Classification of 12-lead ECG With an Ensemble Machine Learning Approach

Matteo Bodini, Massimo W Rivolta, Roberto Sassi

Dipartimento di Informatica “Giovanni Degli Antoni”, Università degli Studi di Milano, Milan, Italy

Abstract

The PhysioNet 2020 Challenge focused on the automatic classification of 27 cardiac abnormalities (CAs) from 12-lead ECG signals. We investigated on a hybrid approach, combining average-template-based algorithms with deep neural networks (DNNs), to build an ensemble classification model. We calibrated the model on the available 40,000+ ECGs, while organizers tested the model on a private test set. Standard ECG preprocessing was applied. For ECGs related to CAs altering the ECG morphology, multi-lead average P, QRS, and T segments were computed. For signals associated with irregular rhythms, time dependent features were computed. The ensemble model comprised of: i) three DNNs to classify morphology-related CAs. ii) a fully connected neural network to classify irregular rhythm; and iii) a threshold-based classifier for premature ventricular beat detection. The organizers designed a score for ranking the models. The ensemble model proposed by our team “BiSP Lab” reached the 40th position, and obtained a score of -0.179 on the private test set. Despite the low performance obtained on the private test set, our ensemble model showed potential for classification of CAs from ECGs.

1. Introduction

The 12-lead clinical ECG is a fundamental diagnostic tool for detecting many cardiac abnormalities (CAs) [1], such as cardiac arrhythmias, myocardial infarction, cardiac axis deviation *etc*. The reliable and automatic detection and correct diagnosis of CAs can largely increase the odds that the treatments become successful [2]. However, the majority of the algorithms available both commercially and in the literature tackles the diagnosis of a relatively small amount of cardiac conditions. Thus, to cover up for their actual vast amount, many algorithms need to be implemented and integrated, with the hassle of merging all their predictions. This complexity might represent one of the main factor for a reduced performance observed in computerized ECG analysis [2].

A similar problem already occurred in the Computer Vision domain where the implementation of dedicated algorithms for the automatic classification of thousands of classes became unfeasible. In addition, in case the problem formulation would change, for instance by adding a new class, all algorithms would require a very delicate time-consuming phase of re-calibration. In this context, deep neural networks (DNNs) tackle the problem using a single mathematical model more flexible to changes and easier to update in case new data become available. However, DNNs require a very large amount of data to perform a proper calibration (a.k.a., training) for achieving a reliable performance.

In the context ECG classification, only several studies investigated the use of DNNs obtaining promising results. Of note, Ribeiro *et al.* [3] proposed a DNN, trained with $>$2,000,000 12-lead ECGs, to classify among 7 different classes. Hannun *et al.* [4] created a DNN using $>$95,000 single lead ECG for detecting 12 different arrhythmias.

The PhysioNet Computing in Cardiology 2020 Challenge asked to identify 27 different CAs from 12-lead ECG recordings, by means of an automatic algorithm [5]. Organizers provided a dataset containing about 40,000 recordings of clinical ECGs, collected from multiple sources, along with their diagnosis. Requested diagnoses could be categorized as affecting the ECG morphology or the rhythm or both. The problem formulation fitted well in the supervised Machine Learning domain, specifically as a multi-class classification problem.

In this study, considering the limited sample size provided, we designed a machine learning algorithm based on an ensemble of four classification models, specifically trained to detect different subsets of CAs. Then, the predictions of each model were concatenated to provide the requested output.

2. Materials and Methods

2.1. Dataset

The dataset provided for the challenge contained 12-lead clinical ECG signals in WFDB format, labeled with one or more CAs, among 111 possible ones, in SNOMED-CT codes [5]. The dataset was obtained merging data from the following sources: i) Southeast University, in-

