ECG classification combining conventional signal analysis, random forests and neural networks – a stacked learning scheme

Martin Baumgartner[1], Martin Kropf[2], Lukas Haider[1], Sai Veeranki[4], Dieter Hayn[1,3], Günter Schreier[1,4]

[1]AIT Austrian Institute of Technology, Center for Health & Bioresources, Graz, Austria
[2] Department of Internal Medicine and Cardiology, Charité University Medicine, Berlin, Germany
[3] Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria
[4] Institute of Neural Engineering, Graz University of Technology, Graz, Austria

Abstract

This year's Physionet Challenge focused on the question how many leads are required to develop a high-quality ECG classification algorithm. We (team name: easyG) propose a stacked learning scheme combining conventional signal analysis, random forests and neural networks. Highly specialized regression random forest models were trained with classical ECG processing where features were derived for each channel of each signal. The outputs were then used in a neural network to achieve a 1D regression vector, which was used to optimize classification thresholds.

We present offline validation results for each lead set and class-specific classification scores to allow for insights into the question how many leads are sufficient.

We have found that lead reduction leads to a minor loss in overall performance. However, variation in class-specific performance with lead reduction exists. Some classes were recognized better with more leads, but in rare cases, the opposite was true. The results suggest that the optimal number of used channels is depending on the setting and goals of the classification.

1. Introduction

"*Will two do*?" is the enticing research question of the 2021 Physionet Challenge [1]. This way of approaching a machine learning problem goes against the current dogma of modern artificial intelligence applications, where ever-increasing amounts of data are collected to achieve improved results. Condensing and reducing information to the utmost essential elements could help to alleviate the omnipresent issue of computational resource limitations.

While it is unlikely, that automated ECG classification algorithms will replace health care professions entirely, applying them can offer a range of benefits. First and foremost, such classifiers could provide medical personnel with a fast, initial assessment of a patient's health condition. In a further scenario, they could be used in long-term observation to warn patients and doctors if a recurring arrhythmia sets in. The fact that this challenge aims at reducing the number of required channels which makes the telehealth setting with wearable ECG recording equipment also a realistic example.

During the 2021 Physionet Challenge, the provided training data (n = 88.253) was comprised of 12-lead ECG data from different and heterogeneous sources. The hidden test set (n = 16.000) consisted of samples from these and additional undisclosed sources.

Past challenges [2], as well as a multitude of academic publications, have shown that both conventional machine learning approaches using feature engineering [3] as well as novel deep learning methods [4, 5] can achieve high-quality classification of ECG pathologies.

The combination of both machine learning paradigms seems appealing. In the 2020 challenge, a high-placed team ("*between a ROC and a heart place*") also combined deep learning residual networks with subjects' meta-features such as age and gender [6]. Our team followed a similar approach last year [7].

In contrast to the parallel use of conventional and deep learning as in the examples described above, this paper describes a combination of these paradigms in sequence.

2. Methods

We present a stacked feature engineering and modelling process that was comprised of four steps: 1) feature extraction, 2) random forest modelling, 3) neural network for channel mapping and 4) determining optimal classification thresholds.

2.1. Feature extraction

We based our feature engineering algorithm on our past signal analysis [8]. The features were calculated in both, time-domain and frequency-domain and can be categorized in different groups: *Averaged beat, single beat, atrial signal properties, rhythm related features, general signal properties, QRS related signals* and *meta-parameters derived as combinations of other features*.

For this work, additional features related to the heart axis concept were calculated. Since leads I and II are part of all channel subsets, these two were used as the axes on which the averaged heartbeats were projected in the frontal plane. From the resulting vector loop, a total of 10 additional parameters were calculated, i.e. mean values of the projections to the X and Y coordinates, the radius and angle of these values, the same procedures applied to the coordinates of the centre of gravity of the loop and, finally, the area and the perimeter of the loop.

Experiments on the impact of adding this set of additional parameters revealed that they were able to significantly increase the predictive power of models based on ECG analysis features.

2.2. Stacked learning scheme

Our three-stage modelling approach is depicted in Figure 1 below and is further explained in the following sections.

