Multilabel 12-Lead Electrocardiogram Classification Using Gradient Boosting Tree Ensemble

Alexander W. Wong[1], Weijie Sun[2], Sunil V. Kalmady[2], Padma Kaul[2], Abram Hindle[1]

[1] University of Alberta, Edmonton, Canada
[2] Canadian VIGOUR Center, Edmonton, Canada

Abstract

Standard 12-lead electrocardiograms (ECGs) are commonly used to detect cardiac irregularities such as atrial fibrillation, blocks and irregular complexes. For the PhysioNet/CinC 2020 Challenge, we built an algorithm using gradient boosted tree ensembles fitted on morphology and signal processing features.

For each lead, we derived features from the heart rate variability, PQRST template shape, and full waveform duration. We concatenated the features for all 12 leads to fit an ensemble of gradient boosting trees and predicted probabilities of ECG instances belonging to each class. We used repeated random sub-sampling by splitting our dataset of 43,101 records into 100 independent runs of 85:15 training/evaluation splits for our evaluation results.

1. Introduction

The electrocardiogram (ECG) is the most effective tool and current best practice strategy for detecting cardiac diseases, outperforming screening history and physical examinations in accuracy and sensitivity [1]. However, ECG interpretation is a complex and highly skilled task with disagreements between cardiologist and non-cardiologist health care staff reference interpretations, with up to a 33% interpretation error rate [2]. Despite active research in computerized interpretations of ECGs, trained human over-reading and confirmation is required and emphasized in published reports [3].

This work classifies standard 12-lead ECGs to their clinical diagnosis as part of the *PhysioNet/Computing in Cardiology Challenge* [4]. We developed a multi-label classification algorithm using natural language and signal processing inspired features and a gradient boosting tree ensemble.

1.1. Dataset & Scoring Criteria

The official phase dataset contains a total of 43,101 ECG records. Each record is labelled with a set of one or more SNOMED CT codes, although not all labels are used in the challenge. Tab. 1 displays the 27 codes, mapped to labels for readability, chosen for evaluation by the challenge.

Table 1. Labels count and percentage in dataset.

Dx	Count	% Total
1st degree av block	2394	5.6%
atrial fibrillation	3475	8.0%
atrial flutter	314	0.7%
bradycardia	288	0.7%
complete right bundle branch block	683	1.6%
incomplete right bundle branch block	1611	3.7%
left anterior fascicular block	1806	4.2%
left axis deviation	6086	14.1%
left bundle branch block	1041	2.4%
low qrs voltages	556	1.3%
nonspecific intraventricular conduction	997	2.3%
pacing rhythm	299	0.7%
premature atrial contraction	1729	4.0%
premature ventricular contractions	188	0.4%
prolonged pr interval	340	0.7%
prolonged qt interval	1513	3.5%
qwave abnormal	1013	2.4%
right axis deviation	427	1.0%
right bundle branch block	2402	5.6%
sinus arrhythmia	1240	2.9%
sinus bradycardia	2359	5.5%
sinus rhythm	20846	48.4%
sinus tachycardia	2402	5.6%
supraventricular premature beats	215	0.5%
t wave abnormal	4673	10.8%
t wave inversion	1112	2.6%
ventricular premature beats	365	0.8%

Given a set of diagnoses $C = \{c_i\}$, we compute a confusion matrix $A = [a_{ij}]$ where a_{ij} contains records that are classified as class c_i and belong to class c_j. The weights $W = [w_{ij}]$, shown in Fig. 1, are set by the challenge to indicate clinical similarity between classes. The scoring function is defined as score $= \sum_{ij} w_{ij} a_{ij}$.

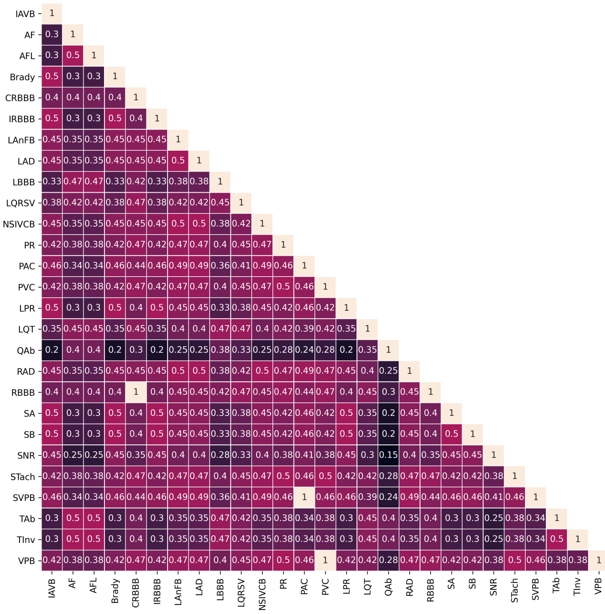

Figure 1. Evaluation scoring function weights per label.

