iScience, Volume *24*

Supplemental information

Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram

Dongdong Zhang, Samuel Yang, Xiaohui Yuan, and Ping Zhang

Transparent Methods

Data source

CPSC2018 The 1st China Physiological Signal Challenge (CPSC) 2018 hosted during the 7th International Conference on Biomedical Engineering and Biotechnology released a freely large multi-label 12-lead ECG database collected from 11 hospitals in China. This database comprises 6877 12-lead ECGs lasting between 6 s and 60 s at a sampling rate of 500 Hz. These ECGs are labeled with 9 diagnostic classes. Patient characteristics and diagnosis class prevalence of the CPSC2018 dataset are shown in Table S1. As shown in Table S1, data imbalance and insufficiency problem is severe for cardiac arrhythmias diagnosis.

Data Preprocessing

The CPSC2018 database comprises multi-label 12-lead ECGs with varying durations between 6 s and 60 s. As the deep neural network requires inputs to be of the same length, we preprocessed the dataset to make all inputs are of the same length nsteps. We tried different values for nsteps, and found that setting nsteps to 15000 (duration of 30 s, sampling rate of 500 Hz) achieved the best performance. For ECGs with a duration of more than 30 s, they will be cropped and the last 30 s ECG data are kept. Otherwise, they will be padded to 30 s with zeros.

Data Augmentation

As shown in Table S1, data imbalance and insufficiency problem is severe for cardiac arrhythmias diagnosis. To address this problem, we applied scaling and shifting for data augmentation during the training phase. Scaling multiplies the ECG signals by a random factor sampled from a normal distribution $N(1, 0.01)$ to stretch or compress the magnitude. Shifting randomly moves the time values a little bit. Data augmentation will introduce noise, but in practice, it can help reduce model overfitting and encourage robustness against adversarial examples.

Network architecture

The overview of the proposed network architecture is illustrated in Figure 2. The proposed network is developed using 1D CNNs. Similar to the original residual neural network for image recognition with 2D CNNs, residual blocks with shortcut connections are utilized in our model to make the model training tractable. The model takes the raw ECG signals $x \in \mathbb{R}^{\{nstep \times 12\}}$ (optimal value for nsteps is 15000) as input and outputs a multi-label classification result $\hat{y} \in \mathbb{R}^{1\times 9}$.

As shown in Figure 2, the network consists of 34 layers. 4 stacked residual blocks are used to extract deep features. Within each residual block, there are two 1D convolutional (Conv1d) layers, two batch normalization (BatchNorm1d) layers, 1 dropout (Dropout) layer, and two rectified linear unit (ReLU) activation layers. Conv1d layers are used to automatically extract features, BatchNorm1d layers to make the model faster and stable, ReLU layers to perform non-linear activation, Dropout layer to reduce overfitting. 1×1 convolution is used to match the dimensions and skip connections. The features extracted by stacked residual blocks are pooled using adaptive max-pooling. The pooling results are sent to the output layer with

sigmoid as activation function to make predictions.

Evaluation metrics

For each diagnostic class, we report Precision, Recall, F1 score (F1), area under the receiver operating characteristic curve (AUC), accuracy score (ACC). For class i, the metrics are calculated with the following equations:

$$Acc_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

$$F_{1i} = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i}$$

where TP_i, TN_i, FP_i, and FN_i represent the number of true positive samples, the number of true negative samples, the number of false positive samples, and the number of false negative samples for class i respectively. Class i can be one of the 9 classes: SNR, AF, IAVB, LBBB, RBBB, PAC, PVC, STD, and STE.

To better evaluate the performance of multi-label classification, we adopt average (AVG) score of each metric on 9 classes (1 normal and 8 abnormal). Average F1 score is used to select the best-performing model. And the final score is the average over classes:

$$Acc = \frac{1}{9}\sum_{i=1}^{9} Acc_i$$

$$Auc = \frac{1}{9}\sum_{i=1}^{9} Auc_i$$

$$F_1 = \frac{1}{9}\sum_{i=1}^{9} F_{1i}$$

Training and Evaluation

For model training and evaluation, we applied a 10-fold cross-validation approach. The CPSC2018 dataset was randomly divided into 10 folds. At each round, 8 folds out of 10 folds are used for training, 1 fold for validation, and 1 fold for testing. The optimal threshold of each class is selected to achieve the best F1 score on the validation dataset. Then the selected thresholds are applied to the test dataset to produce results. The reported results are the average on the test dataset of 10 rounds. Adam optimizer is used as the optimization method and cross-entropy as the loss function to train the model. The optimal values for hyperparameters of the deep neural network are: the length of ECG input is set to 15000; the learning rate is 0.0001; the batch size is 32; the maximum number of epochs is 30; the kernel size of 1D CNNs is 15; the dropout rate of dropout layers is 0.2. Besides, our code is publicly available at https://github.com/onlyzdd/ecg-diagnosis.

