively.

In order to help us select the model with acceptable generalizability, an external dataset from the Hefei Hi-tech Cup ECG Intelligent Competition TIANCHI - Hefei Hi-tech Cup ECG Intelligent Competition (2020) (Hefei dataset in short) was introduced. Hefei dataset consists of 40,000 records of 8-lead ECG signals with a sampling frequency of 500 Hz and length of 10 seconds. Out of all the records, 6,500 records with labels in the 27 types of cardiac abnormalities that we focused were randomly selected and
Identification of 27 cardiac abnormalities from multi-lead ECG signals

Figure 1: Illustration of the framework design. We ensembled two Se-ResNet and one rule-based model to improve the model’s performance and efficiency. Both Se-ResNets are using a special multi-label Sign Loss function to mitigate the class imbalance problem. To enable the model train over 6 different dataset, a series of data preprocessing method was proposed and applied in the framework.

| Databases   | Recordings | Labels | Subset | Extra labels | Extra Recordings |
|-------------|------------|--------|--------|--------------|-----------------|
| CPSC        | 6,877      | 9      | True   | -            | -               |
| CPSC2       | 3,453      | 72     | True   | -            | -               |
| INCART      | 74         | 37     | False  | 11           | 53              |
| PTB         | 516        | 17     | False  | 7            | 46              |
| PTB-XL      | 21,837     | 50     | True   | -            | -               |
| Georgia     | 10,344     | 67     | True   | -            | -               |

Table 1: Basic information of six different datasets that forms the challenge dataset.

formed an external validation set.

Online test dataset. The entire online test data contains 11,630 12-lead ECG recordings that were not represented in the training data. The test data were drawn from three databases shown below.

Test Database 1: A total of 1,463 ECGs from Southeast University, China, including the data from the China Physiological Signal Challenge 2018.

Test Database 2: A total of 5,167 ECGs from the Georgia 12-Lead ECG Challenge Database, Emory University, Atlanta, Georgia, USA.

Test Database 3: A total of 10,000 ECGs from an unspecified US institution comparable to Test Database 2, matched for demographics and prevalence of classes.
Identification of 27 cardiac abnormalities from multi-lead ECG signals

2.2. Multi-label classification with 12-leads ECG

Se-ResNet. One important feature of 12-lead ECG signal is the information contained differs in different leads, due to the difference of signal voltage intensity and variation amplitude at each sampling point. Different cardiac abnormalities may be represented in different leads. An equal importance of different leads could cause information losses, which lead to misdiagnosis or missed diagnosis.

Therefore, in order to enable the model to learn the distinct importance of different leads for different cardiac abnormality, we use Se-ResNet Hu et al. (2018) as our main model to enhance the model performance for multi-leads ECG signal. We integrate a Squeeze-and-excitation (Se) block into the ResNet He et al. (2016) structure. The squeeze operation compresses the global spatial information and produces an embedding of the global distribution of feature responses for each channel, thus all the layers of the network are able to use the information from the global receptive field. The excitation operation takes the embedding as input and capture the channel-wise dependencies, then produce weights for each channel. Consequently, these weights are applied to previously learned feature maps and realize the feature recalibration.

In this way, more important leads could be given higher weights, leading to a better prediction performance for multi-lead ECG classification. Our baseline model was a Se-ResNet model with an input length of 30 seconds. To minimize the effect of padding on the shorter signals, another Se-ResNet model was trained with the input length of 10 seconds and ensembled with the baseline model. The structure of our Se-ResNet model is shown in Figure 2.

Rule-based model. The baseline model did not perform well enough for certain
classes while there were relatively clear clinical rules to follow. One of such classes is bradycardia, which indicates the heart rate is slower than 1 beat per second, or the R-R interval between two heartbeats is longer than 1 second. To detect the R-R intervals, Pan & Tompkins algorithm Pan and Tompkins (1985) was used to detect the R-peaks on lead I, and R-R intervals could be easily calculated. The pseudocode of the rule-based model for bradycardia is shown in Algorithm 1.