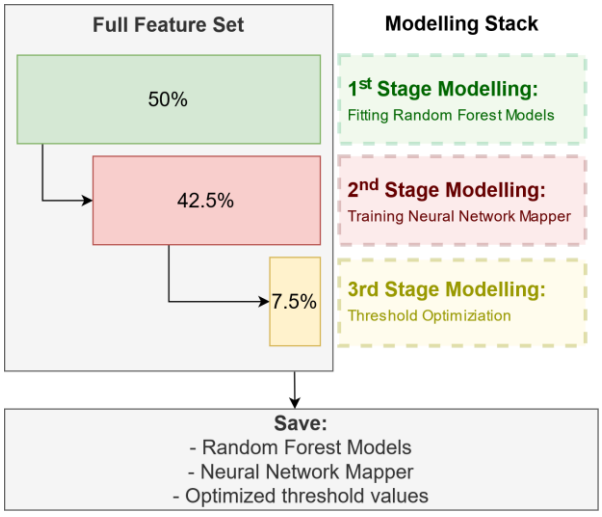

Figure 1. Schematic overview of data splitting and modelling scheme

2.2.1. Random Forest models

The derived feature set was split into two parts in a 50:50 ratio. One portion was used to fit random forest models. Highly specialized bagged regression ensembles with 64 trees each were created for each available data source (S), channel (C) and scored diagnosis (D). This resulted in S x C x D random forest models, which were used to predict the remaining 50% of the feature set, resulting in S x C x D probability scores for each channel of each signal to belong to each scored diagnosis.

2.2.2. Neural Network channel mapping

The regression results of the random forests of each signal had two dimensions [C, D] (No. of channels C, No. of diagnoses D). A multilayer perceptron (MLP) was applied to condense this 2D regression matrix to the required 1D vector of dimension [1, D]. The MLP had three fully-connected layers (units: 256, 128, 64) with a block of batch-normalization, ReLU activation and dropout (rates: 0.3, 0.25, 0.2) in between each layer. The final layer was a sigmoid-activated fully connected regression layer. As fully connected layers require 1D input, the 2D regression matrix was reshaped into a 1D vector of dimension [1, C x D] in order to be used as training data for the neural network. One neural network model was trained for each lead set. The models were trained for 50 epochs on 85% of the 50% (absolute portion 42.5%) portion of the feature set. The remaining 15% (absolute portion 7.5%) were used to determine classification thresholds.

2.2.3. Optimizing classification thresholds

To achieve binary classification results for each diagnosis, thresholds needed to be applied to the regression vector produced by the neural network. A grid search was applied to optimize individual, diagnosis-specific thresholds. The search iterated over each available diagnosis, setting an increasing threshold starting from 0 in 0.01 steps until 1, while setting all other, not yet optimized thresholds to 0.5. The threshold value that achieved the best challenge metric was recorded and fixed to that for the remaining iterations. Threshold optimization was executed for each lead set.

2.3. Querying of models

Since highly specialized models were developed in this stacked approach, each test set was predicted by the respective model (e.g. test samples from CPSC cohort were classified with the CPSC-trained models). However, this was not applicable for test samples of undisclosed sources as no models could be trained for them. In these cases, the models which have been developed with the most positive cases for each available diagnosis were selected. The rationale behind this approach was, that models that have been trained with many examples of that

diagnosis are likely to hold the most knowledge of that class and thus are best suited to classify this specific disease.

3. Results

All results presented in the following chapters are achieved with testing on 1% of each dataset and 100% of the St. Petersburg INCART set (n = 962). The remaining samples were used for training (n = 87.291). Due to technical issues, we cannot present online classification scores. Results for different lead-sets are summarized in Table 1.

3.1. Lead set results

Table 1. Results (F-score, AUROC and challenge metric CM) of the 5 different lead sets on offline data

Lead set	F-score	AUROC	CM	Won classes
12-lead	0.389	0.904	0.683	12
6-lead	0.386	0.903	0.669	5
4-lead	0.365	0.896	0.669	5
3-lead	0.375	0.890	0.658	8
2-lead	0.360	0.888	0.655	2

3.2. Class-specific results

Figure 2 shows the class-specific AUROC results with different lead sets.

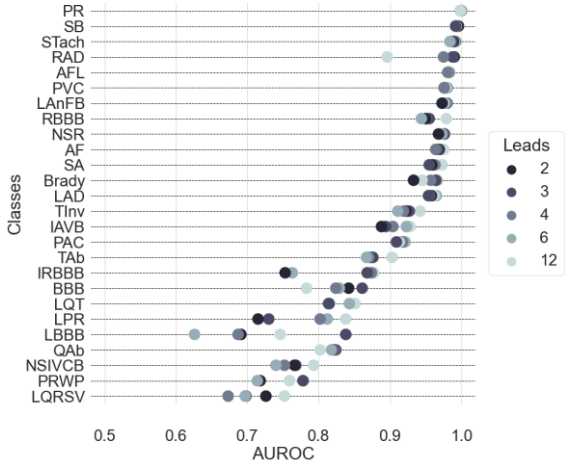

Figure 2. Class-specific AUROC results. The shade of colour indicates the number of channels (darker means fewer channels).

