Multi-Label Cardiac Abnormalities Classification on Selected Leads of ECG Signals

Zhuoyang Xu, Yangming Guo, Tingting Zhao, Zhuo Liu, Xingzhi Sun

Ping An Technology, Beijing, China

Abstract

As part of the PhysioNet/Computing in Cardiology Challenge 2021, Our team, HeartBeats, developed an ensembled model based on SE-ResNet for identifying 30 kinds of cardiac abnormalities from different lead combinations of electrocardiograms (ECGs). At pre-processing stage, ECGs were down-sampled to 500 Hz and each record is normalized using Z-Score normalization. We then employed several residual neural network modules with squeeze-and-excitation blocks to learn from the first 15-second segments of the signals. We designed a multi-label loss to emphasize the impact of wrong predictions during training. We relabelled the dataset which contains only 9 classes using our baseline model build in last year's challenge. Five-fold cross-validation was used to assess the performance of our models. Our classifiers received the scores of 0.58, 0.55, 0.56, 0.53, and 0.53 for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions with the Challenge evaluation metric. Our final model performed well on the test data. However, the results were not officially ranked because our training code may select the offline pre-trained models rather than using the training data if the pre-trained models performed better than the trained models on the training data. The model can therefore fail to learn from new training data.
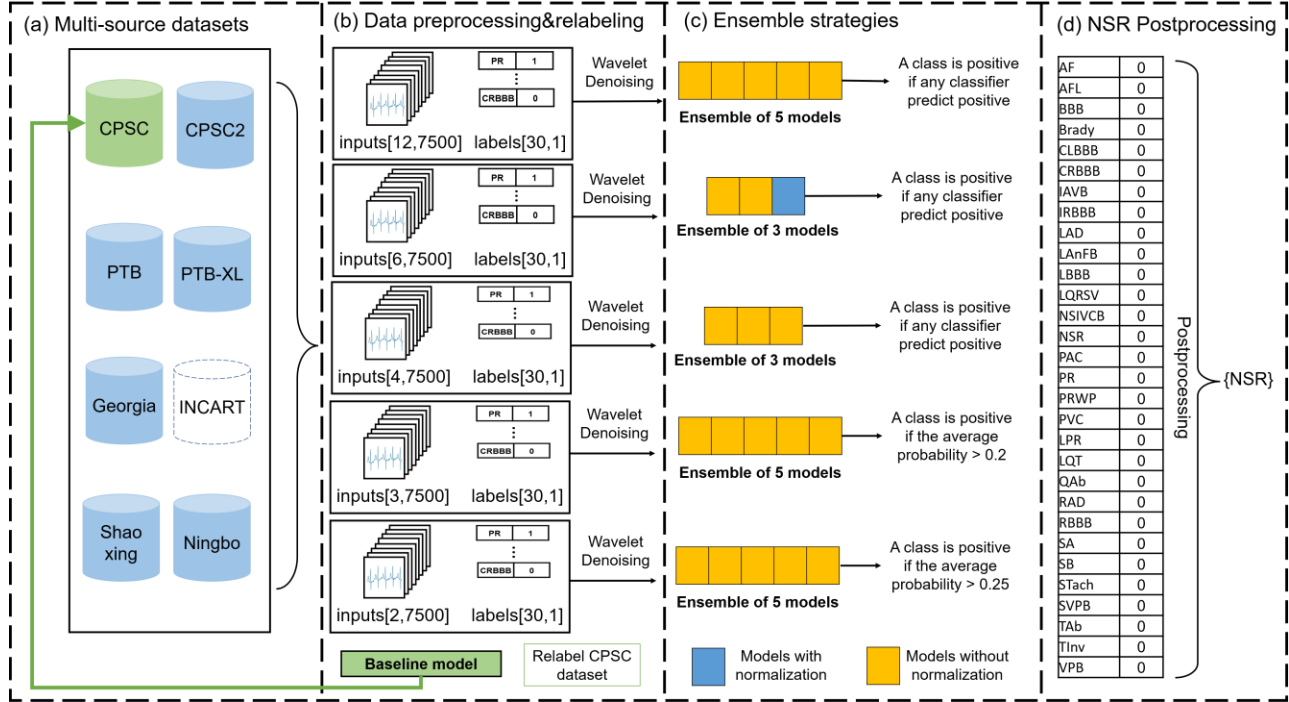
1. Introduction

Cardiovascular disease is one of the primary causes of death globally[1]. Early diagnosis of cardiac diseases may be helpful in preventing premature complications and deaths by enabling successful interventions. 12-lead ECG is a cheap and non-invasive tool to represent the electrical activity of the heart. It is commonly used in clinical care to discern cardiac abnormalities such as atrial fibrillation, bradycardia, arrhythmias and so on. Up to now, interpreting ECG still relies on human efforts, which are time-consuming and error-prone. In addition, the development of ECG devices leads to the rapid growth of recording volume of ECG data. As a result, automatic detection and classification of ECG abnormalities become necessary to reduce the working pressure of physicians and improve the accuracy of diagnoses.

