

Classification of 12-lead ECGs using digital biomarkers and representation learning

David Assaraf^{1,*}, Jeremy Levy^{1,*}, Janmajay Singh², Armand Chocron¹ and Joachim A. Behar¹

¹Faculty of Biomedical Engineering, Technion-IIT, Haifa, Israel

²Independent researcher.

*Equal contribution.

Abstract

Background: The 12-lead electrocardiogram (ECG) is a standard tool used in medical practice for identifying cardiac abnormalities. The 2020 PhysioNet/Computing in Cardiology Challenge addresses the topic of automated classification of 12-lead ECG.

Methods: Two machine learning strategies were implemented: a feature engineering approach based on the engineering of physiological features (or “digital biomarkers”) reflecting the analysis performed in clinical practice and a deep learning approach. Two sets of features were engineered: (1) capturing the interval variation between consecutive heartbeats, commonly called heart rate variability (HRV) measures and (2) using morphological biomarkers (e.g. QT interval, QRS width). A total of 16 HRV and 97 morphological biomarkers were implemented in python for each lead. A random forest (RF) model was trained using 5-fold cross validation to optimize the model hyperparameters. Feature selection was performed using a sequential forward floating selection procedure. For the deep learning approach, a residual neural network (ResNet) architecture was used. The RF and ResNet were also combined in an ensemble learning (EL). The dataset was divided into 80%-20% stratified training-test sets.

Results: on the local test set we achieved a Challenge score of 0.65 using the FE approach, 0.52 using the DL approach and 0.66 using the EL approach. For technical reasons we did not manage to score our models on the Challenge hidden test set.

1. Introduction

The electrocardiogram (ECG) is a non-invasive representation of the electrical activity of the heart. It is measured using electrodes placed on the surface of the torso. The standard 12-lead ECG has been widely used to diagnose a variety of cardiac abnormalities such as cardiac arrhythmia and conduction abnormalities. The early and correct diagnosis of cardiac abnormalities can increase the

chances of successful treatments. However, manual interpretation of the electrocardiogram is time-consuming, and requires skilled personnel with a high degree of training. Automatic detection and classification of cardiac abnormalities can assist physicians in performing an accurate diagnosis. The 2020 PhysioNet/Computing in Cardiology Challenge [1] addresses the topic of automated classification of 12-lead ECG.

2. Methods

In this work we evaluate two machine learning (ML) strategies. The first is a feature engineering (FE) based ML approach that captures known patterns in the physiological time series thus mimicking the cardiologist’s approach to 12-lead ECG analysis. The second is a deep learning (DL) approach that may capture features beyond the traditional ones. Finally, we combine both approaches within an Ensemble Learning (EL) model. A general overview of the individual and the combined approaches is shown in Figure 1.

Signal pre-processing: Similar to the work of Zheng et al. [2], we performed a validation step to detect examples where electrodes were switched. In these instances we inverted their polarity. The signal quality was estimated using the *bSQI* index [3, 4]. Examples with a low signal quality (*bSQI* < 0.8) were discarded. Recordings had different sampling rates (*fs*) across databases (Table 1). We resampled all the databases to *fs* = 500Hz. ECGs were pre-filtered with a Butterworth band-pass filter in order to filter out baseline wander and high-frequency noise. A low cutoff frequency at 0.13Hz and a high cut-off frequency at 88Hz were used. Two Notch filters, with cut-off frequencies at 50Hz and 60Hz respectively were used to remove power-line interference noise.

Databases: The Challenge dataset was composed of 6 databases (Table 1). We split the long ECGs recordings from the INCART database into independent 10-sec 12-lead ECGs, thus totalling 13500 10-sec examples. In order to augment the number of training examples, ex-

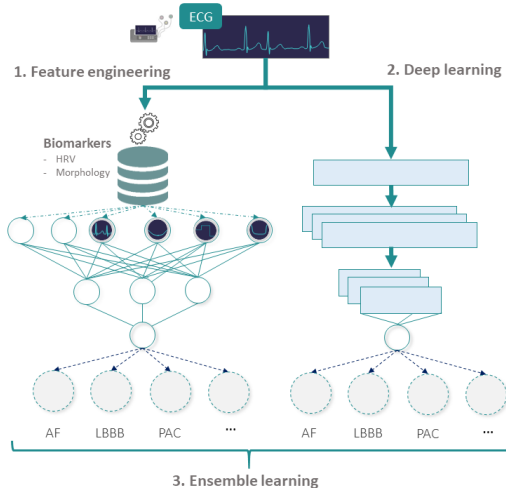


Figure 1. Machine learning strategies: (1) a FE approach, (2) a DL approach and (3) an EL combining (1) and (2).