2. Methodology

Our approach is inspired by existing methods which use feature engineering and shallow learning classifiers [5, 6]. An overview our methodology is presented in Figure 2.

Our feature extraction approach relies on the *NeuroKit2* (version `0.0.40`) neurophysiological signal processing library for ECG signal cleaning, PQRST annotation, signal quality calculation, and heart rate variability metrics [7]. Additionally, we use the general purpose time series feature extraction library *tsfresh* (version `0.16.0`) for analysis of the PQRST heart beat window, as well as the overall lead signal [8].

2.1. Signal Pre-processing

Prior to feature extraction and classification, signal preprocessing was performed to normalize and clean the raw ECG signal. Slow drift and DC offset were removed with a highpass Butterworth filter followed by smoothing using a 50Hz period moving average kernel. We treat each of the cleaned lead signals independently and annotate the PQRST peaks, the PRT onsets and offsets, and the atrial and ventricular systole and diastole phases.

We isolate one candidate heart beat signal for each lead by segmenting heart beat windows as a -0.35, 0.5 second window around each R-peak, shortening this window to -0.25 and 0.4 if the mean heart rate exceeds 80 beats per minute. We use a continuous index of ECG signal quality by interpolating the distance of each QRS segment from the average QRS segment in the data. ECG signal quality is therefore relative for each step in the entire length of the signal, where 1 corresponds to beats that are closest to the average sample and 0 corresponds to beats that are most distant to the average sample. We choose the heart beat window with the highest signal quality as our candidate lead PQRST template.

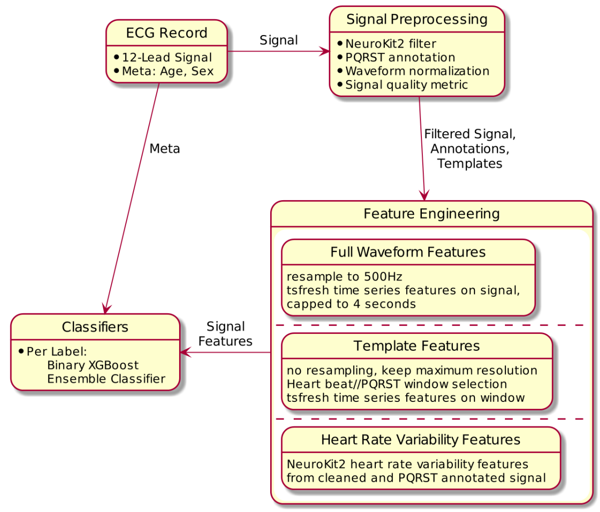

Figure 2. Methodology overview. Feature engineering is performed concurrently for each lead then concatenated.

2.2. Feature Engineering

Our engineered features can be categorized as one of three categories. Full waveform features are derived using the end-to-end ECG signal. Template features are constructed from the extracted heart beat window during preprocessing. Heart rate variability features rely on the relative distances between each R-peak. Each extraction technique was performed independently per lead.

For full waveform features, we used the cleaned ECG signal and applied the *tsfresh* feature extraction library. Due to memory and time constraints for this category of features, the cleaned signal was resampled to 500Hz. We further limited the signal to the middle 4000 samples, or 4 seconds, to reduce the impact of starting and trailing noise. From each lead, we applied the *tsfresh* `ComprehensiveFCParameters` setting to generate 763 full waveform features.

Template features were derived from the isolated heart beat window with highest signal quality, using the *tsfresh* library. No downsampling or truncation was necessary as extracted windows were less than 1 second in length. Similar to the full waveform features, 763 template features were generated per lead.

