Convolutional Recurrent Neural Network and LightGBM Ensemble Model for 12-lead ECG Classification

Charilaos Zisou, Andreas Sochopoulos, Konstantinos Kitsios

Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki,
Thessaloniki, Greece

Abstract

Automatic abnormality detection of ECG signals is a challenging topic of great research and commercial interest. It can provide a cost-effective and accessible tool for early and accurate diagnosis, which increases the chances of successful treatment.

In this study, an ensemble classifier that identifies 24 types of cardiac abnormalities is proposed, as part of the PhysioNet/Computing in Cardiology Challenge 2020. The ensemble model consists of a convolutional recurrent neural network that is able to automatically learn deep features, and LightGBM, a gradient boosting machine that relies on hand-engineered expert features. The individual models are combined using class-specific weights and thresholds, which are tuned by a genetic algorithm.

Results from 5-fold cross validation on the full training set, report the Challenge metric of 0.593 that outperforms both individual models. On the full hidden test set, the proposed architecture by "AUTh Team" achieves a score of 0.281 with an official ranking of 13/41.

1. Introduction

Cardiovascular diseases are the leading cause of death globally [1]. In order to provide an effective treatment, early and accurate diagnosis is of utmost importance, however, it relies on manual ECG inspection by trained professionals, which is a time-consuming and expensive process. Attempts at automating this process are a significant step towards a cost-effective and accessible tool.

Over the years numerous approaches have been proposed, including feature-based [2] and deep learning [3] classifiers, nevertheless, most of them have been tested on relatively homogeneous datasets, with few target classes. PhysioNet/Computing in Cardiology Challenge 2020 promotes research on automatic cardiac abnormality detection by making 12-lead ECG databases from a wide set of sources, publicly available [4].

The authors propose an ensemble model that efficiently combines a convolutional recurrent neural network (CRNN) and a gradient boosting machine (GBM), termed LightGBM [5], with interpretable results as to which model has higher predictability for each class.

The rest of the paper is organized as follows: Section 2 describes the methods used in the proposed analysis, as well as the ensemble model architecture. Section 3 presents the results, while Section 4 discusses model performance. Finally, Section 5 concludes the paper.

2. Methods

The overall pipeline consists of: data relabeling and pre-processing to deal with database format discrepancies, feature extraction for the feature-based LightGBM classifier, training of both individual models CRNN and LightGBM, and finally the creation of the ensemble. The pre-processing steps and the ensemble model architecture are depicted in the form of a block diagram in Figure 1.

2.1. Data Relabeling and Pre-processing

The provided datasets come from multiple sources with different sampling rates, lengths and lead gains and were recorded under imperfect, noisy conditions. The data also contain abnormalities that the Challenge organizers decided not to score, thus the first step of the analysis involves data relabeling and pre-processing, in order to create a single unified dataset.

Out of a total of more than 100 classes, 27 of them are scored, with 6 of them being pair-wise identical. All recordings with unscored diagnoses are removed, while the remaining ones are relabeled, resulting in 24 target classes. The signals are then filtered using a 3rd order low-pass Butterworth filter with a cutoff frequency of 20Hz for the high frequency noise and a Notch filter with a cutoff frequency of 0.01Hz for the baseline wander. Since 99.6% of the relabeled recordings have a sampling rate of 500Hz and 89.2% of the data have a length of 10 seconds, the remaining recordings are also resampled at the same target frequency and truncated or padded to 10 seconds.

