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 Centre, 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 to classify ECG diagnosis.

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

Our methodology generates an average validation split challenge score of 0.484, with a PhysioNet official phase hold out test set score of 0.476 under the team name, CVC.

1. Introduction

The electrocardiogram (ECG) is the current "gold standard" strategy for detecting cardiac diseases, outperforming screening history and physical examinations in accuracy and sensitivity [1]. However, ECG interpretation is a complex task with frequent disagreements between health care staff, 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/CinC 2020 Challenge* [4]. We develop a multi-label classification algorithm using entropy and signal processing inspired features and a gradient boosting decision tree ensemble.

1.1. Dataset & Scoring Criteria

The official phase dataset contains a total of 43,101 ECG records. Each record contains a set of one or more SNOMED CT codes, although not all labels are evaluated in the challenge. Table 1 displays the 27 labels selected for evaluation by the challenge organizers.

Table 1. Labels count and percentage in dataset.

Diagnosis	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%
Q wave 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%

The objective of this challenge is to maximize the metric: $\sum_{ij} w_{ij} a_{ij}$. 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 Figure 1, are set by the challenge to indicate clinical similarity between classes.

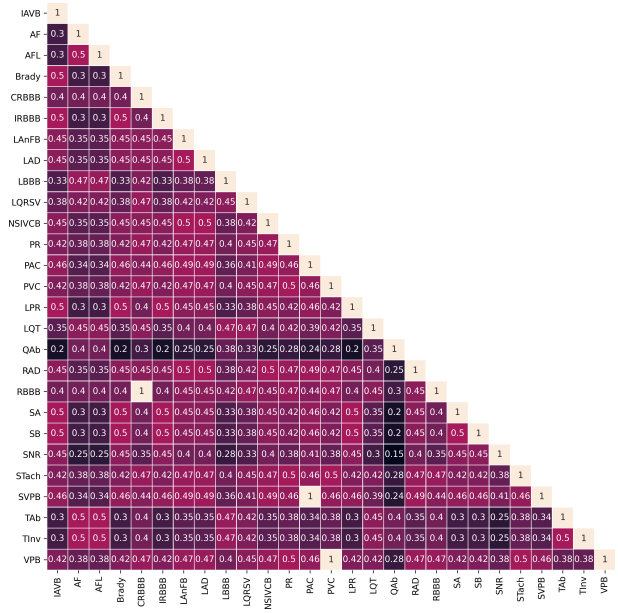

Figure 1. Evaluation scoring function weights per label.

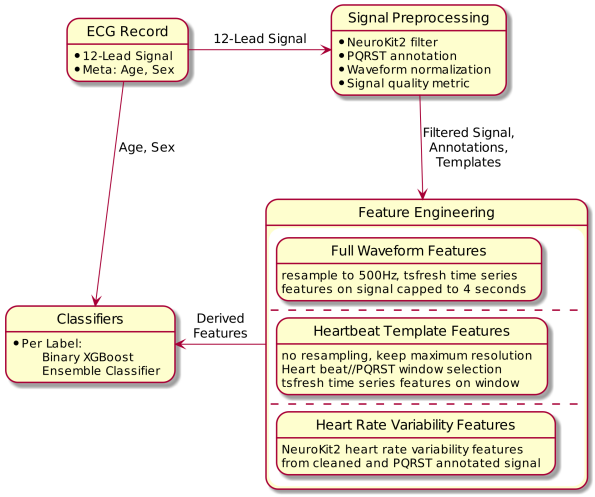

Figure 2. Methodology overview. Feature engineering is performed concurrently for each lead then concatenated.

2. Methodology

Our approach is inspired by existing methods which use feature engineering and shallow learning classifiers [5]. Figure 2 shows an overview of our learning algorithm pipeline from first cleaning and preprocessing the ECG, to then extracting the full waveform, heartbeat template, and heart rate variability features, finally using these features as input to our binary classifiers.

We rely 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 [6]. We also use the time series feature extraction library *tsfresh* (version `0.16.0`) for analysis of the PQRST template and the full waveform [7].