Interpretability

Although deep learning models can achieve state-of-the-art performance in many predictive tasks, deep learning models are usually considered to be black boxes. Due to the

multi-layer nonlinear structure, the decisions made by deep learning models are not traceable by humans. However, understanding the model's behavior when making predictions is as crucial as the accuracy of predictions in many applications, especially in clinical practice. To address this issue, we adopted the SHAP (SHapley Additive exPlanations) method to interpret the model's predictions. SHAP is a game-theoretic approach to explain the model predictions and has been applied to tree-based algorithms to enhance clinical interpretability. SHAP provides a unified way of interpreting predictions of any machine learning models, and satisfies the local accuracy, missingness, and consistency constrains. To be specific, SHAP assigns shap values, a unique additive feature importance measure (ϕ_i), to each feature for a particular prediction:

$$\phi_i = \sum_{S \subseteq F \backslash \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_s)]$$

where F is the set of all features and S is all feature subsets without the ith feature. Model $f_{S \cup \{i\}}$ is trained with that feature present, while f_S is trained with that feature withheld. The difference of predictions of these 2 model $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_s)$ are compared on the input x_s, where x_s represents the values of the input features in the set S. The effect of withholding a feature depends on other features in the model, and the preceding differences are computed for all possible subsets $S \subseteq F \backslash \{i\}$. The shap values are then computed and used as feature contributions. To estimate ϕ_i, the SHAP approach approximates the Shapley value by either performing Shapley sampling or Shapley quantitative influence.

The feature importance analysis can be used for patient level interpretation. Because shap values are directly additive, we eliminated the time factor and calculated the contribution rate of ECG leads towards diagnostic classes via the statistics of shap values. As shown in Figure 4, we applied the SHAP method to the trained deep learning model to interpret the model's behavior at both patient level and population level by utilizing a gradient explainer.

Patient level interpretation Firstly, we focus on patient-level interpretation to understand why the model is making a certain prediction for 12-lead ECG inputs. Given an ECG input $x \in \mathbb{R}^{15000 \times 12}$, the model outputs a multi-label classification result $\hat{y} \in \mathbb{R}^{1 \times 9}$. By applying the gradient explainer, a shap values matrix $sv \in \mathbb{R}^{9 \times 15000 \times 12}$ is generated for each input where $sv_{i,j,k}$ represents the feature contribution of the corresponding ECG input $x_{j,k}$ towards the diagnostic class i. If $sv_{i,j,k} > 0$, then $x_{j,k}$ contributes positively towards the diagnostic class i. For the top-predicted class $l = \operatorname{argmax} \hat{y}$, the submatrix sv_l demonstrates why the deep learning model predicts l given the ECG input x and shows the contribution of features.

Population level interpretation While patient level interpretation explains the model's behavior on a specific ECG input, population level interpretation shows the contribution of ECG leads towards each kind of cardiac arrhythmias over the entire dataset. As shown in Figure 4, population level interpretation is the summarization of patient level interpretation. Given the population of D patients and the shap values matrix $svs \in \mathbb{R}^{D \times 9 \times 15000 \times 12}$, the contribution $c_{i,k}$ of lead k for diagnostic class i is defined as the sum of shap values:

$$c_{i,k} = \sum_{d=1}^{D} \sum_{j=1}^{15000} svs_{d,i,j,k}$$

The normalized contribution rate $r_{i,k}$ of lead k towards class i is calculated as:

$$r_{i,k} = \frac{c_{i,k}}{\sum_{i=1}^{9} c_{i,k}}$$

And the average contribution rate $\bar{r}_k$ of lead k in 12-lead ECG model is:

$$\bar{r}_k = \frac{1}{9}\sum_{i=1}^{9} r_{i,k}$$

The normalized contribution rate $r_{i,k}$ shows which leads are playing an important role in diagnosing a particular cardiac arrhythmia i. The average contribution rate $\bar{r}_k$ reflects the importance of each lead and implies possible feature interactions in the deep model.

Supplemental Figures and Tables

Figure S1. An example of 12-lead ECG with AF. Related to Figure 5.

