An Ensemble Learning Approach to Detect Cardiac Abnormalities in ECG Data Irrespective of Lead Availability

Tim Uhlemann[1], Joshua Prim[1], Nils Gumpfer[1], Dimitri Grün[2], Sebastian Wegener[2], Sabrina Krug[1], Jennifer Hannig[1], Till Keller[2], Michael Guckert[1,3]

[1] Cognitive Information Systems, Kompetenzzentrum für Informationstechnologie, Technische Hochschule Mittelhessen, 61169 Friedberg, Germany
[2] Department of Internal Medicine I, Cardiology, Justus-Liebig-University Gießen, 35390 Gießen, Germany
[3] Department of MND - Mathematik, Naturwissenschaften und Datenverarbeitung, Technische Hochschule Mittelhessen, 61169 Friedberg, Germany

Abstract

Recent research has shown that artificial intelligence can detect heart pathologies when applied to electrocardiogram data. As of now, underlying architectures have mostly been built for specific problems with restricted generalisation to other patterns, e.g. convolutional neural network (CNN)-based models able to detect local patterns do not capture abnormalities with rhythmic dependencies well. Additionally, standard deep learning approaches cannot incorporate knowledge in non-deep learning representations. Aim of this project is to overcome limitations of distinct architectures by using a hybrid ensemble of models, also incorporating expert knowledge.

1. Introduction

The electrocardiogram (ECG) is a ubiquitously available, cost-effective diagnostic instrument for physicians. Machine and deep learning (ML and DL) methods have been successfully applied to detect complex often unspecific patterns that indicate severe problems with the heart [1]. However, these methods require large amounts of high quality data accurately labelled with diagnoses. In their daily routine physicians identify obvious signs for illness in the structures of the ECG by applying rules. Traditional formulas for *Long QT* and *Low Voltage QRS* are still in use and represent cardiological knowledge codified into applicable heuristics [2, 3]. In this paper, we demonstrate how a unified ensemble architecture can incorporate diagnostic engines of different type, e.g. DL models and heuristics, into a single prediction engine that can overcome limitations of each of its components in stand alone application.

2. Method

The core of our concept is an orchestration of architectures in which submodels working with sufficient precision for a subset of diseases are combined into a comprising computer aided diagnostic engine with comprehensible precision for a larger set of pathologies. Submodels working on subsets of available leads are combined into an ensemble which computes aggregated hypotheses for either regression or classification tasks [4]. Submodels can be either be homogeneous or heterogeneous allowing for different types of learners in the ensemble. Aggregation of the predictions may follow various paradigm voting, (weighted) averaging, or meta-learning [5]. With this approach hybrid architectures are possible in which also rule-based inference engines can be incorporated. For proof of concept, we have used a DL model derived from a previously published CNN model that successfully predict patients with myocardial scar based on ECG recordings [6]. With appropriate modifications this model can be applied to multiclass and multilabel tasks. This model is combined with simple exemplary established heuristics for specific diagnoses such as *Long QT* syndrome and *Low Voltage QRS* that fit well as their definition is based on few well defined ECG features. While the output of the DL model is a probability for a given label the heuristics give binary yes-no-answers. The ensemble applies (soft) vetoing, meaning that a non affirmative answer of the heuristics may overrule a positive response of the DL model.

In a multilabel problem for a given domain Z with a set of labels C each $z \in Z$ is labelled with a set of labels taken from C. Let $C(z) \subset C$ denote the set of labels assigned to $z \in Z$. A learner function s that classifies $z \in Z$ with $c \in C$ is defined as: $s\colon Z \times C \rightarrow [0, 1]$. The result of $s(z, c)$ is then transformed into the hypothesis

$H_s(z,c)$ according to:

$$H_s(z,c) = \begin{cases} 1 & \text{if } s(z,c) \geq \tau_a \\ -1 & \text{else} \end{cases} \quad (1)$$

so that $\{c \in C : s(z,c) \geq \tau_a\}$ (τ_a defaulting to 0.5). Let $H_s(z) \subset C$ denote the set of all predicted labels in C assigned to z by s. For a set of S of such learner functions mapping Z with classes C we define an ensemble E_S:

$$E_S(z,c) = sign(\sum_{s \in S} c_s \cdot H_s(z,c)) \quad (2)$$

The weights c_s can be defined, learned or predefined. We assume equal weights for each $H_s(z,c)$. For $s \in S$ we can now define that s has *soft veto priority* over s' in $E_S \setminus \{s\}$. The prediction of E_S is then defined as:

