Multi-label ECG classification using Convolutional Neural Networks in a Classifier Chain

Bjørn-Jostein Singstad[1], Eraraya Morenzo Muten[2], Pål Haugar Brekke[1]

[1]Oslo University Hospital, Oslo, Norway
[2]Institut Teknologi Bandung, Bandung, Indonesia

Abstract

Over the last decade, AI has shown its feasibility in classifying heart-related diagnoses from ECGs. Earlier studies have mainly focused on 12 and 2-lead ECGs, but we aim to classify 26 different diagnoses based on 12, 6, 4, 3, and 2-lead ECGs in this study.

We trained a supervised model on a dataset containing 88 253 ECGs with 26 different diagnoses used as ground truth. The training and classification steps can be separated into three parts. (1) Pan Tompkins algorithm was used to find peaks and calculate the average heart rate. (2) The average heart rate and the Fourier transformed ECG signal was used to train Convolutional Neural Networks (CNN) system that classified the ECGs with regular or irregular rhythms. 9 out of 26 classes were classified in this step. (3) Finally, CNN models in a classifier chain were trained to classify the remaining 17 diagnoses. The classification results from step 2 and the raw ECG signal were used as input to the classifier chain in step 3.

Our team, CardiOUS, achieved a mean PhysioNet Challenge score of 0.49, 0.44, 0.42, 0.45, and 0.44 using the 12, 6, 4, 3, and 2-lead model during 10-fold cross-validation (CV) on the development set. Unfortunately, we were not able to score the model on the hidden validation and test set.

1. Introduction

Cardiovascular diseases (CVDs) are one of the leading causes of death globally, taking an estimated 17.9 million lives each year according to numbers from WHO (WHO, 2016). Better diagnostics may lead to earlier detection of CVDs and may also have an impact on reducing the severity of the disease, and the ECG is one of the most used diagnostic tools for detecting heart diseases [1]. Taking an ECG is a non-invasive method, and the rapid development in wearable technology has also made it available on everyone's wrist. On the other hand, misinterpretations of the ECG are still done frequently by the built-in automatic interpretation software, and doctors often have to read over the raw ECGs [2]. This error is both time-consuming and requires a high degree of expertise by the doctors [3]. Therefore, it has great potential in improving the ECG interpretation algorithms, both to streamline the interpretation process and to detect patterns that doctors cannot see on the ECG.

The last decade has shown a rapid advancement in using deep learning and Convolutional Neural Networks to find patterns in images and signals. The same also applies to ECG, where a review from 2021 found 31 different applications of using Deep Learning to detect different CVDs from ECGs [4]. Some of the most surprising findings have shown that deep learning can detect persons' age and gender [5], and detect silent atrial fibrillation in patients only using their ECGs [6]. Despite the exciting findings and outstanding results in classifying cardiac abnormalities, it is also shown that many of these studies use small datasets with few classes [4]. In addition, many studies have been conducted on single sets of ECG leads. To our extent of knowledge, no study has compared the performance of a classifier on a different number of ECG leads in a single study.

This study shows how classifier chain based CNN-models perform on 12, 6, 4, 3, and 2-lead ECG. The architecture that we use is inspired by a cardiologist's ECG interpretation process. Our architecture first calculates the mean heart rate of a recording, and then it determines the ECG rhythm to classify nine more common diagnoses that can be distinguished by its heart rate and rhythm first. Finally, it classifies the other 17 diagnoses in the classifier chain.

2. Methods

2.1. Data

In this study, we used a dataset containing 88 253 ECG recordings with corresponding information files to train our models. The dataset originates from six different countries and three different continents [7–13]. The informa-

tion files described the recording, patient attributes, and the diagnosis associated with the patient's ECG. 26 out of the 133 diagnoses present in the dataset were used as labels for the supervised models in this study.

To feed the labels to the model during training, we one-hot encoded the diagnoses, such that each ECG recording had a corresponding 26-bit long array of ones or zeros. More than one diagnosis could be associated with the same ECG, making it a multi-label classification problem. In the whole development set, there were 2745 different combinations of diagnoses based on the 26 diagnoses.