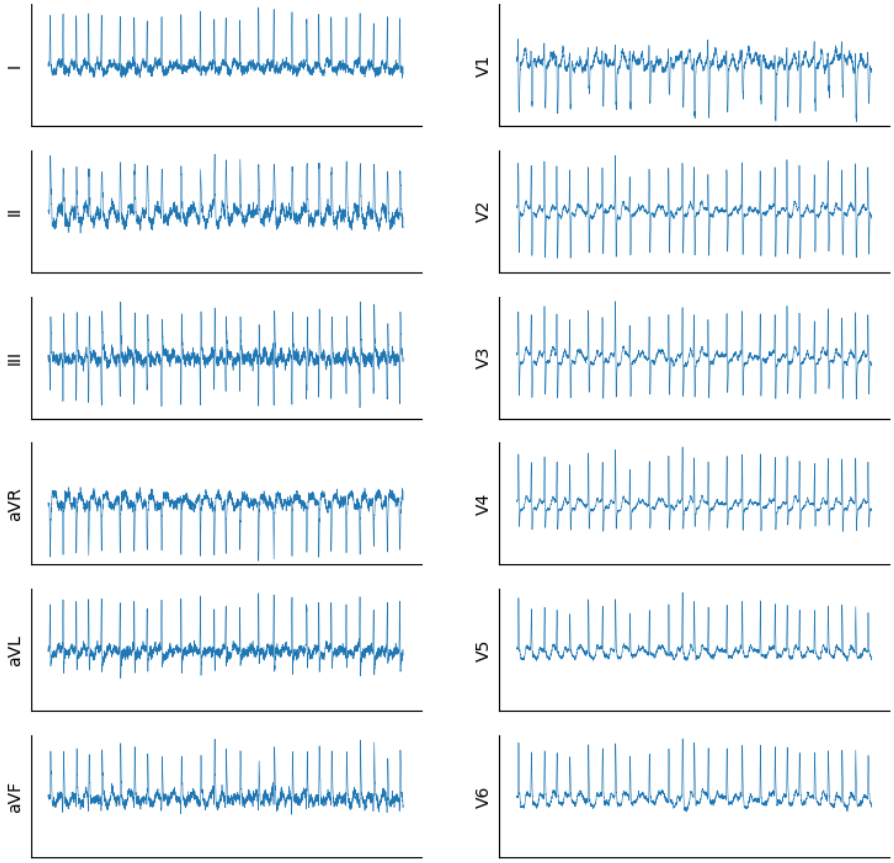

Figure S2. Multi-label confusion matrices of the best validation model predictions and ground truth. Related to Table 1.

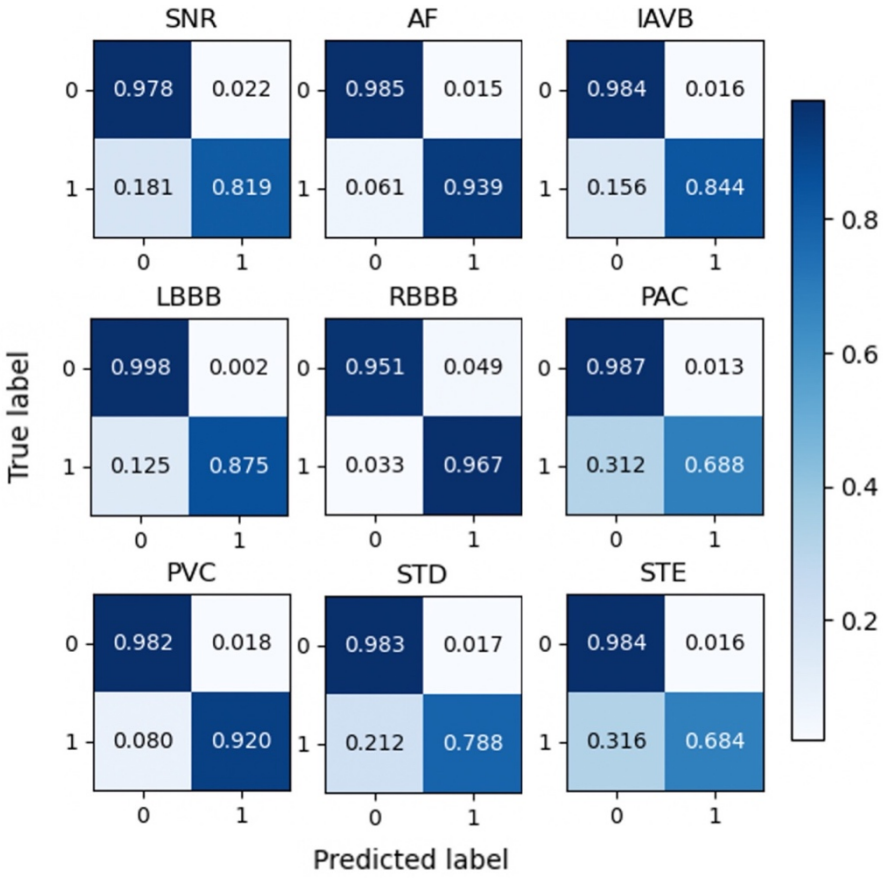

Figure S3. Examples of patient level interpretation. Related to Figure 4.