ternal open-access databases were used. The first one, from Zhang et al. [2] recorded at the Chapman University and Shaoxing People’s Hospital, denoted “CUSPH” contains 10646 examples. The second, by Zheng et al. [5] recorded at the Chapman University and Ningbo First Hospital, denoted “CUNFH”, is mostly composed of premature ventricular contractions (PVC) and contains 334 examples. A total of 27 classes were considered in the Challenge among which 26 cardiac abnormalities and normal sinus rhythm (SNR). All examples that had no label among the 27 classes considered were excluded because these were not scored by the Challenge performance measure. This resulted in $n=53730$ examples (Table 1). Three pairs of classes were considered strictly equivalent by the Challenge scoring system: CRBBB with RBBB, PVC with VPB and PAC with SVPB. In summary, we considered a 24 multilabel classification problem.

Database	Included	Excluded	fs (Hz)
CPSC-1	5279	1598	500
CPSC-2	1278	2175	500
Georgia	9458	886	500
PTB	97	419	1000
PTB-XL	21604	233	500
INCART	5940	7380	257
CUSPH (<i>add</i>)	9749	897	500
CUNFH (<i>add</i>)	325	9	2000
TOTAL	53730	12691	

Table 1. Examples that were included and excluded for each database based on the presence of at least one label within the 27 classes that were considered by the Challenge performance measure.

2.1. Feature engineering approach

Waveform analysis and features extraction: Fiducial points on the ECG waveform were extracted using the open source wavedet algorithm by Martínez et al. [6]. QRS complex, P-wave and T-wave delineations were obtained by applying wavedet to a representative ECG cycle [7] for each lead. The QRS duration and frequency morphology distortion were extracted to capture premature atrial and ventricular contractions and bundle branch blocks, [8], [9], [10], [11]. Atrioventricular block were detected through the statistical analysis of the PR interval [12]. A total of 113 morphological biomarkers were engineered for each lead or their vectorcardiogram representation [13]. HRV features implemented in Chocron et al. [14] were used. Finally age and sex were included. Features were standardized by subtracting the mean and dividing by the standard deviation.

Machine learning strategy: The database was divided between training and test sets following a 80%-20% split. Figure 2 shows the class distribution for the training and test sets. Of note, for the purpose of producing this figure, an example with multiple labels was virtually considered as a count of +1 for each of the classes with a positive label. 5-fold cross validation was used for hyperparameters tuning. To address the multilabel classification challenge, we used a one versus the rest approach, first by training a classifier for SNR against the rest, then by training another 23 different classifiers for each cardiac abnormality against all other abnormalities (but not SNR). The minority classes were proportionally over-weighted to account for the imbalance of the classes. Feature selection was performed in order to prevent overfitting and ensures the model’s generalization. Minimum Redundancy Maximum Relevance (mRMR) [15] was used and subsequently, a floating recursive feature addition (RFA) algorithm was used. The feature selection was stopped when the Challenge score did not improve significantly on the average validation set F_β score. For each cardiac abnormality we selected the decision threshold that maximized the average F_β score for this class on the validation sets.

2.2. Deep learning approach

The same preprocessing as for the feature engineering approach was used. Two deep learning (DL) networks were trained. The first network is a binary classifier for SNR against the rest. The second network is used to classify for any of the 23 cardiac abnormalities using a multi-head model. Padding was performed, to the constant length of 40000 samples (i.e. 80 seconds at $fs=500Hz$).

The feature extraction part of both networks is a Convolutional Neural Network (CNN) model. Each CNN cell was composed of: Conv1d, BatchNorm, ReLu, Conv1d,