$$E_S(z,c) = sign(E_{-s,s'}(z,c) + (c_s + c_{s'})H_{s,s'(z,c)}) \quad (3)$$

where $E_{-s,s'}$ denotes the ensemble with prediction algorithms $S \setminus \{s, s'\}$ and $H_{s,s'(z,c)}$ is the hypothesis based on $s'(z,c) - \big(\tau_v - \tau_a\big)(1 - s(z,c))$. Furthermore, τ_v is a threshold above which an affirmative answer of s' will be accepted. The name soft veto priority indicates that such a predictor overrules affirmative results of another classifier which are too weak, i.e. for which the probability is too low.

Our ensemble $E = \{s_1, s_2, s_3\}$ applied to the ECG signals consists of a DL model s_1 based on our previously published CNN architecture adapted to the given multi-label task [6], a heuristic s_2 for detection of *Long QT*, and a second heuristic s_3 for *Low Voltage QRS*. The ensemble can process ECG recordings with different lead sets such as 12, 6, 4, 3, and 2 lead ECGs. While s_1 predicts a probability for each label in C, i.e. $s_1(z,c) \in [0,1]$ transformed to be -1 or 1, s_2 and s_3 are label specific, i.e. they return -1 for every label but the one representing *Long QT* syndrome and *Low Voltage QRS*, respectively. In that case $s_x(z,c)(x = 2,3)$ is either 0 or 1 indicating the presence or absence of the disease/symptom. In the ensemble s_2 an s_3 both have soft veto priority over s_1.

3. Final Architecture

Before ECG signals with e.g. different lead sets are fed into our ensemble model preprocessing is performed. During this preprocessing the sampling frequency and the length of all leads of the ECG signal are normalised so that heterogeneous input data can be used. Higher frequencies potentially lead to better results but induce higher hardware requirements. The sampling rate also determines the size of the input layer of our model. Table 1 describes details of that model for an example frequency of 200Hz. As this architecture is designed for multi-label problems, the output layer uses sigmoid activation instead of softmax. Note that the output layer contains 58 neurons as the model is designed to assign 29 different labels. Each pair of neurons therefore represents one of the possible labels. The first neuron in each pair (even index) represents the positive pressure while the second one (odd indexes) represents the negative pressure for its corresponding label.

Layer	Outputsize	Channels
Input	1000	12
Conv	1000	64
Pool	500	64
Conv	500	64
Pool	250	64
Dropout(15%)	250	64
Conv	250	64
Pool	125	64
Dropout(15%)	125	64
Conv	125	64
Pool	62	64
Dropout(15%)	62	64
Conv	62	64
Pool	31	64
Dropout(15%)	31	64
GAP	64	1
Dense	256	1
Dropout(31.5%)	256	1
Dense	256	1
Dropout(31.5%)	256	1
Dense	256	1
Dropout(31.5%)	256	1
Dense	58	1
Dense(Output)	58	1

Table 1. Architecture of the DL model s_1

Finally, the output of the s_1 is combined with the s_2 and s_3. Both exemplarily used heuristics depend on annotated ECG signals in which Q-, R-, S-, and T-Peaks and T-offsets are located. We annotate the raw signals using the *NeuroKit2* library (version 0.1.1) [7]. These annotated peaks are then used to estimate the length of the QT-interval (*Long QT*) and the voltages the QRS-complexes (*Low Voltage QRS*). Both heuristics have soft veto priority and overrule affirmative results of the DL model if the corresponding probabilities are below the threshold value τ_v. τ_v can be tuned during hyperparameter optimisation.

4. Results and Discussion

In the official phase of PhysioNet/CinC Challenge, an earlier version of our proposed ensemble achieved a challenge metric score of 0.318, 0.246, 0.214, 0.259, and 0.257 for the 12, 6, 4, 3, and 2 lead inputs, respectively (Team