ISSN: 2325-887X DOI: 10.22489/CinC.2020.406

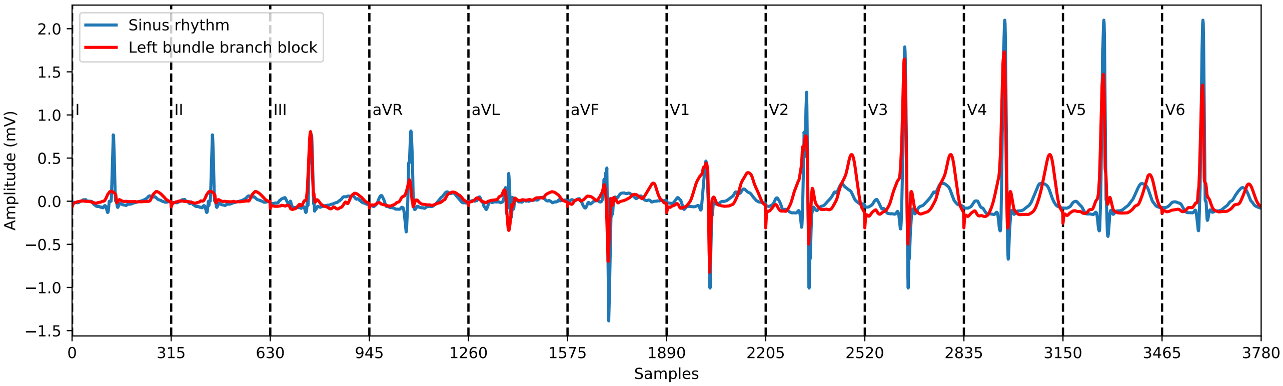

Figure 1: Example of average 12-lead PQRST template for sinus rhythm (blue line) and left bundle branch block (red line).

cluding the data from the China Physiological Signal Challenge 2018; ii) St. Petersburg Institute of Cardiological Technics; iii) The Physikalisch Technische Bundesanstalt; and iv) Georgia 12-Lead ECG Challenge Database. ECG recordings lasted from 6 seconds to 30 minutes and sampling rates ranged from 257 Hz to 1000 Hz, where the majority was sampled at 500 Hz. A total number of 43101 ECG signals was available where each was characterized by 111 possible classes. We only used the dataset provided for the challenge.

A private test set handled by the challenge organizers was used for evaluating the algorithms proposed by the participants.

2.2. Preprocessing and feature extraction

ECG signals were downsampled or upsampled to 500 Hz according to their actual sampling rate and filtered with a bandpass Butterworth filter (3rd order, zero phase, and pass-band: 0.67–30 Hz) to reduce powerline interference, baseline wandering and high frequency noise. Only the first 1 minute segment (or less, depending on the signal) of each ECG was further processed.

Beats were detected on the vector magnitude[1] (VM) using the *gqrs* algorithm [6] and beat positions were refined using the Woody algorithm applied to the VM [7]. Signal quality was assessed computing the average crosscorrelation between each QRS complex and an average QRS template. ECGs were further considered only when the signal quality was higher than 0.9 for each lead. After quality check, 4752 signals were detected as bad quality and discarded from the dataset.

Depending on the CA to detect, we processed the ECG signals differently. First, given the fact that CAs altering the ECG morphology were not transient, we created an average PQRS template, *i.e.*, from R peak -260 ms to R+370 ms, for each lead that were concatenated afterwards. Figure 1 reports two examples of such concatenated vector. Second, for the rhythm-related CAs, from the inter-beat time interval series (RR), we extracted the following features: RR median, RR standard deviation, RR minimum distance, RR maximum distance, and root mean square of successive differences of RR. Third, for detecting Premature Ventricular Contractions (PVCs), we computed the maximum amplitude on the VM signal.

2.3. The ensemble model

We designed an ensemble model comprising of four neural networks and a threshold-based classifier. Figure 2 reports the complete scheme of the ensemble model.

Three convolutional neural networks (CNNs), *i.e.*, P−CNN, QRS−CNN and T−CNN, were designed to classify CAs altering the morphology of the P, QRS and T segments, respectively. Each network classified different classes:

- P−CNN classified I-AVB and LPR;
- QRS−CNN classified CRBBB, IRBBB, LAnFB, LAD, LBBB, LQRSV, NSIVCB, QAb, RAD and RBBB;
- T−CNN classified LQT, TAb, and TInv.

The input features of the three CNNs were the respective concatenated P, QRS and T average segments taken from each lead of the average beat. Specifically, P segments spanned in the range (R-260 ms, R-150 ms), QRS complexes were taken in the range (R-50 ms, R+50 ms) and T segments ranged in (R+100 ms, R+370 ms).

Each CNN comprised of one or more convolutional layer, a fully connected layer and an output layer whose dimension depended on the number of classes to classify. The structure of the three CNNs is shown in Fig. 3.