**Algorithm 1: Rule-based bradycardia classifier**

**Input:** List of R-R intervals  
**Output:** Classification of bradycardia  

brady_beats = 0;  
foreach R-R interval  
    if 1s ≤ length of interval ≤ 1.6s  
        brady_beats += 1  
    if brady_beats / # of R-R intervals ≥ 0.5  
        return True  
    else  
        return False

However, the final bradycardia prediction of the system was not purely decided by the rule-based model. A very high recall and low precision were observed when doing so, and this could be attributed to the sub-optimal label quality of the datasets. Therefore, the prediction of the rule-based model was only taken when its output is negative. The pseudocode for the final bradycardia prediction is shown in Algorithm 2.

**Algorithm 2: Final bradycardia prediction**

**Input:** Prediction from ensembled model  
**Output:** Prediction from rule-based model  

if Prediction from rule-based model is False  
    return False  
else  
    return Prediction from ensembled model

Model Ensembling. Considering that the actual application scenarios and the final evaluation test set contain new data sources, it is necessary to improve the generalization ability of the model as much as possible. The fusion model is a common and effective method to improve the robustness of the model. The idea of fusion in this paper is to fuse two models that receive different length input signals. Different data lengths give the model multiple views to capture valid information. The two lengths selected in this model are 10s and 30s (corresponding to the data length of 5000 and 15000). Signals that were predicted to be negative for all classes were revised to be positive for the default normal class, sinus rhythm (SNR).
Identification of 27 cardiac abnormalities from multi-lead ECG signals

Figure 3: Total counts of 27 cardiac abnormalities in the original dataset, from which we observed a significant class imbalance of the dataset. The number of samples of Sinus rhythm (NSR) is more than 20000, while there are fewer than 200 samples for Premature ventricular contractions (PVC). Such imbalance could undermine the model’s performance, as the model is likely to learn the pattern from categories with a large number of samples while ignoring the minorities.

2.3. Class imbalance

Sign Loss. A significant issue observed in our data was class imbalance, shown in 3, which resulted in predictions biased towards the majority class. To mitigate this issue, we designed a multi-label sign loss for our model training. The loss is defined as follows:

\[
\text{sign}(p) = \begin{cases} 
  y - 2py + p^2, & |y - p| < 0.5 \\
  1, & |y - p| \geq 0.5 
\end{cases}
\]

\[
\text{Loss}(p, y) = \text{sign}(p) \times \text{BinaryCrossEntropyLoss}(p, y)
\]

For the correctly classified labels, a coefficient smaller than 1 was multiplied to the default binary cross-entropy loss. By doing so, the accumulated loss from a large number of true negative labels became smaller, and the loss from the misclassified labels became more prominent. Furthermore, the gradient of the loss changes significantly around 0.5, which enables our models to capture this change, and thus the optimal binarization threshold will also be close to 0.5 and more robust.

2.4. Multi-source data preprocessing

The raw data were sampled from different sources, which varies in sampling rate, signal amplitude, noise level, etc. To better prepare the data for model training, we adopted the following data preprocessing techniques.
Processing original data. INCART dataset was excluded from our training data since it has only 74 30-minutes records with a sampling frequency of 257 Hz and is significantly different from other datasets. All data without a label in the 27 scored classes were excluded as well. PTB dataset was downsampled from 1000 Hz to 500 Hz to make the sampling frequency of all training data unified. Since lead III, aVR, aVL, and aVF are linearly dependent on other leads and can be calculated based on Einthoven’s Law Kligfield (2002) and Goldberger’s equations Goldberger et al. (2018), these 4 leads were also excluded.

Truncating & padding. For the baseline model, all input signals were fixed at 30 seconds in length. This was done by truncating the part exceeding the first 30 seconds for longer signals and padding the shorter signals with zero. For the other ensembled model, the input length was fixed at 10 seconds with the same preprocessing method.