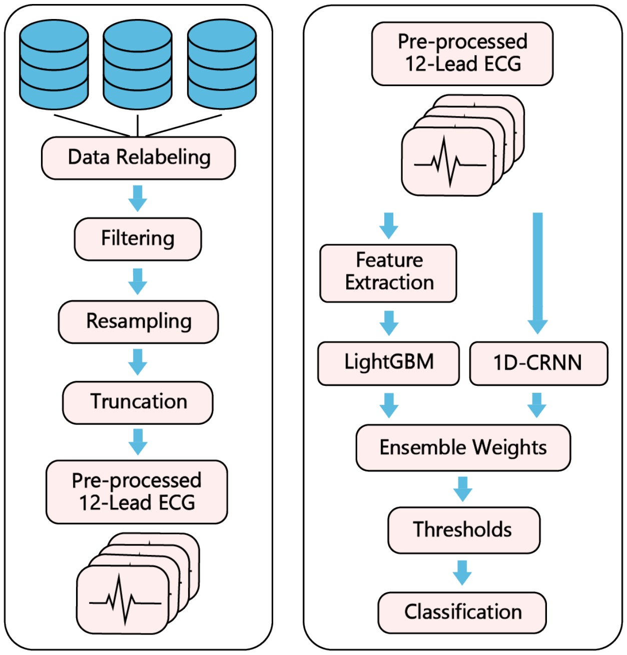
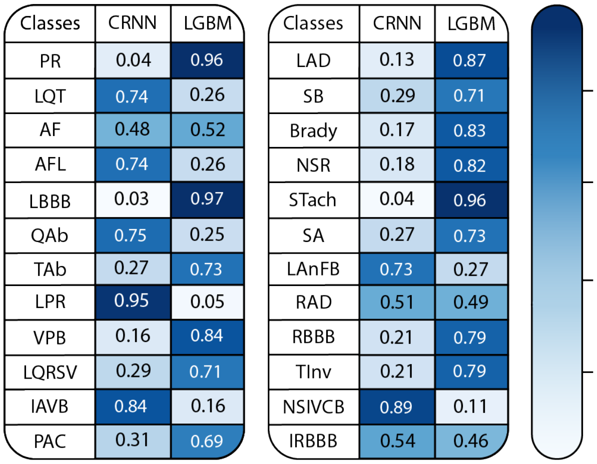


Figure 1. Pre-processing and ensemble architecture.

2.2. Feature Extraction

After pre-processing, feature extraction is performed so as to produce the input for LightGBM, the feature-based classifier. ECG signals contain oscillatory modes, thus wavelet multiresolution analysis (WMRA) occupies quite an important role. The signal is decomposed into multiple scales/resolutions using the discrete wavelet transform (DWT) and information about the presence of modes associated with specific abnormalities, is extracted through statistical measures, i.e. standard deviation, skewness and kurtosis. The selected mother wavelet is ’sym5’ and the number of scales is empirically determined to be 8. Scales 1, 2 and 8 contain noise and baseline wander residue and are, therefore, excluded from the analysis.

Under the existence of white noise, traditional signal processing techniques often resort to assumptions of linearity and Gaussianity. However, ECG signals, like most real-life data, are inherently non-linear and non-Gaussian. For that purpose, the proposed analysis involves higher order statistics (HOS). Specifically, the wavelet bispectrum (WBS) is computed using the Gaussian complex mother wavelet due to its time localization property [6]. From the WBS, highest peak: amplitude, 1st and 2nd frequency and standard deviation along the axis of one frequency keeping the other fixed, are extracted. Also, signal processing features such as signal energy, autocorrelation function values, signal-to-noise ratio (SNR) and R peak amplitude statistics are calculated. All of the aforementioned features are extracted from every lead.



Classes	CRNN	LGBM
PR	0.04	0.96
LQT	0.74	0.26
AF	0.48	0.52
AFL	0.74	0.26
LBBB	0.03	0.97
QAb	0.75	0.25
TAb	0.27	0.73
LPR	0.95	0.05
VPB	0.16	0.84
LQRSV	0.29	0.71
IAVB	0.84	0.16
PAC	0.31	0.69

Classes	CRNN	LGBM
LAD	0.13	0.87
SB	0.29	0.71
Brady	0.17	0.83
NSR	0.18	0.82
STach	0.04	0.96
SA	0.27	0.73
LAnFB	0.73	0.27
RAD	0.51	0.49
RBBB	0.21	0.79
TInv	0.21	0.79
NSIVCB	0.89	0.11
IRBBB	0.54	0.46

Figure 2. Genetic algorithm optimal weights.

Following [7], standard time domain, frequency domain and non-linear heart rate variability (HRV) features are computed. These include: RR interval and successive differences statistics, power spectral density features, Poincaré plot features, HRV triangular index and sample entropy. Finally, patient sex and age are also added.