A feed-forward neural network (FFNN) was designed to classify the CAs related to irregular rhythms (hereafter, named as Rhythm−NN). The classes were AF, AFL, Brady, PR, PAC, SA, SB, STach, and SVPB. The input

[1] Square root of the sum of the squared ECG leads

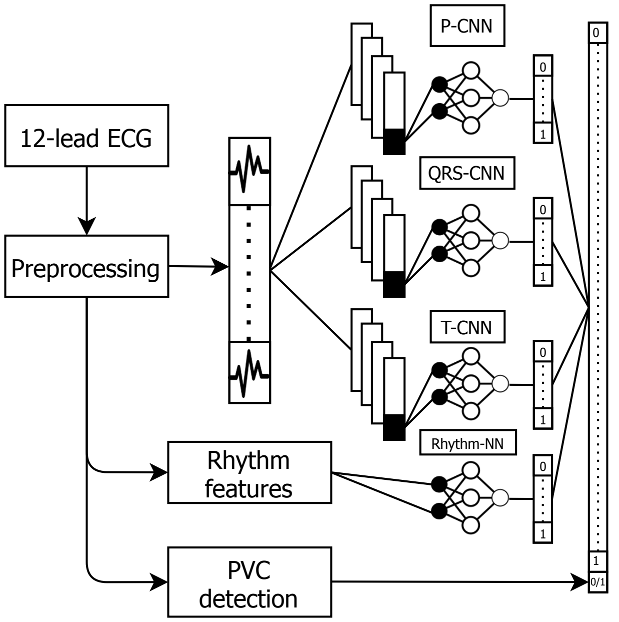

Figure 2: Scheme of the ensemble model.

features were those extracted from the RR series (see sec. 2.2). The network had two hidden layers with 64 and 32 neurons, respectively, and an output layer with 9 neurons, equivalent to the number of rhythm classes.

ECG containing PVCs were classified using a threshold calibrated by means of a Receiving Operating Curve analysis performed on the maximum value of the VM signal. The optimal cut-off was selected as that one balancing the true positive and negative rates.

For all the networks, the Rectified-Linear unit activation function was used for the fully connected layers, and the Sigmoid activation function in the output layer. No activation functions were set after the convolutional layers. Batch Normalization and Dropout (with a rate starting at 0.1 in the first layer and with a 0.1 increase each further layer) were used in all the layers, except the last one, as regularization techniques. The Adam algorithm was used as optimizer ($\epsilon = 10^{-8}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$) and the average binary cross-entropy across classes was set as loss function. The batch size was set to 64 samples.

Table 1: Confusion matrices for the four networks composing the ensemble model. Values were normalized by the row.

	P−CNN		QRS−CNN	
	Pred+	Pred−	Pred+	Pred−
Act+	0.71	0.29	0.87	0.13
Act−	0.12	0.88	0.19	0.81

	T−CNN		Rhythm−NN	
	Pred+	Pred−	Pred+	Pred−
Act+	0.76	0.24	0.74	0.26
Act−	0.15	0.85	0.15	0.85

For training the ensemble model, four datasets were built containing only the input features within to the considered subset of CAs. Then, each dataset was randomly sampled with stratification using a 70/30 training/validation split. Models were trained separately for 1000 epochs on their respective training set. Metrics were computed on the validation set to assess the performance.

The model submitted for the evaluation on the private test set was trained using all the available data without splitting and using the same configuration.

Given the fact that the output of the CNNs and FFNN were sigmoid functions, resembling then the conditional probability of observing a given class, the final decision was taken by setting a 0.5 threshold for such probabilities.

The output vector was obtained by concatenating all the decisions obtained by the CNNs, FFNN and the threshold-based classifier. In addition, the decision related with the detection of a "normal" ECG was also concatenated to the output vector. It is worth mentioning that the ECG were classified as "normal" only if no other CA was detected.

3. Results

Given the multi-label classification problem, the confusion matrices for the four neural networks composing the ensemble model were computed in class-wise manner and results are reported in Table 1. Positive classes contained CAs specific to the neural network under evaluation, while the negative classes contained all the others. Confusion matrices were normalized by row, *i.e.*, dividing by the total number of samples for each class.

We computed the recall values for all the 27 scored classes provided with the dataset. The three highest recall values were obtained by the QRS−CNN for RBBB and LBBB (0.93 and 0.85, respectively) and by P−CNN for I-AVB (0.88). The worst values were achieved by the Rhythm-CNN for Pacing (0.71) and Flutter (0.74), and for the normal ECG detection (0.74).

The area under the ROC curve for the PVC detection was 0.82, obtaining true positive and negative rates of 0.72. The identified threshold was 1.44 mV.