2.2. Preprocessing

The ECG recordings in the development dataset were different in length and had different sampling frequencies. To be able to feed the ECG signals to the CNN models, we used two strategies:

1. Fourier transforms the ECG signal to represent the entire signal as a magnitude of the different frequencies between 0 and 100 Hz. The Fourier transformed signal was then down or upsampled to a length of 5000 samples.
2. Upsample or downsample the raw ECG signal to 500 Hz and then pad or truncate the signal to a length of 5000 samples, which is equivalent to 10 seconds. Resampled recordings with length $n > 5000$ samples were truncated to the first 5000 samples. While, those of length $n < 5000$ were padded with zeros until they reach 5000 samples.

2.3. Model

2.3.1. Heart rate detection

The first step in the model architecture was to detect heart rate. This detection was done by finding the R-peaks using the Pan-Tompkins algorithm on the raw ECG signal at its full-length [14]. The time interval between the peaks was then used to calculate the mean heart rate.

2.3.2. Rhythm classification

A Fourier transformed ECG signal and the mean heart rate were used as input in the rhythm classification. This classification was done in two steps.

1. Categorize the rhythm as regular or irregular.
2. Classify the ECG with diagnoses considered as regular or irregular based on the previous categorization.

In the training of the rhythm categorizer, ECGs labeled with atrial flutter, pacing rhythm, sinus rhythm, sinus bradycardia, sinus tachycardia, and sinus arrhythmia were extracted from the dataset and categorized as regular rhythms. In contrast, ECGs labeled with atrial fibrillation, ventricular premature beats/premature ventricular contractions, and supraventricular premature beats/premature atrial contraction were extracted from the dataset and categorized as irregular rhythms. If an ECG was categorized as a regular rhythm in step 1, it was then classified into one or more of the regular rhythms in step 2, and then all irregular rhythms were classified as false. If an ECG was categorized as an irregular rhythm in step 1, it was then classified into one or more of the irregular rhythms in step 2, and then all regular rhythms were classified as false.

2.3.3. Classifier chain

The rhythm classifier classified 9 out of 26 classes, and the remaining 17 classes were classified using a classifier chain [15]. The classifier chain trained one classifier for each of the 17 classes ordered in the chain. The first classifier in the chain was trained on the 5000 samples long ECG signals and nine labels from the rhythm classifier. The n-th classifier in the chain was trained on the 5000 samples long ECG signals, 9 labels from the rhythm classifier, and the $n - 1$ predictions from the previous classifiers in the chain. This process implies that the order of the classes is not indifferent since more data is given to the last model in the chain than the first. The 17 diagnoses were classified in the following order by the classifier chain: bundle branch block, bradycardia, 1st-degree av block, incomplete right bundle branch block, left axis deviation, left anterior fascicular block, left bundle branch block, low qrs voltages, nonspecific intraventricular conduction disorder, poor R wave Progression, prolonged pr interval, prolonged qt interval, qwave abnormal, right axis deviation, right bundle branch block, t wave abnormal, t wave inversion.

2.4. CNN architecture and model parameters

The CNN model used in both the rhythm classification and the classifier chain used an Encoder [16]. This model architecture showed its feasibility in classifying ECG in PhysioNet/CinC Challenge 2020 [17]. The final layer of this CNN model used sigmoid activation and binary cross-entropy as the loss function. [1].

2.5. Model validation and hyperparameters

The models were trained and validated on the development dataset using 10-fold CV on both 12, 6, 4, 3, and 2-lead ECG recordings, with a random seed = 42 to make the folds equal and the results reproducible. A stratified CV was used to split the data such that the distribution of diagnoses was similar in both the train and validation data.