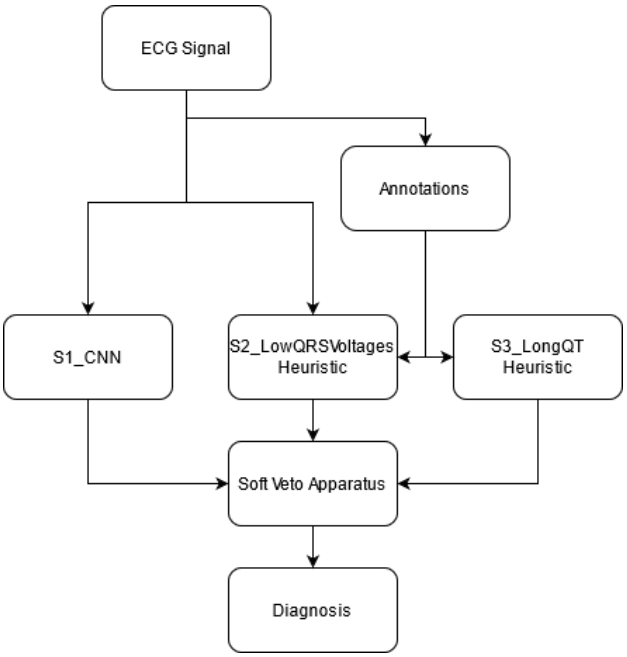


Figure 1. Ensemble Architecture

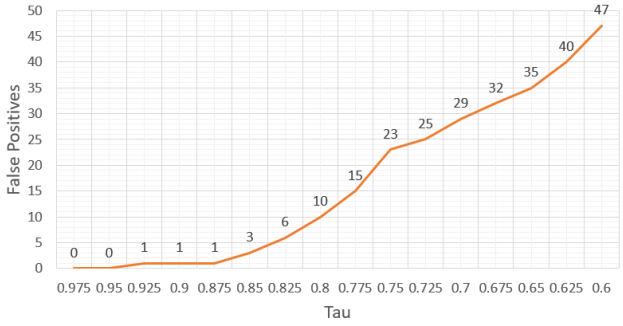
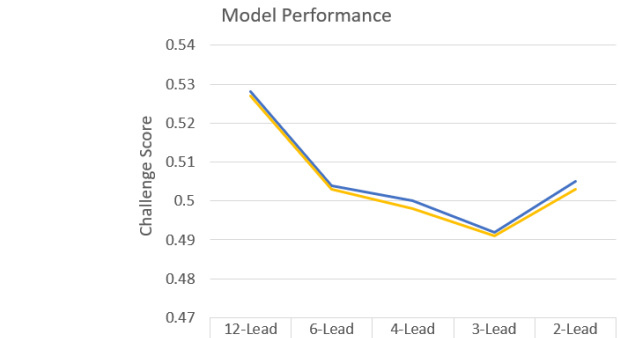


Figure 2. False positives in Long QT and Low Voltage QRS for values of τ_v.

CardioIQ). During this challenge the model was trained with data from multiple datasets: CPSC2018, CPSC2018-Extra, St Petersburg, PTB Diagnostic, PTB-XL, Georgia 12-Lead ECG Challenge Database, Chapman University (Shaoxing People's Hospital), and Ningbo first Hospital. Due to hardware limitations imposed by the challenge infrastructure ECG data had to be downsampled to 100Hz (for s_1) with a sample length of five seconds (only for s_1, s_2 and s_3 use full length).

In local tests the St Petersburg data source was skipped, as this dataset only contains 75 ECG recordings of a lower native sampling rate than any other database. Data used in local experiments is resampled to 200Hz with a sample length of five seconds. Data was split into training- (75%) and validation set (25%).

Figure 3 shows the results achieved by the ensemble



	12-Lead	6-Lead	4-Lead	3-Lead	2-Lead
Challenge Metrik(Only S1)	0.528	0.504	0.5	0.492	0.505
Challenge Metrik(S1, S2, S3)	0.527	0.503	0.498	0.491	0.503

Figure 3. Results for s_1 and the ensemble

with the standalone DL model on the 12-, 6-, 4-, 3-, and 2-lead datasets. On 12-Lead validation sets our ensemble achieved values of (0.527 for the 2021 PhysioNet challenge metric [1]) and an AUROC of 0.932.

Our approach strongly depends on the precision of the annotation algorithm. While the *NeuroKit2* library achieves high precision in determining R-peaks, we observed imprecise or missing locations of the Q-, S-peaks, and T-offsets that impair the results of our heuristics. Figure 4 shows how annotations may be wrong or lack precision when applied to reals-world ECG signals from the PhysioNet challenge datasets that deviate from an ideal ECG signal. Here, Q-peaks are missing and a T-peak has mistakenly been annotated as an R-peak. As the heuristics estimate interval lengths, e.g. the QT interval, and measure voltages, e.g. maximal voltages in QRS-complex, by using the annotations the quality of the diagnoses obviously suffers from such inaccuracies.