The challenge scoring system made use of a metric depending on the recognition performance of each class in a weighted manner. Organizers made available the scoring system: we obtained a score of 0.241 on the validation set and a score of -0.179 on the private test set.

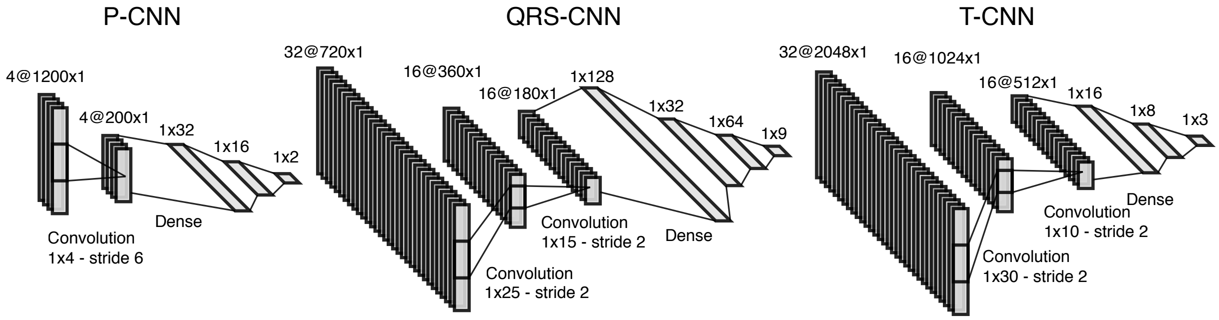

Figure 3: Network architectures for the three CNNs of the ensemble model.

4. Discussion

The ensemble model reached intermediate classification performance. The QRS−CNN was the best among the four NN models and it reached the top highest recall values (up to 0.93 for RBBB) computed over all the 27 classes. The other three networks and the PVC detector showed moderate performance, reporting the worst recall values. We noticed that the worst performance were correlated with classes having a low number of samples. For instance, the P−CNN model was trained using only the available 340 samples with prolonged PR, achieving one of the lowest recall (0.74).

Several are the improvements that can be implemented. First, the Rhythm−NN and PVC detector can be substituted with more efficient models. In fact, Hannun *et al.* [4] and Zhou *et al.* [8] recently demonstrated that DNNs can achieve high recall values for both rhythm and PVC detection. Second, the low recall values of the T−CNN might be due to the preprocessing step implemented. Indeed, the average PQRST template did not account for changes in the heart rate within the considered 1-minute segment, while it is well known the heart-rate dependency of the T-wave duration. RR-binning can be used to improve this aspect instead of averaging beats within the entire segment. Third, the recall value for normal ECG detection was among the lowest ones. The detection by elimination, *i.e.*, when the final output was the zero vector, was sensitive to misclassification of any of the other classes. For example, if misclassifications were statistical independent between the 26 classes and the error rate was just random at 1% (but we are still far from this value for many classes), the misclassification of normal ECG would be approximately 23%, leading to an extremely high false positive rate. A possible solution might be designing and adding another CNN in the ensemble model, whose input is the average PQRST template and several rhythm-related features, capable of recognizing normal ECGs.

Differently from previous studies on DNNs, where ECG signals were roughly injected in the model, we tested a hybrid approach, merging average-template-based algorithms, known to be effective, with the state-of-the-art for classification in deep learning, using an ensemble model. The approach seemed suitable to deal efficiently with the challenging multi-class problem of ECG classification and the limited sample size available.

References

[1] Kligfield P, Gettes LS, Bailey JJ, at al. Recommendations for the standardization and interpretation of the electrocardiogram. J Am Coll Cardiol 2007;49(10):1109–1127.

[2] Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms. J Am Coll Cardiol 2017;70(9):1183–1192.

[3] Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11:1760.

[4] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–69.

[5] Perez Alday EA, Gu A, Shah A, et al. Classification of 12-lead ECGs: the Physionet/Computing in Cardiology Challenge 2020. Physiol Meas 2020 (Under Review);.

[6] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[7] Woody CD. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Med Biol Eng 1967;5(6):539–554.

[8] Zhou X, Zhu X, Nakamura K, Mahito N. Premature ventricular contraction detection from ambulatory ecg using recurrent neural networks. In Conf Proc IEEE Eng Med Biol Soc. 2018; 2551–2554.

Address for correspondence:

Matteo Bodini
Dipartimento di Informatica “Giovanni Degli Antoni”, Università degli Studi di Milano, Via Celoria 18, Milan 20133, Italy
matteo.bodini@unimi.it