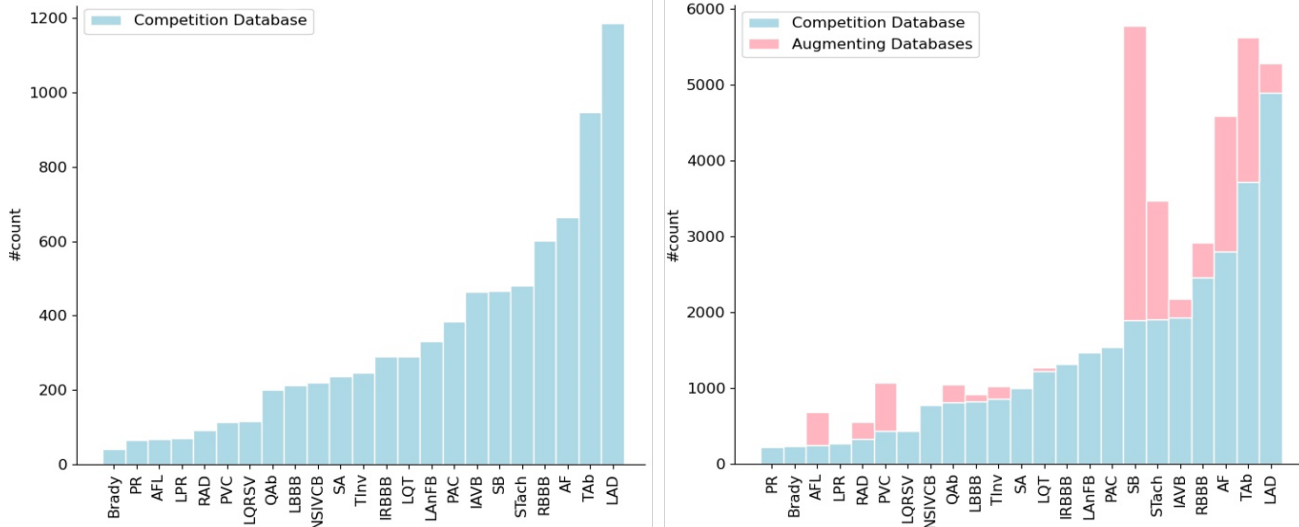


Figure 2. 80%/20% Training/Testing split with augmenting databases highlighted.

BatchNorm, ReLu, MaxPool, Dropout. The network was trained with shortcut connections to avoid vanishing gradients. A total of 8 CNN cells were used, with shortcuts every 2 cells, followed by a Gated Recurrent Unit (GRU) cell. The GRU takes as input a total of 1250 features produced by the CNN. A multi-head model was used, which means a fully connected layer for each of the 23 cardiac abnormality. Stratified 5-fold cross validation validation was performed to find the best hyperparameters of the model. A combination of two losses was used: BCEWithLogit-sLoss, and the F-Beta score. The weight of each loss was considered a hyperparameter of the model. The Adam optimizer was used.

2.3. Ensemble Learning Model

In order to combine the strengths of the FE and the DL approaches an EL model was evaluated. It consisted in adding a logistic regression unit taking as the input the probability outputs of the FE and DL models.

3. Results

Signal quality evaluation using *bSQI* excluded a total of 3213 examples (2647 from the Challenge original dataset and 566 from the additional databases). A total of 140 features were selected (Figure 3) out of the 1360 features computed for each example. The results of the hyperparameter tuning for the one versus the rest RF classifier were: 750 estimators and a max depth of 70. The results for each of the three models are reported in Table 3, for the 5-fold cross validation and the local test set. For technical reasons we did not manage to score our models on the Challenge hidden test set.

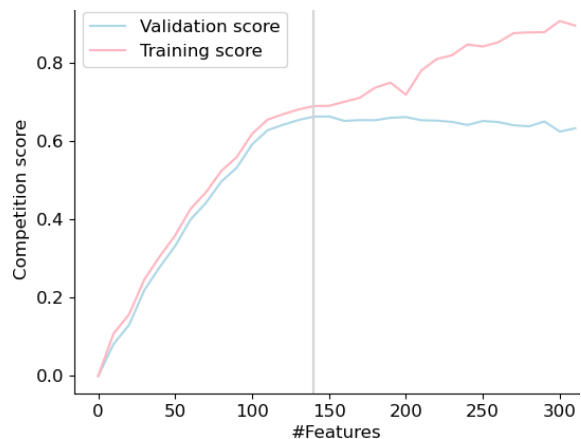


Figure 3. Feature Selection process. Performance corresponds to the Challenge score.

Model	CV	Local Test
FE	0.66+/-0.02	0.65
DL	0.54+/-0.04	0.52
EL	0.66+/-0.03	0.66

Table 2. 5-fold cross validation and local test set scores of the different models. Scores on the hidden test set are not reported, due to technical difficulties with the submission.