Wavelet denoising. To reduce the noise in ECG signals, biorthogonal wavelet transformation was applied. The numbers of vanishing moments for the decomposition and reconstruction filters were 2 and 6 respectively. The level of refinement was set to be 8.

Relabelling CPSC data. CPSC dataset was relabeled due to the fact that the labels cover only 9 classes and the class distribution is significantly different from other datasets. A baseline model was first trained on the original training set, and used for inference on the CPSC dataset. For each signal, among all the classes with inference output probability higher than 0.8, the classes that were not in the original 9 classes but in the 27 scored classes were added as a new label. To check the validity of our relabelling strategy, out of all the relabelled data with inference output probability higher than 0.95, 11 records were reviewed by a clinician. The feedback that most of the new labels were valid testified that the CPSC dataset has missing labels.

2.5. Evaluation Metrics

We followed the official evaluation metrics from PhysioNet/Computing in Cardiology Challenge 2020 Alday et al. (2020). There are 111 kinds of abnormal ECG signals, the subtle differences in the characteristics of ECG signals lead to some connection between some abnormalities. In the real-world clinical practice, the negative effects of confusing abnormal A into abnormal B may be less harmful than that of confusing abnormal A with another abnormal C. Hence, the evaluation metrics used in this competition are expected to conform to the real-world clinical practice as much as possible, thus some of the misdiagnosis that close to the ground truth should also be rewarded positively. Therefore, a weight matrix is introduced in the calculation of the confusion matrix. When abnormal A is misclassified as other different types of abnormalities, different weights between zero and one are obtained, which are provided by the organizer. Only 27 anomalies of the total 111 anomalies in 6 datasets participated in the final evaluation.

To be more specific, $C = [c_i]$ defined as the collection of our predictions. The multiclass confusion matrix is $A = [a_{ij}]$, where $a_{ij}$ is the normalized number of recordings.
Identification of 27 cardiac abnormalities from multi-lead ECG signals

| Online Test Dataset                  | Score  | Rank |
|--------------------------------------|--------|------|
| Official Validation Set              | 0.682  | 3    |
| Official Test Database 1             | 0.852  | 3    |
| Official Test Database 2             | 0.649  | 2    |
| Official Test Database 3             | 0.396  | 3    |
| The entire hidden test database      | 0.514  | 3    |

Table 2: Performance of our best online model on online datasets.

in a database that were classified as belonging to class $c_i$ but actually belong to class $c_j$, where $c_i$ and $c_j$ can be the same class or different classes. A reward metrics $W = [w_{ij}]$ is defined, where $w_{ij}$ denotes the reward for a positive classifier output for class $c_i$ with a positive label $c_j$. The unnormalized score will be calculated by equation 1.

$$Unnormalized\ Score = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} a_{ij}$$ (1)

$$Normalized\ Score = \frac{Unnormalized\ Score - Inactive\ Score}{Correct\ Score - Inactive\ Score}$$ (2)

Then this score will be normalized by equation 2 so that a classifier that always outputs the true class or classes receives a score of 1 and an inactive classifier that always outputs the normal class receives a score of 0. The Inactive.Score is the score for an inactive classifier that always outputs a normal class, while Correct.Score is the score for the model that always predicts the true class. The detailed calculation of Normalized.Score shown in *PhysioNet/Computing in Cardiology Challenge 2020 Evaluation* (2020).

2.6. Training Setup

The ensembled model was trained for 19 epochs with a batch size of 16 on a machine with 117 GB RAM, 4-core CPU, and one NVIDIA V100 GPU. The model parameters were optimized with the Adam optimizer Kingma and Ba (2015). The learning rate during training was set as 0.001, and rescheduled to 0.0001 at the 13th epoch. The optimal binarization threshold was found to be 0.36 on the offline test set.