[1] All code developed in this study are available here: `https://github.com/CardiOUS/PhysioNetChallenge2021-CNN`

All CNN models were trained using Adam optimizer, a learning rate of 0.0001, and a batch size of 30. All rhythm models were trained for 10 epochs, and all models in the classifier chain were trained for 5 epochs. The order of the data, feed to the model, were shuffled after each epoch.

3. Results

Figure 1 shows the Challenge scores obtained by the 12, 6, 4, 3, and 2-lead models on the validation split of the development data during the 10-folded CV.

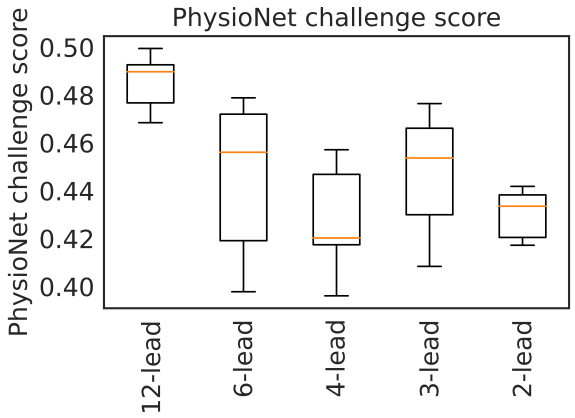

Figure 1. The boxplots show the PhysioNet Challenge score obtained by the 12, 6, 4, 3, and 2-lead models on the validation split from the 10-folded CV.

The model validation on the development set was done in Google Colab Pro with 16.28 MB GPU and 25.46 GB RAM available. The runtime for the 12, 6, 4, 3, and 2-lead models are presented in table 1. Unfortunately, we were not able to test the model on the hidden validation and test set. This was due to the extensive runtime that exceeded the maximum runtime allowed by the PhysioNet/CinC organizers.

Model	Runtime (minutes pr CV fold)	
12-lead	913	82
6-lead	896	90
4-lead	832	77
3-lead	780	75
2-lead	755	81

Table 1. The measured runtime while training 12, 6, 4, 3, 2-lead models on one CV fold represented as mean and standard deviation calculated from all 10 folds

4. Discussion and conclusion

In this study, we developed a model inspired by the ECG interpretation procedure of a cardiologist. The model is based on a heart rate calculation, a rhythm classifier, and a classifier chain used to classify 26 different diagnoses.

From figure 1 we see that the 12-lead model is significantly better than the models with a reduced amount of leads measured in terms of Challenge score on the development data.

Further, we see from the results in figure 2 that sinus rhythm, left axis deviation, and t wave abnormal show slightly lower accuracy than the other classes. A striking observation is that these diagnoses also have some of the highest prevalences in the dataset (sinus rhythm (n = 28971), left axis deviation (n = 7631), and T wave abnormal (n = 11716)). A possible explanation for this might be that the accuracy score has some drawbacks when scoring imbalanced datasets like the one present in this study. The accuracy score tends to favor the classes with low prevalence and give a relatively lower score to the balanced classes [18].

We padded and truncated the signals to 5000 samples necessary to get the same dimensions on the data, which is a premise to feed the signal to a normal CNN model. The disadvantage of this method is that important information might be clipped off from the ECG and thus not used by the model. On the other hand, the heart rate calculation and the rhythm classifier used features from the whole ECG.

The hyperparameters used in this model, such as epoch, batch size, and learning rate, were selected experimentally. The extensive runtime of these models made it too computationally heavy for hyperparameter tuning. Further experiments with more computational resources and a greater focus on hyperparameter tuning will probably increase the model's accuracy. On the other hand, it's likely that the model presented in this study is highly generalizable since it has not been overfitted due to hyperparameter tuning.

Another source of limitation of this study is that the order of the 17 classes in the classifier chain was picked randomly. In future studies, this could be tuned by finding the optimal order of classification in the classifier chain.

In summary, the 12-lead model seems to perform better than the models with a reduced amount of leads on the development set used in this study. On the other hand, there was no significant difference between the 6, 4, 3, and 2-lead models. This may have implications for future studies of wearable ECGs.