Heart rate variability (HRV) features reported by *NeuroKit2* were calculated using the cleaned signal and corresponding signal annotations. The relative R-peak distances were the primary inputs for these generated features. For each lead, 53 different HRV features were generated.

For each 12-lead record we combined all three categories of engineered features with the age and sex parsed from the ECG record metadata. We arrive at a $12 * (763 + 763 + 53) + 2 = 18,950$ length feature vector per 12-lead record. For undefined features, such as the HRV feature set on signals where no PQRST annotations could be generated, `NaN` placeholders were set.

2.3. Classification

We trained a XGBoost binary classifier for each clinical diagnosis, using `xgboost@1.1.1` [9]. We used the `gbtree` booster method with the `gpu_hist` tree method and `gradient_based` sampling method. All other classifier model parameters were left to their default values.

The evaluation scoring function weights were used during training as instance sample weights, threshold set to 0.5. For example, when training the 1st degree av block (IAVB) classifier we also considered instances of bradycardia (Brady), incomplete right bundle branch block (IRBBB), prolonged PR interval (LPR), sinus arrhythmia (SA), and sinus bradycardia (SB) as positive examples with 0.5 weight. Other labels that had scoring function weights below 0.5 were treated as negative examples with a sample weight of 1. To account for the dataset label imbalance, we further scaled the positive example weight using the number of negative examples over the positive examples in the training set split.

For our evaluation results, we performed repeated random sub-sampling of our total dataset, randomly splitting our 43,101 records into an 85:15 training/evaluation set split. Our results contain the aggregate of 100 independent runs of this configuration.

3. Results

Using 100 independent runs of an 85:15 training and evaluation split of the provided dataset, our methodology attained a mean challenge metric score of 0.478 on the evaluation set. Additionally, we attained a mean values for AUROC of 0.886, AUPRC of 0.383, accuracy of 0.260, overall F_1 score of 0.364, F_β of 0.416, and G_β measure of 0.220. An overview of the experiment classification metrics can be found in Figure 3.

Our model's top three best classified labels are normal sinus rhythm (SNR, F_1 mean: 0.921), right bundle branch block (RBBB, F_1 mean: 0.854), and left bundle branch block (LBBB, F_1 mean: 0.838). The summary of the each of our 100 experiments F_1 scores for each label can be found in Figure 4.

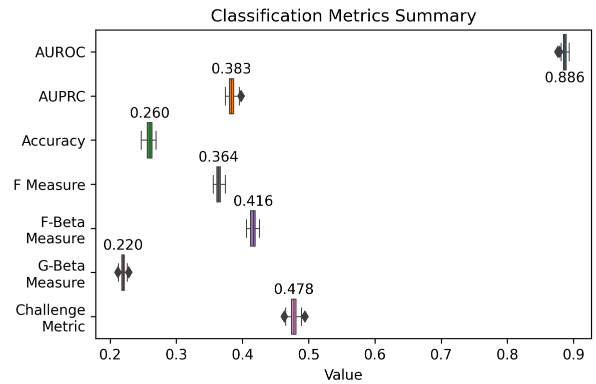

Figure 3. Summary of classification metrics over 100 experiments on all labels. Annotations indicate mean value.

Furthermore, we ran a Pearson correlation coefficient test between the label F_1 means and the label counts within our dataset. The statistical test revealed a Pearson correlation coefficient of 0.569 at a p-value of $3.7 * 10^{-3}$. This result suggests that a positive linear correlation exists between the label occurrence in our dataset and our classification model's F_1 score.

On the official phase hold out test set, our methodology achieved a challenge score of *TODO: Official phase results when released. How to discuss feature importance, or OMIT?*.

4. Discussion & Future Work

Despite the label specific scaling of our dataset training weights, the correlation between the label occurrence with the F_1 scores revealed further improvements are necessary to address the existing label imbalance. The label imbalance may be addressed by adding more low occurrence disorders into the existing corpus of ECG records. Synthesizing new records of low occurrence disorders to use as training data may also prove promising. Additionally, exploration of new features to use as classifier inputs may reveal common characteristics of specific heart disorders that are currently unused or missing.

Our approach, although applicable to 12-lead ECGs, perform feature extraction on each lead separately before concatenating the features together for classification. We believe that further improvements can be made utilizing feature extraction approaches capable of handling multidimensional time series data.

