# Classifying different dimensional ECGs using deep residual convolutional neural networks

Wenjie Cai, Fanli Liu, Xuan Wang, Bolin Xu, Yaohui Wang

University of Shanghai for Science and Technology, Shanghai, China

## Abstract

*Introduction: The electrocardiogram (ECG) is the most common diagnostic tool for screening cardiovascular diseases. PhysioNet/Computing in Cardiology Challenge 2021 aims to classify cardiac abnormalities from twelve-lead, six-lead, four-lead, three-lead, and two-lead ECGs. Methods: ECGs were downsampled to 250 Hz and then applied with a bandpass filter to reduce noise. The unscored label named ventricular ectopics was transformed to premature ventricular contractions. The ECGs labled as af in the Ningbo Database were relabeled as afl or af. All ECGs were randomly shuffled and divided into a training set and a validation set at 4:1. Five models based on a deep residual convolutional neural network were proposed to make classification from different dimensions of ECGs. A novel loss calculation method was proposed to balance the different labeling tendency of different source data sets. Results: After training, the performance of five models was evaluated on the local validation set and got the challenge metric of 0.610, 0.595, 0.610, 0.612, and 0.589 on twelve-lead, six-lead, four-lead, three-lead, and two-lead ECGs, respectively. Our team, USST_Med, received a test score of 0.597, 0.583, 0.536, 0.552, and 0.535 on five test datasets, respectively. Conclusion: The proposed models performed well on classifying ECGs and have potential for clinical application.*

## 1. Introduction

Cardiovascular disease is responsible for most deaths worldwide in the world [1]. Early detection and diagnosis are of great significance for reducing mortality. Electrocardiography is the most popular and non-invasive method for screening heart disease. Standard electrocardiogram (ECG) has 12 leads including 6 limb leads and 6 chest leads showing heart electrical activity transmission in the coronal plane and the transverse plane. Cardiologists check the ECG lead by lead and beat by beat to make a diagnostic conclusion. ECG diagnosis is a time-consuming technical task, and usually doctors need years of training. Even so, high-intensity diagnostic workloads are prone to misdiagnosis. With the rise of wearable devices, more and more electrocardiography devices have entered daily life. Many of them can achieve 24-hour ECG monitoring, so a large number of ECGs are generated at all times. It is unrealistic to rely solely on doctors for diagnosis. Fully automatic diagnosis can greatly reduce the workloads of doctors and is a useful supplement to manual diagnosis. With the development of artificial intelligence, many automatic ECG classification algorithms emerge [2-5]. It is reported that the method based on deep residual neural network can surpass the cardiologists in single-lead ECG classification [5]. However, different ECG acquisition devices, different placement positions of leads, ethnic differences, and geographic differences will challenge the robustness of automatic diagnosis algorithms.

Some wearable electrocardiographs can only produce ECGs with a limited number of leads. Whether the information contained in these ECGs can be comparable to standard 12-lead ECGs, and whether 2 or 3 leads can meet the basic clinical needs, are interesting questions to be explored. The PhysioNet/Computing in Cardiology Challenge 2021 [6, 7] aims to classify cardiac abnormalities from 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead ECGs. This study attempts to develop robust and high-precision algorithms for ECG classification.

## 2. Methods

The overall flow chart of this study is shown in Fig. 1. Briefly, the data are preprocessed to have the same resolution and remove noises. Some labels that have the same scoring weights or medical significance were merged together. The deep learning model designed for 12-lead data was trained for the first round and then evaluated all data to compare the results and the labels. The recordings with high probabilities of wrong labels were screened out to create the mask. The model was retrained on the training set and the optimal threshold was determined on the validation set.

### 2.1. Datasets

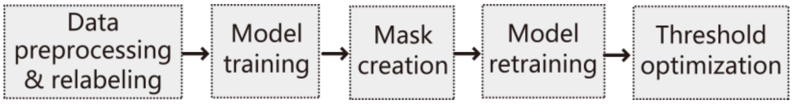

Figure 1. The overall flow chart of this study.