4. Discussion and conclusion

In our experiment the FE approach outperformed the DL approach. The EL approach did not improved significantly the results over the FE approach alone. However, by improving further our DL performance we expect that the EL approach will significantly improve over the individ-

Model	FE	DL	EL
F_{β}^{AF}	0.91	0.41	0.92
F_{β}^{AFL}	0.46	0.27	0.44
F_{β}^{Brady}	0.18	0.13	0.18
$F_{\beta}^{I AVB}$	0.76	0.38	0.77
F_{β}^{IRBBB}	0.16	0.26	0.17
F_{β}^{LAnFB}	0.52	0.26	0.49
F_{β}^{LAD}	0.81	0.69	0.83
F_{β}^{LBBB}	0.80	0.58	0.81
F_{β}^{LQRSV}	0.03	0.09	0.05
F_{β}^{NSIVCB}	0.59	0.23	0.57
F_{β}^{PR}	0.03	0.09	0.07
F_{β}^{PAC}	0.72	0.38	0.71
F_{β}^{PVC}	0.52	0.42	0.51
F_{β}^{LPR}	0.17	0.12	0.22
F_{β}^{LQT}	0.75	0.68	0.73
F_{β}^{QAb}	0.27	0.24	0.27
F_{β}^{RAD}	0.62	0.60	0.63
F_{β}^{RBBB}	0.80	0.68	0.81
F_{β}^{SA}	0.50	0.29	0.46
F_{β}^{SB}	0.80	0.32	0.80
F_{β}^{SNR}	0.80	0.90	0.86
F_{β}^{STach}	0.90	0.63	0.90
F_{β}^{TAb}	0.66	0.32	0.64
F_{β}^{TInv}	0.71	0.24	0.71

Table 3. F_{β} scores for individual cardiac abnormalities considered in the Challenge and for the local test set. FE: feature engineering approach, DL: deep learning approach and EL: ensemble learning approach.

ual models.

Among the five lowest F_{β} scores, four (namely Brady, LQRSV, PR and LPR) had less than 500 examples in the training set (Figure 2). This suggests that augmenting the number of examples in these classes will likely improve the performance of the model for these classes.

Among the features selected in the FE approach, the most relevant features were: cosEn, SD1, AFEv, Origin-Count, PACEV and IRmax for the HRV features and Dqrsmax, Dqrsmean, Dqrsmed, Dqrsstd (which relate to QRS duration for PAC, PVC and Bundle Branch Blocks) for the morphological biomarkers.

References

- [1] Perez Alday EA, Gu A, 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 eegs: the physionet/computing in cardiology challenge 2020. *Physiological Measurement Under Review* 2020;.
- [2] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data* 2020; 7(1):1–8.
- [3] Li Q, Mark RG, Clifford GD. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological measurement* 2007;29(1):15.
- [4] Behar J, Oster J, Li Q, Clifford GD. Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE transactions on biomedical engineering* 2013; 60(6):1660–1666.
- [5] Zheng J, Fu G, Anderson K, Chu H, Rakovski C. A 12-lead ecg database to identify origins of idiopathic ventricular arrhythmia containing 334 patients. *Scientific Data* 2020; 7(1):1–10.
- [6] Martínez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A wavelet-based ecg delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering* 2004;51(4):570–581.
- [7] Oster J, Behar J, Sayadi O, Nemati S, Johnson AE, Clifford GD. Semisupervised ecg ventricular beat classification with novelty detection based on switching kalman filters. *IEEE Transactions on Biomedical Engineering* 2015;62(9):2125–2134.
- [8] et a.l. VTK. Automatic detection of premature atrial contractions in the electrocardiogram. *Electrotechniques Electronics E E* 2006;9(10).
- [9] Hao Chen JBea. Automatic identification of premature ventricular contraction using eegs. *International Conference on Health Information Science* 2019;143–156.
- [10] M.L. Talbi AC. Pvc discrimination using the qrs power spectrum and self-organizing maps. *Computer methods and programs in biomedicine* 2009;94(3):223–231.
- [11] Martín-Yebra A, Martínez JP. Automatic diagnosis of strict left bundle branch block using a wavelet-based approach. *PLoS one* 2019;14(2):e0212971.
- [12] Luning Mao HCea. Automated detection of first-degree atrioventricular block using eegs. *International Conference on Health Information Science* 2018;156–167.
- [13] Dawson D, Yang H, Malshe M, Bukkapatnam ST, Benjamin B, Komanduri R. Linear affine transformations between 3-lead (frank xyz leads) vectorcardiogram and 12-lead electrocardiogram signals. *Journal of electrocardiology* 2009; 42(6):622–630.
- [14] Chocron Armand ea. Remote atrial fibrillation burden estimation using deep recurrent neural network. *arXiv preprint arXiv:200802228* 2020;.
- [15] Hanchuan Peng Member IFLea. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 2005;27(8):1226–1239.

Address for correspondence:

Joachim A. Behar
jbehar@technion.ac.il