3. Results

3.1. Online Testing Results

Table 2 shows the online evaluated challenge scores and rank on a) Official Validation Set, b) Official Test Database 1 and, c) Official Test Database 2, d) Official Test Database 3 and e) the entire hidden test database.
Table 3: Performance of different models on offline validation datasets, all the models are developed based on our baseline model, from this table we select Model 6 as our best model.

### 3.2. Offline Validation Results

Table 3 shows the offline performance of different models we have tried based on our baseline model. Model 1 is our baseline model that uses Se_ResNet as framework. In Model 2, we apply wavelet denoising, and add the relabeled CPSC data to training data, the performance of Model 2 improved in both our offline validation set and Hefei validation set, compared to our baseline model. However, we found that the problem of threshold shifting still remained. In order to stable the threshold and enhance the generalization of our model, we introduce Sign Loss to Model 3 and apply SNR Post-Processing. Though Model 3 shows an inferior performance on our offline validation set, it shows better performance on the Hefei validation set. To some extent, it can be explained that Sign Loss can improve the generalization ability of the model. In order to improve the efficiency of model training, Model 4 and Model 5 only use 8 leads signal data in 12 leads. The length of Model 4 signal input is 15000, and the length of Model 5 signal input is 5000. The training time of both models is less than that of Model 3, and the performance remains unchanged. Considering the training efficiency, model performance, and model generalization, Model 6 integrates Model 4 and Model 5, with a score of 0.683 on the offline validation set and a score of 0.319 on the Hefei validation set. The performance is better than model 4 and Model 5 on both our offline validation
set and Hefei validation set. Meanwhile, the difference between the online validation score and the offline score is only 0.001 (online: 0.682), which shows a strong ability of generalization. Finally, we select Model 6 as our best model.

3.3. Detailed Model Performance Analysis

There are 27 kinds of diseases in the original evaluation, of which 3 pairs were treated as the same cardiac abnormality when calculating the score. These 3 pairs are complete right bundle branch block (CRBBB) and right bundle branch block (RBBB), predictive atomic contract (PAC) and superventional predictive beats (SVPB), predictive viral contracts (PVC), and venturial predictive beats (VPB). Based on this mechanism, we detailed analyze our model’s performance in 24 categories. Figure 4 shows the performance of our proposed method on each cardiac abnormality, from which we are able to find the factors that affect the model’s performance, shown below.

1) **Partial label**: The AUC of each cardiac abnormality is at a high level as a whole, while the f1-score of some cardiac abnormality is at a low level. It is likely that some actual anomalies in the data are not marked out, which leads to an excessive prediction of false positive. Here are two possible reasons. Reason 1: there are 6 datasets in total, each dataset has only partial classes, and no dataset has a complete 27 classes. For example, atrial fibrillation and sinus rhythm appear in all six data sets, with complete annotation and good overall model performance. However, premature ventricular contractions and low qrs voltages only appear in two data sets, thus the model’s performance is relatively poor. Reason 2: annotation error. A certain data set has a certain cardiac abnormality annotation, but there are still many samples that are cardiac abnormalities not marked.

2) **Hard detected features**: Some features of cardiac abnormalities are hard to detect. For example, for some cases in low QRS voltages, we found that the amplitude of the signal differs, this could due to the signal have been processed with different scales in different hospitals.

3) **Feature confusion**: We also found that features between two cardiac abnormalities could be too similar for the model to classify. For example, the features of bradycardia are similar to sinus bradycardia.

As we mentioned previously, the SE module can obtain the importance of each feature by learning, and then enhance the useful features according to the importance, and suppress the features that are not useful for the current task, so the performance of the model can be improved compared with the original ResNet.