The PhysioNet/Computing in Cardiology Challenge 2021[2] is an extension of last year's Challenge [3] in which the task is to classify cardiac abnormalities from twelve-lead ECGs. This year the participants are asked to build the models that can classify cardiac abnormalities from not only twelve-lead, but designated six-lead, four-lead, three-lead, and two-lead ECGs. The details about the challenge including rules, data and metrics are described in[2].

In last year's Challenge, we proposed a deep learning framework to automatically identify 27 types of cardiac abnormalities using standard 12-lead ECG signal [4], which placed us 3rd out of 40 teams in the official ranking. Our model outperformed the top two teams [5][6] in the hidden validation and two test sets, but fall behind them in the third test set. Considering the relationship between the two competitions and the unfavorable performance in the third test set, we followed the main framework of our last year's approach but made some adjustments. Our focus this year is to improve the generalizability of the model and tune the recall threshold of the model to achieve higher Challenge scores. We tackle the problem by a variety of strategies, such as threshold adjustment, different ensemble approaches for different selected-lead data, signal normalization and so on.

2. Method

Our objective was to propose models that could accurately classify 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead ECG recordings into multiple classes of 30 clinical diagnoses. In this section, we will describe in detail our approaches in data preprocessing, model construction and postprocessing. The framework design of the proposed method is seen in **Figure 1**. We first introduce the multi-source datasets we used. Then we present the data preprocessing techniques that reduce the data divergence, the Sign Loss was introduced to solve the class imbalance problem, and the design of the ensemble models for each reduced-lead ECG signals that enables robustness of the classifications. Finally, we present our training setups.

Figure 1. The framework design of the proposed method. **(a)** The 7 datasets used in training stage. i.e. data from the China Physiological Signal Challenge 2018 (CPSC and CPSC2), the Physikalisch-Technische Bundesanstalt databases (PTB and PTB-XL), the Georgia 12-lead ECG Challenge (Georgia) database, Chapman University, Shaoxing People's Hospital (Shaoxing) and Ningbo First Hospital (Ningbo) database. St.Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART) was excluded from our training data. **(b)** Data pre-processing and relabeling CPSC dataset: the input length was fixed at 7500 and different reduced-lead ECG are made as the inputs of the models after wavelet denoising. CPSC dataset is relabeled by our baseline model in last year's Challenge. **(c)** Ensemble strategies: different ensemble policy for different reduced-lead data. For example, 5 models without data normalization (Z-Score normalization) were ensembled for 12-lead ECG data, a class is positive if any classifier predict positive. **(d)** NSR Post-processing: signals that were predicted to be negative for all classes in the final result would be classified as positive for sinus rhythm (NSR).

2.1. Datasets & Labelling

The public challenge training data consists of 88,253 12-lead ECG signals from eight different datasets. The recordings were of varying frequency of the range of 257 Hz - 1000 Hz and the length of the signals varies from 6 s to 30 min. There are 133 abnormalities in the eight dataset, 30 types of abnormalities are considered in the challenge scoring metrics.

Processing original data.

INCART dataset was excluded from the training data since it has only 74 30-minutes records with a sampling frequency of 257 Hz and is significantly different from other datasets. Data without a label in the 30 scored classes were excluded as well. Signals were re-sampled to have a sampling rate of 500 Hz to make the sampling frequency of all training data unified. In the rest of the data, we randomly split 80% as the training set and 20% as the offline test set. The final sizes of the training set and test set are 65,765 and 16,445 respectively.

Relabelling CPSC data.

There are only 9 types abnormalities in CPSC dataset. It

was found that there was indeed label omission in CPSC data after manually checked by physician. A baseline model[4] trained in last year's Challenge (the baseline model in **Figure 1**.) was used for inferring pseudo labels on the CPSC dataset. For each ECG signal in the CPSC dataset, inferred pseudo labels were added as new labels if (1) the inference output probability was higher than 0.8, (2) the labels not in the original nine labels, and (3) the labels were in the 30 officially scored labels. The re-labeled CPSC will be replace the original CPSC and be used as a part of training dataset. Pseudo labels added to CPSC dataset is shown in **Table 1**.