2.3. LightGBM

The extracted features are used to train a LightGBM model. LightGBM is a gradient boosting decision tree (GBDT) [8] algorithm implementation, proposed by Microsoft [5]. It introduces algorithmic optimizations such as the histogram method, gradient-based one-side sampling and exclusive feature bundling. These algorithms, designed to reduce time complexity, offer a well-balanced trade-off between accuracy and efficiency, especially in large-scale, high-dimensional data.

Since boosting algorithms are not affected by feature scaling/normalization, no such processing is performed, however, leads corrupted with significant motion artefacts are imputed with zeros. Boosting algorithms do not directly support multi-label classification, therefore, 24 individual binary classifiers are built using one-vs-rest logic. To address binary class imbalance, the synthetic minority over-sampling technique (SMOTE) [9] is employed, and the minority class is over-sampled at 10% of the majority class. The authors in [9] recommend combining SMOTE with random under-sampling of the majority class, therefore random under-sampling by 50% is performed. This way, a more balanced class distribution and, ultimately, a more robust model is achieved. The training process uses the binary logloss function as the objective, a learning rate of 0.075 and early stopping based on the validation set area under the receiver operating characteristic curve (AUROC).

Description	AUROC
Basic training	0.920 ± 0.001
SMOTE	0.930 ± 0.002
SMOTE + BO	0.935 ± 0.002

Table 1. 5-fold cross validation scores before applying anything (Basic training), after applying the synthetic minority over-sampling technique (SMOTE) and after using the Bayesian optimizer (BO).

Performance is further improved by hyper-parameter tuning. Due to its ability to tackle expensive-to-evaluate functions, Bayesian optimization (BO) [10] is used. Some abnormalities may be harder to identify than others, thus parameters are tuned independently, in order to allow each binary classifier to select the appropriate model complexity for optimal performance. The parameters that are tuned are: *feature_fraction*, *lambda_l1*, *lambda_l2*, *max_depth*, *min_child_weight*, *min_split_gain* and *num_leaves*. Scores before and after applying SMOTE and BO are compared in Table 1

2.4. Convolutional Recurrent Neural Network

In deep learning theory, the convolutional neural network (CNN) is one of the most developed areas, widely used in image recognition for its spatial feature extraction properties. Frequently employed in time series data, the recurrent neural network (RNN) is a different type of neural network that has the ability to capture sequential, time domain information. A CRNN, i.e. a RNN stacked on top of a CNN is able to capture both temporal and spatial features, making it suitable for ECG applications.

The model used in this study is a modified CRNN from [11] with 24 convolutional filters per layer to account for the extra number of target classes. The convolutions are 1D, treating the 12 leads as channels and the activation function is the LeakyReLU. The architecture includes batch normalization (BN) to reduce internal covariate shift [12], dropouts [13], a self-attention mechanism [14] and a bidirectional gated recurrent unit (Bi-GRU) which is a type of long short-term memory (LSTM) network.

The network is trained using the 10 second preprocessed data. For faster convergence and stability during training, it is common practice to perform input scaling, thus each lead is normalized to zero mean, unit variance. The neuron weights are initialized using the Xavier method, the Adam optimizer is selected, and the training is performed with early stopping based on the validation AUROC score.

Metric	CRNN	LightGBM	Ensemble
AUROC	0.908	0.935	0.946
Challenge metric	0.511	0.549	0.593

Table 2. Average scores from 5-fold cross validation on the training set. All models use GA optimized thresholds in order to have comparable Challenge metrics.

2.5. Ensemble

As described in the previous sections, during both LightGBM and CRNN training, the Challenge metric is not monitored. The goal is to create classifiers that are as accurate as possible, because the Challenge metric is optimized during the final merging process. The ensemble is a class-depended weighted average, i.e. 24 different weights and thresholds are assigned to each class so that:

$$p_{ens}[c] = p_{crnn}[c] \cdot w_c + p_{lgbm}[c] \cdot (1 - w_c) \qquad (1)$$

$$l[c] = \begin{cases} 1, & \text{if } p_{ens}[c] \geq thr_c \\ 0, & \text{if } p_{ens}[c] < thr_c \end{cases} \qquad (2)$$

where w_c is the weight, thr_c is the threshold, $p_{ens}[c]$, $p_{crnn}[c]$ and $p_{lgbm}[c]$ are the probabilities of the ensemble, CRNN and LightGBM, respectively, and $l[c]$ is the label of class c. These weights and thresholds are tuned by a genetic algorithm (GA), that uses the negative value of a 5-fold Challenge metric average as the fitness function.