The public training set contains 88,253 recordings from 8 databases including the China Physiological Signal Challenge (CPSC) 2018, the St. Petersburg INCART 12-lead Arrhythmia Database, the Physikalisch-Technische Bundesanstalt (PTB) database, PTB-XL database, the Georgia 12-lead ECG Challenge (G12EC) Database, the Chapman-Shaoxing database and the Ningbo database. Among them, INCART database has 74 recordings and PTB database has 516 recordings. These two databases have fewer recordings and longer sampling time compared with other databases. And they provide very limited labels that are valuable for scoring. Considering the complexity of data processing and the performance of the model, we discarded these two databases. The remaining 87,663 recordings were used as the local training set and validation set, with a ratio of 4 to 1.

## 2.2. Data preprocessing

ECGs were loaded with SciPy module in Python. For each lead, data values were divided by the amplitude resolution and subtracted by the mean value of that lead. These values then represent the voltage in mV. Values greater than 20 mV indicate abnormal recorded signals. They exist in some recordings in CPSC database. These outliers appear like spikes, and were replaced by the normal values next to them [8]. Data were then downsampled to 250 Hz with fast Fourier transformation and filtered by an FIR bandpass filter with bandwidth between 0.5 Hz and 45 Hz.

## 2.3. Data relabeling

There are 30 of 133 diagnoses used for evaluating challengers' algorithms. According to the scoring algorithm provided by the challenge organizer, some labels are considered as the same diagnosis. So we merged labels Complete right bundle branch block (CRBBB) and Right bundle branch block (RBBB), labels Complete left bundle branch block (CLBBB) and Left bundle branch block (LBBB), labels Premature atrial contraction (PAC) and Supraventricular premature beats (SVPB), and labels Premature ventricular contractions (PVC) and Ventricular premature beats (VPB). Ventricular ectopics (VEB) is not scored in the scoring algorithm, but we think it has the same medical meaning with VPB or VEB. So we merged VEB with VPB and PVC.

There are 4 diagnoses that have fewer than 700 ECG recordings, namely Bundle branch block (BBB), Bradycardia (Brady), Poor R wave progression (PRWP) and Prolonged pr interval (LPR). We removed them from the labels of all recordings. Thus there are 22 classes to be classified.

In Ningbo database, 7,615 recordings are labelled as Atrial flutter (AFL) whereas none is labelled as Atrial fibrillation (AF). Local cardiologists believe that most of them are wrong. To relabel these recordings, we constructed a binary classification model and trained this model with all data except the data from Ningbo database. Then we predicted the AFL recordings of Ningbo database with the trained model and relabelled these recordings with predicted results.

In CPSC database, there are only 6 diagnoses belong to scoring labels along with VEB. We believe that these diagnoses are incomplete. To supplement the missing labels, we constructed a deep learning model that can classify the remaining 15 scoring diagnoses and trained the model with all data except CPSC data. Then all recordings in CPSC database were predicted with the trained model. If the inference result showed that the probability of one class was greater than 0.8, the new diagnosis was added to the corresponding labels.

## 2.4. Deep learning models architecture

This year's challenge is to classify ECGs of five different numbers of leads. We built five models that share the same architecture except for the input layer that has different shapes. The main architecture consists of one CNN layer, one max pooling layer, eight Residual blocks, one global max pooling layer and two fully connected layers (Fig. 2). Each Residual block used SENet (Squeeze-and-Excitation Networks) [9] to get channel attention and used Mish as the activation function. Sigmoid was used as the activation function in the last layer. The dimension of the model input is consistent with the number of leads, and the length is set to be variable to adapt to different lengths of ECG. The 4 discarded classes are always set to 0.

## 2.5. Model training and the loss functions

The models were trained using Keras with TensorFlow as the backend. Adam was selected as the optimizer. The batch size was set to 256. Warm start training strategy was applied. In detail, the learning rate was set to 0.0001 in the first two epochs, and then set to 0.001 for other epochs. Early stopping strategy was used to prevent model overtraining. The training process stopped when the loss