We experimented with removing features with variance threshold under 0, 10, 20, and 30%, as well as feature selection using a separate XGBoost classification model to calculate feature importances. Although our classification input dimension used a large number of features, we found no statistically significant improvements in challenge score

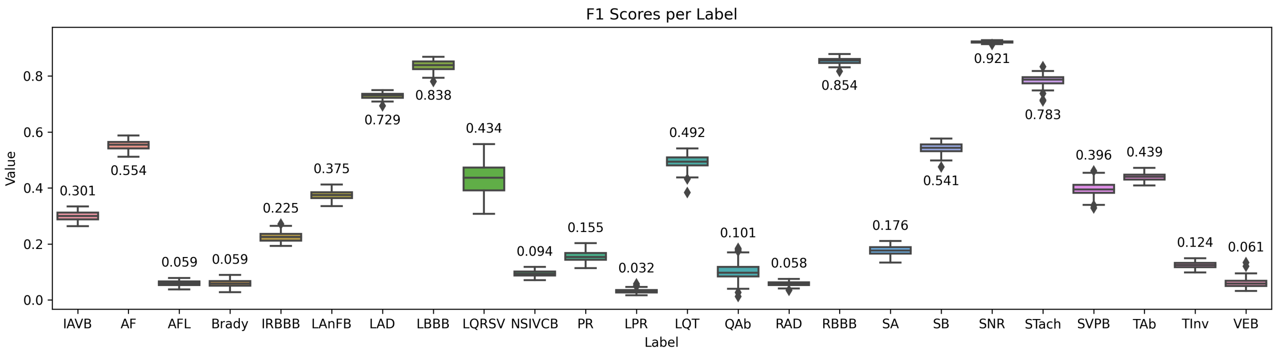

Figure 4. Label specific F1 classification score over 100 experiments. Annotations indicate mean value.

by incorporating feature selection into our machine learning pipeline.

5. Conclusion

We created an algorithm for the classification of 27 heart conditions using natural language and signal processing inspired feature engineering and an XGBoost tree ensemble classifier. We combined a set of 18,950 features from full waveform, template, and heart rate variability groups. Using 100 repeated random sub-sampling, our evaluation challenge score was 0.478. The official phase challenge score on PhysioNet's hold out test set was *TODO*.

Acknowledgments

We would like to thank Eric Ly and Leiah Luoma of the Canadian VIGOUR Center for their help and guidance during our research journey.

References

[1] Harmon KG, Zigman M, Drezner JA. The effectiveness of screening history, physical exam, and ECG to detect potentially lethal cardiac disorders in athletes: A systematic review/meta-analysis. Journal of Electrocardiology May 2015;48(3):329–338. ISSN 0022-0736.

[2] Mele P. Improving electrocardiogram interpretation in the clinical setting. Journal of Electrocardiology September 2008;41(5):438–439. ISSN 0022-0736.

[3] Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms: Benefits and Limitations. Journal of the American College of Cardiology August 2017;70(9):1183–1192. ISSN 0735-1097, 1558-3597. Publisher: Journal of the American College of Cardiology Section: The Present and Future.

[4] Reyna M, Alday EAP, Gu A, Liu C, Seyedi S, Rad AB, Elola A, Li Q, Sharma A, Clifford G. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. PhysioNet, 2020. DOI: 10.13026/f4ab-0814.

[5] Tziakouri M, Pitris C, Orphanidou C. Identification of AF and Other Cardiac Arrhythmias from a Single-lead ECG Using Dynamic Time Warping. September 2017; .

[6] Goodwin A, Goodfellow S, Eytan D, Greer R, Mazwi M, Laussen P, Goodfellow S. Classification of Atrial Fibrillation Using Multidisciplinary Features and Gradient Boosting. September 2017; .

[7] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, S H Chen A. Neurokit2: A python toolbox for neurophysiological signal processing, 2020.

[8] Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). Neurocomputing 2018;307:72 – 77. ISSN 0925-2312.

[9] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. San Francisco, California, USA: Association for Computing Machinery. ISBN 978-1-4503-4232-2, August 2016; 785–794.

Address for correspondence:

Alexander W. Wong
Department of Computing Science
2-32 Athabasca Hall, University of Alberta
Edmonton, Alberta, Canada
T6G 2E8
alex.wong@ualberta.ca