In our proposed method, we integrate two Se_ResNet, which take 8-Lead ECG data with a length of 5,000 and 15,000 as input respectively. The integrated model outperforms the two sub-models, which shows that the integrated model effectively combines the advantages of the two different input length settings. Firstly, the length of the ECG samples in the datasets ranges from 2,500 to over 100,000. The larger input length can contain more information in samples with longer signal length, while the
Identification of 27 cardiac abnormalities from multi-lead ECG signals

Figure 4: The performance of the best model on each cardiac abnormality, in terms of Area Under Curve (AUC) and F1 Score. The AUC measures the model’s ability in identifying the positive and negative samples when considering each cardiac abnormality. While the F1 Score evaluates the overall performance of multi-label classification, where different categories could affect the model’s performance. From the figure we can see that the AUC for each cardiac abnormality is relatively high, indicating that our model can classify each cardiac abnormality well. The fluctuation of the F1 score shows that different cardiac abnormalities could have similar feature space, thus confuse the model.

smaller input length can make samples with short length free from the information losses of padding in training. Secondly, some of the cardiac abnormalities show a characteristic of continuous repeat, while some occasionally appear. Hence, The larger input length can capture the intermittent abnormal signal, while the smaller input can reduce the difficulty of the neural network for anomaly detection.

4. Discussion

From the results above, our ensembled approach demonstrated its ability to classify the cardiac abnormalities despite the challenges presented, e.g. noise in the signals and labels. The score on the offline validation set is 0.683, only differed the online validation score by 0.001, suggesting good generalizability and little overfitting.

In this study, we also experimented with several other ideas. Two ideas that we would like to share and hopefully inspire further explorations are the segmentation of abnormal heartbeats.

Abnormal heartbeats segmentation via 1D U-net. Some of the cardiac abnormalities are associated with individual heartbeats. If the model can learn to recognize these individual heartbeats that directly lead to the abnormality label, it may perform and generalize better, especially when the abnormal heartbeats are rare and
Identification of 27 cardiac abnormalities from multi-lead ECG signals

| Model               | Val score | Hefei val score |
|---------------------|-----------|-----------------|
| Model1: Baseline    | 0.682     | 0.241           |
| Model6: Best model  | **0.683** | **0.319**       |
| Model7: Catboost    | 0.677     | 0.249           |
| Model8: 1D U-net    | 0.671     | 0.247           |

Table 4: The performance of the Catboost and the 1D U-net, compared with the baseline model and the best model.

Based on the preliminary analysis of the prediction results and advice from a clinician, we selected data with premature ventricular complex (PVC) labels to annotate. The raw lead-II signals were transformed into images by plotting on a grid background. The images were then imported into Colabeler, and we manually annotated all suspected PVC heartbeats by specifying their x-axis spans. Lastly, we translated the x-axis spans back to the actual locations on the signals. Due to time constraints, only 160 PVC data were annotated and used for subsequent training. We adapted U-net Ronneberger et al. (2015), a popular segmentation model in medical imaging to segment the PVC heartbeats. The 160 annotated PVC records were used as positive samples and 500 randomly selected records without PVC labels were used as negative samples to train the model.

**Feature engineering with Catboost.** Deep learning model can only automatically learn abstract features according to given tags, while feature engineering can be a supplement to abstract features. We use biosppy software package to extract r-peaks of ECG signals. Based on the extracted r-peaks, we then use the hrvanalysis software package to extract 32-dimensional features, which including 16-dimensional time-domain features and 7-dimensional frequency-domain features. Age and gender are added as features as well.

Table 4 shows the performance of two models we have attempted, compared to our baseline model. Model 7 combines the engineered features of the ECG signal with the deep features from the neural network, and train a binary classifier for each cardiac abnormality with Catboost. The offline score of the model is 0.677. The advantage of Model 7 is that it uses the engineered features of the ECG signal, which helps to classify certain cardiac abnormalities more accurately. The disadvantage is that the framework of the model is complex and the extraction of ECG signal features takes a long time. It requires too much computing resources and its time consuming, thus we have only tested this framework on our offline validation set.