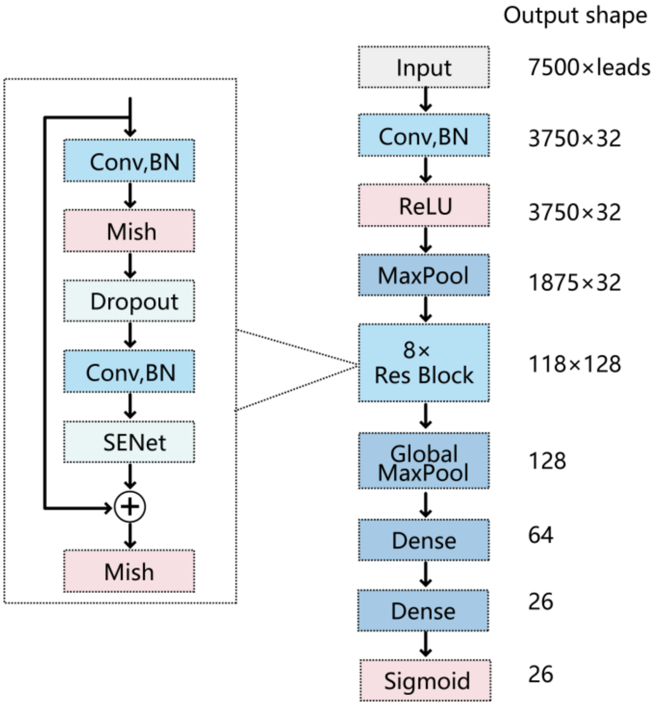

Figure 2. The architecture of the deep learning model.

hadn't decreased for 4 epochs.

The models have gone through two rounds of training. The focal loss [10] was used in the first round of training. It's equation is shown as follows:

$$L_{focal} = \begin{cases} L_+ = (1-p)^2 * \log(p) \\ L_- = 0.1 * p^2 * \log(1-p) \end{cases} \quad (1)$$

Where $p$ is the model's output probability, $L_+$ is the loss for positive labels and $L_-$ is the loss for negative labels.

All recordings were fed into the deep learning model and the outputs were compared with the corresponding labels to generate masks. In detail, the masks had the same dimensions with the labels and were initially set to 1 for all values. The model's output denotes the probabilities for 26 classes of one recording. If the model predicted that one recording was classified as one class with the probability greater than 0.9 whereas this class was not in the labels, we set the mask for this class to 0. Similarly, when the model predicted that one recording was classified as one class with the probability less than 0.1 whereas this class was in the labels, we set the mask for this class to 0, too.

The second round of training used the modified asymmetric loss [11] which discarded the loss where the mask value was 0. The loss function is shown as follows:

$$P_{neg} = \max(p - 0.05, 0) \quad (2)$$

$$P_{pos} = 1 - \max(0.99 - p, 0) \quad (3)$$

Table 1. Performance of proposed 12-lead model on local validation set.

| Thresholds | Challenge Score |
|---|---|
| 0.5 | 0.536 |
| optimized | 0.610 |

Table 2. Performance of proposed models on the local validation set and official validation set.

| Model | Challenge Score | |
|---|---|---|
| | Local validation set | official validation set |
| 12-lead | 0.61 | 0.597 |
| 6-lead | 0.595 | 0.583 |
| 4-lead | 0.61 | 0.536 |
| 3-lead | 0.612 | 0.552 |
| 2-lead | 0.589 | 0.535 |

$$L = \begin{cases} L_+ = m * \left(1 - P_{pos}\right) * \log\left(P_{pos}\right) \\ L_- = m * P_{neg}^2 * \log\left(1 - P_{pos}\right) \end{cases} \quad (4)$$

Where $P_{neg}$ is the output probability when the true label is 0, $P_{pos}$ is the output probability when the true label is 1, and $m$ is the mask.

## 2.6. Model inference

The ECG recordings were pre-processed as described at section 2.1. Since Our models were designed with adaptive input length, the ECG data can be directly fed into the deep learning models without any segmentation. Because the training data are extremely imbalanced, the optimal thresholds for each class are different. We performed two step search that was proposed by Zhao et al.[12]. First, we set the thresholds of all classes to the same value, ranging from 0.1 to 0.9, with an increase of 0.1 each time, and determined the optimal threshold after calculating the scores respectively. Second, only the threshold of one class was changed from 0.2 to 0.8, with an increase of 0.01 each time. The thresholds of other classes used the best thresholds determined in the first step. After calculating the scores separately, the threshold corresponding to the highest score was the optimal threshold for that class.