2.1. Signal Pre-processing

First we perform signal pre-processing to normalize and clean the raw ECG signal. Slow drift and DC offset are removed with a Butterworth highpass filter followed by smoothing using a moving average kernel of 0.02 seconds. Each of the cleaned leads are independently annotated with the PQRST peaks, the PRT onsets, and PRT offsets.

We isolate one candidate heart beat signal for each lead by segmenting heart beat windows as a -0.35 to 0.5 second window around each R-peak, shortening to a -0.25 to 0.4 second window if the mean heart rate exceeds 80 beats per minute. We create an ECG signal quality metric 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 QRS and 0 corresponds to beats that are most distant to the average QRS. We use the PQRST beat window with the highest signal quality as our candidate lead heartbeat template.

2.2. Feature Engineering

Our engineered features are 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 PQRST window during pre-processing. Heart rate variability features rely on the relative distances between each R-peak. Each extraction technique is performed independently per lead.

For full waveform and heartbeat template features, we use the cleaned ECG signal and apply the *tsfresh* feature extraction library. For full waveform features, we cap the signal sampling rate to a maximum of 500Hz before limiting the signal to the middle 2,000 samples to remove starting and trailing artifacts. Template features are derived from the isolated heart beat window with highest signal quality. Using the default feature extraction settings, we generate 763 template and 763 full waveform features per lead. The extracted features include autoregressive model coefficients, change quantiles, aggregate linear least-squares regression trends, peak counts, sample/approximation entropy, energy, continuous waveform transform coefficients, fast fourier transform coefficients, and other descriptive statistics of the signal.

Heart rate variability (HRV) features are generated from the cleaned signal and corresponding R-peak annotations using *NeuroKit2*. We used the default feature extraction

settings and generated 53 different HRV features per lead. HRV features include: mean, median, standard/absolute deviation, and interquartile range of the RR intervals; standard deviation of the successive differences between RR intervals; proportion of RR intervals greater than 50/20ms over total RR intervals; and geometric indices measuring triangular interpolation of the RR interval distribution.

For each 12-lead record we combine all three categories of engineered features with the age and sex parsed from the ECG record metadata. We arrive at a $12 \cdot (763 + 763 + 53) + 2 = 18,950$ length feature vector per 12-lead record.

2.3. Classification

We train a XGBoost binary classifier for each of the 27 clinical diagnosis, using `xgboost@1.1.1` [8]. We sample each training instance with a selection probability proportional to the regularized absolute value of the gradients. Early stopping is set to 20 rounds with binary logistic regression as our objective function.

We use the evaluation scoring weights as instance sample weights, capping positive examples to a 0.5 threshold. For example, when training the 1st degree atrioventricular block (IAVB) classifier we consider 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 have scoring function weights below 0.5 are treated as negative examples with a sample weight of 1. To account for the dataset label imbalance, we further scale the positive weight using the number of negative samples over the positive samples in the training set split. We use repeated random sub-sampling of our total dataset, randomly splitting our 43,101 records into an 85:15 training/validation set split.

We run 100 experiments using the full 18,950 features to determine feature importances. Feature importance is the model reported gain in accuracy contributed by the feature over all branches in the decision tree. We average the importances outputted by the 27 classifiers in each experiment to get the mean importance for each feature. We generate our challenge classification results by running 100 additional experiments, using only the top 1,000 most important features ranked by mean feature importance.

3. Results

A categorical visualization of our top 1,000 features, broken down by lead and feature type, is shown in Figure 3. We see that heartbeat template features are particularly important, with emphasis on leads `aVR` and `V1`.

Our methodology attained a mean challenge metric score of 0.484 on our validation splits. Additionally, we attained mean values for AUROC of 0.890, AUPRC of 0.390, accuracy of 0.251, overall F_1 score of 0.370, F_β of 0.427, and G_β measure of 0.222 using $\beta = 2$. An overview of these classification metrics can be found in Figure 4.