Model 8 introduces U-net to improve the classification performance for PVC. The PVC classification is considered positive if there is any positive PVC signal output. When we incorporated U-net into our system, the PVC predictions were solely determined by U-net. Our experiments showed that incorporating U-net increased the
F\textsubscript{beta} and G\textsubscript{beta} measures in the evaluation metrics, but no challenge score improvement was observed. Due to the additional training time required, we did not incorporate U-net in our final system.

In the future, we will try to improve our work in the following aspects. The first is to combine ECG medical knowledge with deep learning to construct a classification model. For example, machine learning methods are used to extract medical features such as RR interval instead of traditional methods that are not accurate enough. These features can be integrated with the features extracted from Se_ResNet. Meanwhile, medical knowledge can effectively improve the interpretability and generalizability of the model. The second is to develop a better abnormal signal location model. Many ECG abnormalities occur intermittently, it is a challenging task to locate the time period of abnormal signal in a long period of ECG data. If the abnormal signal location function is introduced, the classification model can extract the features of abnormal signals more effectively, and maintain good performance on ECG data with different lengths.

5. Conclusions

In this paper, we proposed a deep learning framework to automatically identify multiple cardiac abnormalities. Compared to previous works, the main contribution of our methods is three-fold. Firstly, our proposed framework is able to classify 27 number of cardiac abnormalities on 12-lead ECG signals, while previous works focused on at most 9 number of cardiac abnormalities, and mainly using single lead ECG signals. Secondly, we introduce a Sign Loss to solve the class imbalance problem and thus improve the generalizability of the model. Thirdly, our framework is developed on 6 different datasets across the world, and we proposed a number of preprocessing methods to eliminate the diversity of data from a different source, while previous works mainly use a small dataset from a single source. Since the framework is developed on real-life datasets, we believe that it has great potential to be deployed in the actual clinical practice.

Acknowledgments

This work is partially supported by the National University of Singapore Start-up grant (R-608-000-172-133).

References

Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S. et al. (2020). Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020, Physiological Measurement.

Alexakis, C., Nyongesa, H., Saatchi, R., Harris, N., Davies, C., Emery, C., Ireland, R. and Heller, S. (2003). Feature extraction and classification of electrocardiogram (ecg) signals related to hypoglycaemia, Computers in Cardiology, 2003, IEEE, pp. 537–540.
Andreotti, F., Carr, O., Pimentel, M. A., Mahdi, A. and De Vos, M. (2017). Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg, 2017 Computing in Cardiology (CinC), IEEE, pp. 1–4.

Artis, S. G., Mark, R. and Moody, G. (1991). Detection of atrial fibrillation using artificial neural networks.

Bickerton, M. and Pooler, A. (2019). Misplaced ecg electrodes and the need for continuing training, British Journal of Cardiac Nursing 14(3): 123–132.

Billeci, L., Chiarugi, F., Costi, M., Lombardi, D. and Varanini, M. (2017). Detection of af and other rhythms using rr variability and ecg spectral measures, 2017 Computing in Cardiology (CinC), IEEE, pp. 1–4.

Bin, G., Shao, M., Bin, G., Huang, J., Zheng, D. and Wu, S. (2017). Detection of atrial fibrillation using decision tree ensemble, 2017 Computing in Cardiology (CinC), IEEE, pp. 1–4.

Chen, Y., Wang, X., Jung, Y., Abedi, V., Zand, R., Bikak, M. and Adibuzzaman, M. (2018). Classification of short single-lead electrocardiograms (ecgs) for atrial fibrillation detection using piecewise linear spline and xgboost, Physiological measurement 39(10): 104006.

Gaziano, T. A., Bitton, A., Anand, S., Abrahams-Gessel, S. and Murphy, A. (2010). Growing epidemic of coronary heart disease in low- and middle-income countries, Current Problems in Cardiology 35(2): 72 – 115. Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries.