For proof of concept, we therefore have experimented with *perfect* heuristics that just mirror the label that has manually been assigned to the recordings. Figure 2 shows the number of FP (false positive) classifications of our ensemble for different values of τ_v when using the *perfect* heuristics for *Long QT* and *Low Voltage QRS*.

Overall, our results support the hypothesis that single model architectures that perform well on specific diseases can be improved by amending restrictions of their generalisation power on elementary different diseases by combining specifically trained expert models together with disease specific simple algorithmic implementations in an ensemble with an appropriate aggregation. This is a promising approach to outperform singular models.

[1] https://physionetchallenges.org/2021/

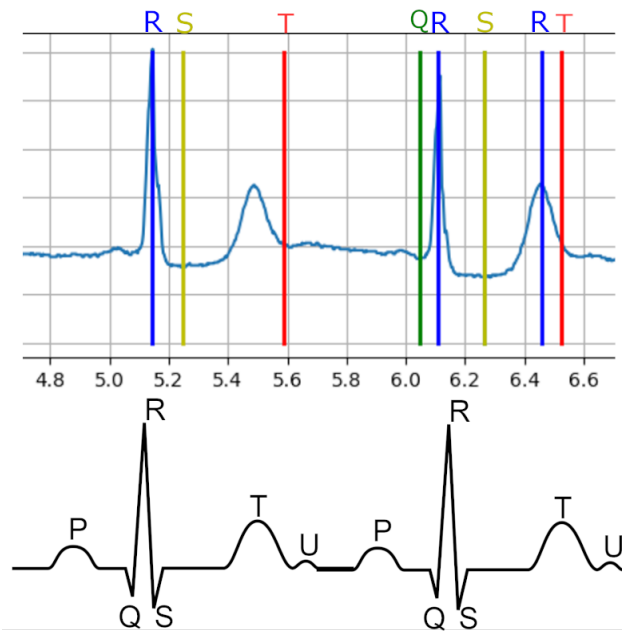


Figure 4. Automated annotations (upper figure) vs. ideal ECG signal (lower figure).

5. Conclusion and Future Work

In this paper, we discussed an architecture that can incorporate different types of diagnostic algorithms working on ECG recordings with various lead sets into a single prediction engine. Diagnostic algorithms can either be rule based, ML or DL models or simple rules of thumb taken from the daily routine of the cardiologist. Our proof of concept shows that the ensemble built with a DL model and two simple heuristics can potentially have a high diagnostic leverage. Heuristic and rule based approaches are particularly sensitive to the real world data problem. Automated annotation algorithms must become more robust to deviation from an ideal ECG signal. Here, we further see potential for DL algorithms for improving such annotation algorithms. Furthermore, future research will extend the set of prediction mechanisms to be included in the ensemble and elaborate on how the results will be combined. Using additional knowledge of cardiological experts encoded in machine readable form, e.g. in ontologies, can potentially be used to control and fine tune the ensemble mechanisms. For example, knowledge about a disease and the affected anatomy - e.g. information about and how and where a disease is mirrored in the ECG - can be used to determine which predictions of the elements of the ensemble to use and how to combine them into a single result.

Acknowledgements

This work has partially been funded by Flexi Funds (Forschungscampus Mittelhessen).

References

[1] Grün D, Rudolph F, Gumpfer N, Hannig J, Elsner LK, Von Jeinsen B, Hamm CW, Rieth A, Guckert M, Keller T. Identifying Heart Failure in ECG Data with Artificial Intelligence-a Meta-Analysis. Frontiers in Digital Health 2020;2:67.

[2] Bazett HC. An analysis of time relations of electrocardiograms. Heart 1920;7:353–367.

[3] Fridericia LS. Die systolendauer im elektrokardiogramm bei normalen menschen und bei herzkranken. Acta Med Scand 1920;53:469–486.

[4] Opitz D.; Maclin R. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 1999; 11:1689–1696.

[5] Zhou H. Ensemble Methods - Foundations and Algorithms. Chapman and Hall - CRC Press, 2012.

[6] Gumpfer N, Grün D, Hannig J, Keller T, Guckert M. Detecting myocardial scar using electrocardiogram data and deep neural networks. Biological Chemistry 2020;402(8):911–923.

[7] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA. NeuroKit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods feb 2021;53(4):1689–1696.

Address for correspondence:

Michael Guckert
Wilhelm-Leuschner-Straße 13, 61169 Friedberg (Hessen)
michael.guckert@mnd.thm.de