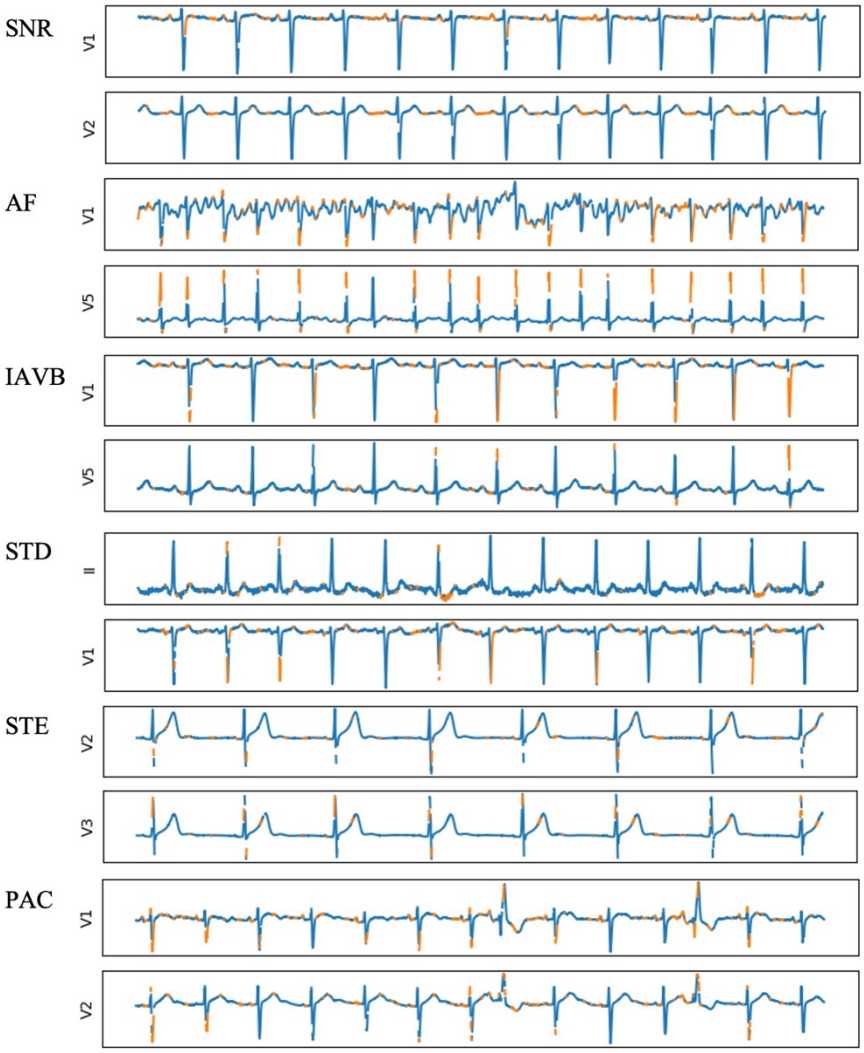

Figure S4. Failed cases when the model makes incorrect predictions (Ground truth → incorrect prediction). In this figure, (a) The ECG shows mild ST elevations in V1-V3 with ST depressions in II, III, and aVF, consistent with poor oxygenation of the cardiac muscles. The mild ST elevations in V1-V3 were not picked up by the model; (b) Both IAVB and RBBB are seen in this example. In the figure provided, the model selected RBBB as the predominant diagnosis; (c) There is a clear PVC in the second QRS in the rhythm. The p-waves are not consistent with PAC. There is some artifact in the ECG (usually due to patient movement) which could be leading to incorrect classification; (d) This example shows LBBB (confirmed by deep S wave in V1 and monophasic R wave in V6) with STE (V1-V4). As previous examples showed, ECG interpretation is complex and multiple diagnoses may exist in a single study. Related to Figure 4.