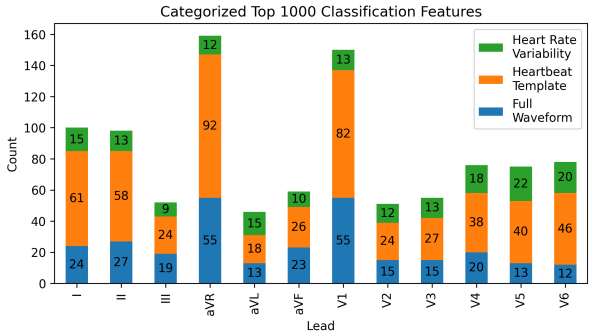

Figure 3. Count of lead and feature categories comprising the top 1,000 features. Age, but not sex, is important meta.

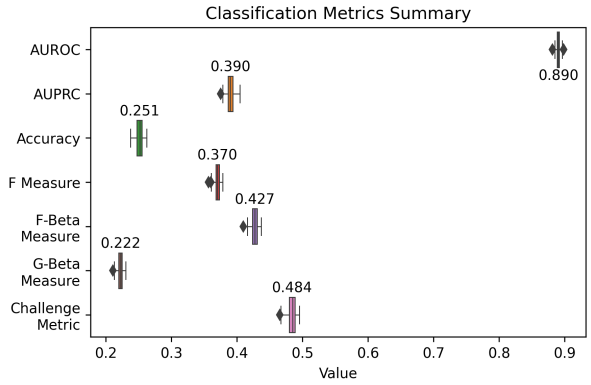

Figure 4. Summary of classification metrics over 100 experiments on all labels. Annotations indicate mean value.

Our model's top three best classified labels are normal sinus rhythm (SNR, F_1 mean: 0.920), left bundle branch block (LBBB, F_1 mean: 0.836), and sinus tachycardia (STach, F_1 mean: 0.779). A summary of our 100 experiment F_1 scores for each label is in Figure 5.

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 0.599 at a p-value of $9.7 \cdot 10^{-4}$. 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 0.476 .

4. Discussion & Future Work

Despite the label specific scaling of our dataset, the correlation between the label occurrence with the F_1 scores suggest further improvements are necessary to mitigate la-

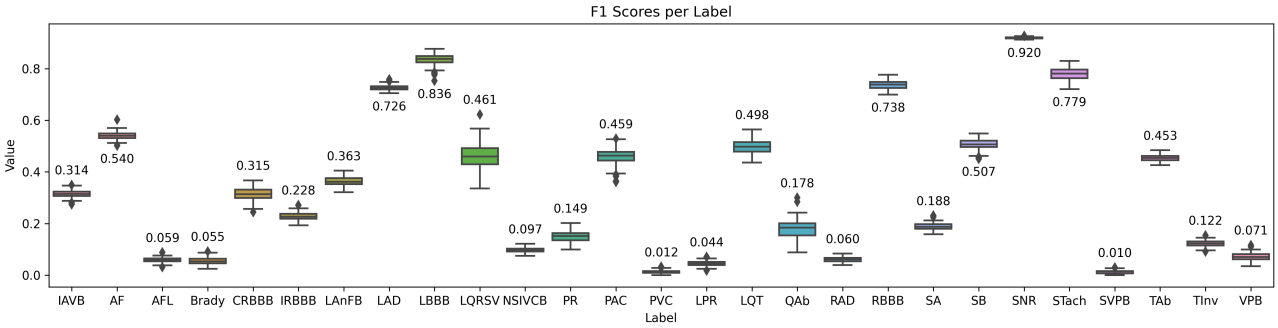

Figure 5. Label-wise F_1 scores over 100 experiments. Annotations indicate mean value.

bel 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 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 multi-dimensional time series data.

5. Conclusion

We create an algorithm for the classification of 27 heart conditions using signal processing inspired feature engineering and an XGBoost tree ensemble classifier. We combine a set of 18,950 features from full waveform, heartbeat template, and heart rate variability groups. Using 100 repeated random sub-sampling of 85:15 train/validation, we train models to get feature importances and distilled out 1,000 most important features. Using this reduced set of 1,000 features, we retrain our models and achieve a mean challenge score of 0.484 on our validation split. The official phase challenge score on the *PhysioNet/CinC* hold out test set is 0.476 for our team, *CVC*.

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] 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; .

[6] 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.

[7] 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.

[8] 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, 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