## 3. Results

After models training, we evaluated model performance on our own validation set. In order to make the challenge score generated on the validation set consistent with the official test set as much as possible, we only selected the part of the validation set that belongs to the G12EC database. As shown in Table 1, when the thresholds were fixed at 0.5, the challenge score for 12-lead model was 0.536. After thresholds optimization, the challenge score reached 0.610 which was 0.074 higher than before. And the challenge scores for 6-lead, 4-lead, 3-lead, 2-lead models on the validation set were 0.595, 0.610, 0.612 and 0.589, respectively (Table 2).

These models' performance was further tested on the hidden validation set. As shown in Table 2, the five models for different leads ECGs achieved the challenge scores of 0.597, 0.583, 0.536, 0.552, and 0.535, respectively. The first two scores were under optimal thresholds. And the other three scores were with the fixed thresholds set at 0.5.

## 4. Discussion and Conclusions

The results shown in Table 2 indicate that the challenge scores of our models on the local validation set and the official online validation set are very close, especially for the 12-lead models and 6-lead models. It must be pointed out here that our ideal code did not run successfully online due to some technical reasons, so the version of the code used in the competition is sub-optimal. The thresholds used for 4-lead, 3-lead and 2-lead models in this version are set to 0.5, which are apparently not optimal. While other thresholds including the ones used in local validation set are all optimized. So the online results of models with less than 6 leads are much lower than the results of 12-lead model and 6-lead model.

The results show that the model trained with 12-lead data has the best performance. The results of the models trained with 4-lead and 3-lead data are comparable to the results of the 12-lead model, while the results of the model trained with 6-lead and 2-lead data are slightly worse. It could be explained by the leads positions. As we mentioned in the introduction, the standard 12-lead ECG has 6 limb leads and 6 chest leads. The 12-lead, 4-lead and 3-lead data used in this competition contains at least one limb lead and one chest lead, whereas the 6-lead and 2-lead data contains limb leads only. The hybrid combination can evaluate the ECG from different angles and help the correct classification of the ECG.

There are several limitations in our study. Firstly, the effects of different loss functions need to be studied systematically. The proposed loss function could be further tuned. Secondly, the submitted code is not the best version, which affects the further analysis of the results. Thirdly, the role of this mask in model performance needs further exploration.

In conclusion, the proposed deep learning models and data processing method showed great potential for clinical application in automatically classifying ECGs.

## References

[1] S. S. Virani *et al.*, "Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association," *Circulation*, pp. e1-e458, Jan 29 2020.

[2] L. L. Yang, J. J. Zhu, T. H. Yan, Z. Y. Wang, and S. S. Wu, "A Modified Convolutional Neural Network for ECG Beat Classification," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 3, pp. 654-660, Mar 2020.

[3] J. X. Cai, W. W. Sun, J. F. Guan, and S. You, "Multi-ECGNet for ECG Arrythmia Multi-Label Classification," *Ieee Access*, vol. 8, pp. 110848-110858, 2020.

[4] S. S. Xu, M. Mak, and C. Cheung, "Towards End-to-End ECG Classification With Raw Signal Extraction and Deep Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1574-1584, 2019.

[5] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat Med*, vol. 25, no. 1, pp. 65-69, Jan 2019.

[6] G. A. Perez Alday EA, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA., " Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Under Review*, 2020.

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

[8] W. Cai and D. Hu, "QRS Complex Detection Using Novel Deep Learning Neural Networks," *IEEE Access*, vol. 8, pp. 97082-97089, 2020.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2019.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. M. He, and P. Dollar, "Focal Loss for Dense Object Detection," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, Feb 2020.

[11] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, and L. Zelnik-Manor, "Asymmetric Loss For Multi-Label Classification," 2020.

[12] Z. Zhaowei *et al.*, "Classification of Cardiac Abnormalities From ECG Signals Using SE-ResNet," *2020 Computing in Cardiology (CinC)*, Conference Paper pp. 4 pp.-4 pp., 2020 2020.

Address for correspondence:

Wenjie Cai
School of Medical Instrument and Food Engineering,
University of Shanghai for Science and Technology,
516 Jungong Road, Yangpu Distric, Shanghai, China
wjcai@usst.edu.cn