URL: http://www.sciencedirect.com/science/article/pii/S0146280609001273

Goldberger, A. L., Goldberger, Z. D. and Shvilkin, A. (2018). Goldberger's clinical electrocardiography: a simplified approach, Elsevier.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hu, J., Shen, L. and Sun, G. (2018). Squeeze-and-excitation networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Huang, C., Zhao, R., Chen, W. and Li, H. (2019). Arrhythmia classification with attention-based res-bilstm-net, Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, Springer, pp. 3–10.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization, in Y. Bengio and Y. LeCun (eds), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

URL: http://arxiv.org/abs/1412.6980

Kligfield, P. (2002). The centennial of the Einthoven electrocardiogram, Journal of Electrocardiology 35(4): 123–129.
REFERENCES

Kligfield, P., Gettes, L. S., Bailey, J. J., Childers, R., Deal, B. J., Hancock, E. W., Van Herpen, G., Kors, J. A., Macfarlane, P., Mirvis, D. M. et al. (2007). Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology a scientific statement from the American Heart Association electrocardiography and arrhythmias committee, council on clinical cardiology; the American college of cardiology foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology, Journal of the American College of Cardiology 49(10): 1109–1127.

LeCun, Y., Bengio, Y. et al. (1995). Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361(10): 1995.

Luo, C., Jiang, H., Li, Q. and Rao, N. (2019). Multi-label classification of abnormalities in 12-lead ecg using 1d cnn and lstm, Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, Springer, pp. 55–63.

Mahmoodabadi, S., Ahmadian, A. and Abolhasani, M. (2005). Ecg feature extraction using daubechies wavelets, Proceedings of the fifth IASTED International conference on Visualization, Imaging and Image Processing, pp. 343–348.

Martínez, J. P., Almeida, R., Olmos, S., Rocha, A. P. and Laguna, P. (2004). A wavelet-based ecg delineator: evaluation on standard databases, IEEE Transactions on biomedical engineering 51(4): 570–581.

Minami, K.-i., Nakajima, H. and Toyoshima, T. (1999). Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network, IEEE transactions on Biomedical Engineering 46(2): 179–185.

Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm, IEEE Transactions on Biomedical Engineering BME-32(3): 230–236.

PhysioNet/Computing in Cardiology Challenge 2020 Evaluation (2020).
  URL: https://github.com/physionetchallenges/evaluation-2020/blob/master/evaluate_12ECG_score.py

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation, Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 p. 234–241.

Sodmann, P., Vollmer, M., Nath, N. and Kaderali, L. (2018). A convolutional neural network for ecg annotation as the basis for classification of cardiac rhythms, Physiological measurement 39(10): 104005.

TIANCHI - Hefei Hi-tech Cup ECG Intelligent Competition (2020).
  URL: https://tianchi.aliyun.com/competition/entrance/231754/introduction

Vafaie, M., Ataei, M. and Koofigar, H. (2014). Heart diseases prediction based on ecg signals' classification using a genetic-fuzzy system and dynamical model of ecg signals,
Biomedical Signal Processing and Control 14: 291 – 296.
URL: http://www.sciencedirect.com/science/article/pii/S1746809414001311

Wang, C., Yang, S., Tang, X. and Li, B. (2019). A 12-lead ecg arrhythmia classification method based on 1d densely connected cnn, Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, Springer, pp. 72–79.

Warrick, P. A. and Homsi, M. N. (2018). Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection, Physiological measurement 39(11): 114002.

Xia, Z., Sang, Z., Guo, Y., Ji, W., Han, C., Chen, Y., Yang, S. and Meng, L. (2019). Automatic multi-label classification in 12-lead ecgs using neural networks and characteristic points, Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, Springer, pp. 80–87.

Xiong, Z., Nash, M. P., Cheng, E., Fedorov, V. V., Stiles, M. K. and Zhao, J. (2018). Ecg signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network, Physiological measurement 39(9): 094006.