Overall, 12-lead ECG classification provided the highest score in 12 classes, the 6-lead and 4-lead sets in 5, 3-lead classification in 8 and 2-lead classification in 2 classes. This is summarized in Table 2.

Table 2. Lead sets in class-specific comparison. The second column indicates how many classes each respective lead set produced the highest score in.

Lead set	Won classes
12-lead	12
6-lead	5
4-lead	5
3-lead	8
2-lead	2

4. Discussion

Due to the fact, that unfortunately no valid submission was realized, all presented results are achieved with offline training data, which constitutes the main limitation of our work (our last submissions were successfully trained remotely, but ran into an error during testing, most probably only within the reduced lead sets). Therefore, the main focus of this work is to highlight an innovative methodology.

The feature engineering algorithm was mainly derived from previous work of our research group [8]. Originally, it was developed for single-lead ECG signals during the PhysioNet Challenge 2017 and received the second-best score against the hidden test set in the Physiological Measurement Focus Issue follow-up 2017. For this year's challenge, this algorithm was applied to all 12 available channels, which ultimately resulted in a large number of features (447 per channel, 5.364 in total). Many features were severely correlated. Some features provided hardly any benefits. As a next step, we will explore the influence of different approaches for feature selection on the results.

4.1. Lead set results

When examining Figure 2 it becomes obvious, that disparities in classes existed. Most of the classes that produced high AUROC scores are consistent among different lead sets. A notable exception was right axis deviation (RAD), which interestingly suffered heavily from using all 12 channels. It appears as if the 5 channels used in the 12-lead analysis only (V1, V3, V4, V5 and V6) were providing misleading features.

Other classes however did clearly benefit from the full 12-ECG. For example, right bundle branch blocks (RBBB) were distinctly better classified if all channels were available. Both T wave-specific classes T wave inversion (TInv), T wave abnormalities (TAb), as well as prolonged PR interval (LPR), nonspecific intraventricular conduction disorders (NSIVCB) and low QRS voltages (LQRSV) all clearly benefitted from using all 12 ECG leads.

Left branch bundle blocks (LBBB) constituted a curious case: 3-lead classification provided the best results, followed by 12-lead with a distinct margin. A similar

behaviour was found in the poor R wave progression class (PRWP). Further research is needed to provide a reasonable explanation for this phenomenon.

4.2. "Will two do?"

To answer the question of this challenge - "will two do?" - the scenario matters. Our results indicate, that using only two ECG channels is inferior to using the full 12-lead ECG although the margins are small (see Table 1). We have found that lead requirements differ between classes. Even though 2-lead classification provided the highest scores only in 2 classes, the main sentiment of lead reduction should not be rejected as 3-lead classification proved to be highly effective (see Table 2).

In our opinion, the answer is not entirely conclusive. While overall classification does not suffer from lead reduction as much, specific classes exist, that benefit from using more information (i.e. more leads). The found results suggest that the selection and number of channels should be adapted to the specific application of automated ECG classification.

References

[1] Reyna M., Sadr N., Gu A., et al., "Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet - Computing in Cardiology Challenge 2021", PhysioNet, 2021.

[2] Perez Alday EA., Gu A., J Shah A., et al., "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020.", Physiological Measurement, vol. 41, no. 12, pp. 124003, 2021.

[3] Maršánová L., Ronzhina M., Smíšek R., et al., "ECG features and methods for automatic classification of ventricular premature and ischemic heartbeats: A comprehensive experimental study", Scientific Reports, vol. 7, no. 1, pp. 11239, 2017.

[4] Zhai X. and Tin C., "Automated ECG Classification Using Dual Heartbeat Coupling Based on Convolutional Neural Network", IEEE Access, vol. 6, pp. 27465–27472, 2018.

[5] Wang T., Lu C., Sun Y., et al., "Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network.", Entropy (Basel, Switzerland), vol. 23, no. 1, 2021.

[6] Zhao Z., Fang H., Relton SD., et al., "Adaptive Lead Weighted ResNet Trained With Different Duration Signals for Classifying 12-lead ECGs", in 2020 Computing in Cardiology, 2020, 1–4.

[7] Baumgartner M., Hayn D., Ziegl A., et al., "Multi-Stream Deep Neural Network for 12-Lead ECG Classification", in 2020 Computing in Cardiology Conference, 2020.

[8] Kropf M., Hayn D., Morris D., et al., "Cardiac anomaly detection based on time and frequency domain features using tree-based classifiers", Physiological Measurement, vol. 39, no. 11, 2018.

Address for correspondence:

Martin Baumgartner
AIT Austrian Institute of Technology, Center for Health & Bioresources, Giefinggasse 4, 1210, Vienna, Austria
martin.baumgartner@ait.ac.at