Finally, in order to reuse all training data, but also to improve model robustness, 10-fold bagging is performed, i.e. 10 different ensemble models are trained, and their predicted probabilities are averaged for the final submission.

2.6. Validation

The provided datasets have severe class imbalances, therefore, in order to perform reliable early stopping, hyper-parameter tuning and model evaluation, all cross validations must be done in a stratified manner. Since the problem is multi-label, an iterative-stratification cross-validation procedure is used as proposed in [15].

3. Results

The results from 5-fold cross validation on the training set are compared in Table 2. Figure 2 illustrates the optimal ensemble weights, as determined by the GA. The final Challenge submission achieves a metric of 0.281 on the full hidden test set.

4. Discussion

From Table 2 it is evident, that LightGBM performs better, with a score of 0.549 compared to CRNN's score of

0.511. An overall comparison shows that the ensemble outperforms both individual classifiers, achieving a score of 0.593. This suggests that both models bring value to the final classification, by complementing each other's weaknesses. Since LightGBM relies on feature-engineering, its advantage is that it can benefit from expert knowledge. On the contrary, the CRNN does not need any a-priori information extraction, because it can automatically learn deep features. The drawback is that it cannot benefit from domain-specific knowledge, because making informed adjustments in deep architectures is a non-trivial process.

The proposed approach, does not need any manual weight adjustment, since it is done automatically by the GA. The optimal weights also offer some form of interpretability. Figure 2 shows that the GA generally gives higher preference to LightGBM for HRV-related cardiac arrhythmias and to CRNN for QRS morphology-related abnormalities.

For future work, more domain-specific feature engineering will be pursued. Another potential avenue is to use segmentation techniques, to address classes that contain information in specific regions during long recordings.

5. Conclusion

This paper proposed an ensemble model for automatic 12-lead ECG classification of 24 types of cardiac abnormalities, as part of the PhysioNet/Computing in Cardiology Challenge 2020. Both a feature-based and a deep learning approach were implemented and efficiently combined, using a genetic algorithm. The comparative evaluation results demonstrate the effectiveness of the ensemble process, as an automatic and interpretable method for combining models.

Acknowledgments

The authors would like to thank Prof. Leontios J. Hadjileontiadis and the Signal Processing and Biomedical Technology Unit (SPBTU) for their valuable contributions and support.

References

[1] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al. Heart Disease and Stroke Statistics – 2019 Update: a report From the American Heart Association. Circulation 2019;.

[2] Alickovic E, Subasi A. Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. Journal of Medical Systems 2016; 40(4):108.

[3] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv170701836 2017;.

[4] 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 ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Under Review 2020;.

[5] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. 2017; 3146–3154.

[6] Taplidou SA, Hadjileontiadis LJ. Nonlinear analysis of heart murmurs using wavelet-based higher-order spectral parameters. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2006; 4502–4505.

[7] Camm AJ, Malik M, Bigger JT, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 1996;93:1043–1065.

[8] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001;1189–1232.

[9] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16:321–357.

[10] Pelikan M, Goldberg DE, Cantú-Paz E, et al. Boa: The bayesian optimization algorithm. In Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99, volume 1. Citeseer, 1999; 525–532.

[11] Chen TM, Huang CH, Shih ES, Hu YF, Hwang MJ. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. Iscience 2020;23(3):100886.

[12] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv150203167 2015;.

[13] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 2014;15(1):1929–1958.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In Advances in Neural Information Processing Systems. 2017; 5998–6008.

[15] Sechidis K, Tsoumakas G, Vlahavas I. On the stratification of multi-label data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2011; 145–158.

Address for correspondence:

Charilaos Zisou
School of Electrical and Computer Engineering
Aristotle University of Thessaloniki
charilaz@ece.auth.gr