Class	Recordings	Class	Recordings
AFL	3	LQT	9
Brady	49	SA	2
CRBBB	1	SB	14
IRBBB	4	STach	122
LAD	5	SVPB	57
LAnFB	2	TAb	13
LQRSV	1	VPB	6
PVC	46		

Table 1. Pseudo labels added to CPSC dataset.

2.2.   Data Preprocessing

Truncating & padding.
Recordings were extract as the length of T=15 seconds, 7500 points. This was done by truncating the part exceeding the first T seconds for longer signals and applying zero-padding, if the sequence length is less than T seconds.

Normalization.
Each recording is normalized using Z-Score normalization and then multiply 100 so that each channels' signal lies within the range of -100 to 100.
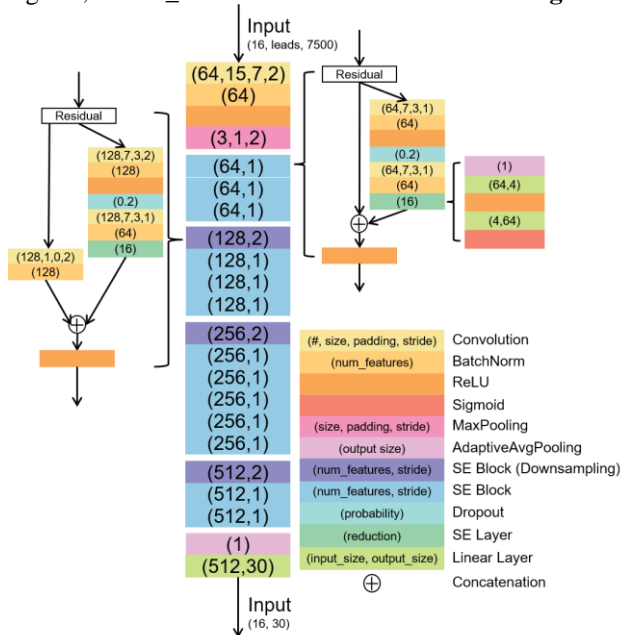
Wavelet denoising.
Biorthogonal wavelet transformation (bior2.6) was applied to reduce the noise in ECG signals. The numbers of vanishing moments for the decomposition and reconstruction filters were 2 and 6, respectively. The level of refinement was set to be 8, where the high-frequency coefficients in level 1, level 2, and level 8 were set to zero.

2.3.   System Architecture

SE-ResNet.
One important feature of the 12-lead ECG signal is that the information contained differs in different leads due to the difference in signal voltage intensity and amplitude variation. Different ECG abnormalities may be more apparent in specific leads. An equal importance of different leads could cause information losses, leading to misdiagnosis. SE_ResNet [7] is good at capturing the distinctive information in each of the multi-leads ECG signals, the SE_ResNet architecture is shown in **Figure 2**.

Figure 2. Architecture of the SE_ResNet model. The parameter *leads* varies according to reduced-lead ECG data, which can be 12, 6, 4, 3 and 2 for 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead

versions data.

Sign Loss.
Sign Loss function was designed to solve class imbalance problem, inspired by Sun et al [8]. The improved multi-label Sign Loss is defined as below:

$$Sign(p) = \begin{cases} y - 2py + p^2, & |y - p| < 0.5 \\ 1, & |y - p| \geq 0.5 \end{cases} \quad (1)$$

$$Loss = \sum_{i=1}^{30} Sign(p_i) \times BCE(p_i, y_i) \quad (2)$$

where y denotes the ground truth and p denotes the model's estimated probability for y = 1, BCE indicates Binary Cross Entropy Loss.

NSR Postprocessing.
If all the classes are predicted as negative, we will make the sinus rhythm (NSR) to be positive. That's because the signal contains at least one type of abnormality and NSR accounts for a high proportion of the dataset. This setting allows us to get a positive challenge score instead of 0 when all classes are negative.

2.4.   Model Ensemble & Threshold Adjustment

In order to improve the robustness of the classifications, we created different ensemble strategies for different reduced-lead ECG data and models were trained via five-fold cross validation. For 12-lead ECG data, 5 models without data normalization were ensembled, ECGs were classified positive if any of the models predicted positive. For 6-lead ECG data, 2 models without data normalization and 1 model with data normalization were ensembled, ECGs were classified in the same way like 12-lead models. Specially, the model with data normalization not predict low qrs voltages and poor R wave Progression since data normalization will destroy features of these two abnormalities. For 4-lead ECG data, 3 models without data normalization were ensembled, ECGs were classified in the same way like 12-lead models, too. For each model ensembled in 12-lead, 6-lead and 4-lead models, the threshold is 0.35. For 3-lead and 2-lead ECG data, 5 models without data normalization were ensembled, ECGs were classified according the average probabilities of the 5 models, and the threshold were 0.2 and 0.25, respectively.