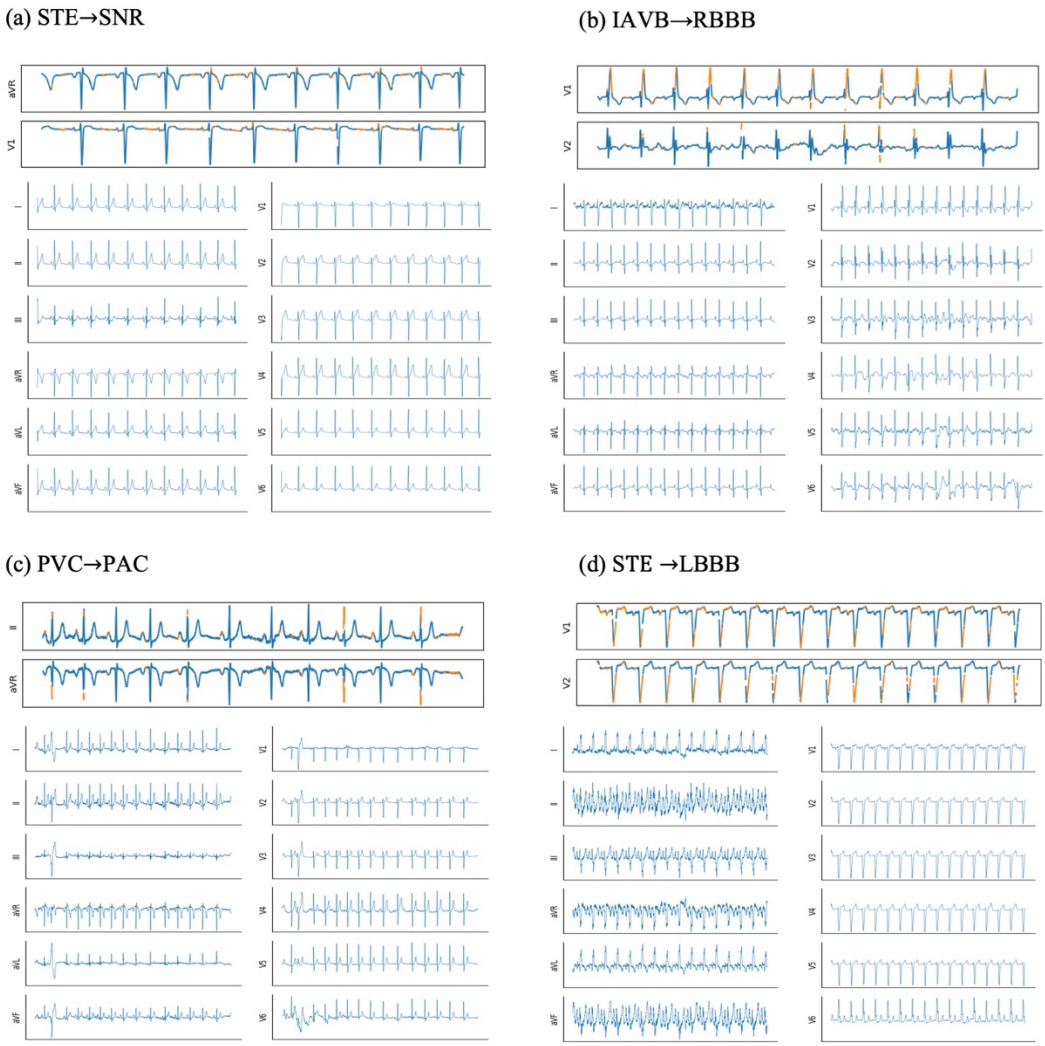

Table S1. Patient characteristics and diagnostic class prevalence on the CPSC2018 dataset. Related to Figure 5.

Class	Count (%)	Male (%)	Age	Duration
SNR	918 (13.35%)	363 (39.54%)	41.56 (18.45)	15.43 (7.64)
AF	1221 (17.75%)	692 (56.67%)	71.47 (12.53)	15.07 (8.73)
IAVB	722 (10.50%)	490 (67.87%)	66.97 (15.67)	14.42 (7.08)
LBBB	236 (3.43%)	117 (49.58%)	70.48 (12.55)	15.10 (8.10)
RBBB	1857 (27.00%)	1203 (64.78%)	62.84 (17.07)	14.73 (9.00)
PAC	616 (8.96%)	328 (53.25%)	66.56 (17.71)	19.30 (12.39)
PVC	700 (10.18%)	357 (51.00%)	58.37 (17.90)	20.84 (15.39)
STD	869 (12.64%)	252 (29.00%)	54.61 (17.49)	15.65 (9.79)
STE	220 (3.20%)	180 (81.82%)	52.32 (19.77)	17.31 (10.74)

Mean and standard deviation are reported for age and ECG duration (s).