2.5.   Training Setup

The proposed model was trained with a batch size of 32 for 17 epochs as the validation loss was not further decreasing. The model parameters were optimized with the Adam optimizer Kingma and Ba [9]. During training, the learning rate was set as 0.001 and rescheduled to 0.0001 at the 15th epoch.

3.   Results

Our classifiers received scores of 0.58, 0.55, 0.56, 0.53, and 0.53 for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions with the Challenge evaluation metric (see **Table 2**).

Leads	Training	Validation	Test
12	0.725	0.669	0.58
6	0.700	0.647	0.55
4	0.685	0.650	0.56
3	0.628	0.594	0.53
2	0.667	0.615	0.53

Table 2. Challenge scores for our final selected entry (team *HeartBeats*) using 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

4.   Discussion

In offline test set, our model performed constantly well on AFL, CRBBB, NSR, PR, SB and Stach. The most serious problem observed was, classification of some abnormalities reached high AUC score but low F-1 score. For these abnormalities, there were many false-positives – labelled 0 although both the model and the human judged as 1. It was an interesting question to construct well generalized models from the partial labelled datasets.

Performance of LQRSV and PRWP, were significantly destroyed by the data normalization technique. Their features may be closely related to the signal scale. However, data normalization benefited LPR and was expected to reach more robust prediction to face different situations in real world. PRWP and IRBBB performed much worse in 2-lead models than 12-lead models. There seemed to be important features indicating these abnormalities besides the 2-lead signals.

The overall performance of 6-lead models was worse than 4-lead models. The most likely cause of this observation was lead V2 was much more important and contained much more potential features than aVR, aVL, aVF leads for abnormalities identification.

5.   Conclusions

In this paper, we developed a deep neural network architecture and ensemble models for multi-label classification of cardiac abnormalities from 12-lead and reduced-lead ECGs. In order to improve the robustness of the classifications, we adopted signal normalization, threshold adjustment, different ensemble approaches for different reduced-lead data and introduced Sign Loss for tracking class imbalance problem. Our final models performed well on the test data. However, the results were not officially ranked because our training code may select the offline pre-trained models rather than using the training data if the pre-trained models performed better than the trained models on the training data. The model can therefore fail to learn from new training data. Another possible reason of not being ranked is that we relabelled some data, which can prevent the training code from being repeated.

References

[1] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart Disease and Stroke Statistics – 2021 Update: a Report from the American Heart Association. Circulation 2021;143(8):e254–e743.

[2] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Li Q, Sharma A, Clifford GD. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021; 48: 1-4.

[3] Perez EA, Gu A, J Shah A, Robichaux C, Ian Wong AK, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas. 2021 Jan 1;41(12):124003. doi: 10.1088/1361-6579/abc960. PMID: 33176294; PMCID: PMC8015789.

[4] Zhu Z, Lan X, Zhao T, Guo Y, Kojodjojo P, Xu Z, et al. Identification of 27 Abnormalities from Multi-lead ECG Signals: an Ensembled SE_ResNet Framework with Sign Loss function. Physiol Meas. 2021 Jun 29;42(6). doi: 10.1088/1361-6579/ac08e6. PMID: 34098532.

[5] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification. Computing in Cardiology 2020; 47:1-4.

[6] Zhao Z, Fang H, Relton S, Yan R, Liu Y, Li Z, et al. Adaptive Lead Weighted ResNet Trained with Different Duration Signals for Classifying 12-lead ECGs. Computing in Cardiology 2020; 47:1-4.

[7] Hu J, Shen L, Albanie S, Sun G and Wu E. Squeeze-and-Excitation Networks. in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[8] X. Sun et al. System-level Hardware Failure Prediction Using Deep Learning. 2019 56th ACM/IEEE Design Automation Conference (DAC), 2019, pp. 1-6.

[9] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In Bengio Y, LeCun Y (eds.), 3rd International Confer- ence on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Address for correspondence:

Xingzhi Sun
3 Xinyuan Road, Chaoyang District, Beijing, China
xingzhi_sun@hotmail.com