Can Supervised Learning Be Used to Classify Cardiac Rhythms?

Marcus Vollmer[1], Philipp Sodmann[1], Leonard Caanitz[1], Neetika Nath[1,2], Lars Kaderali[1]

[1] Institute of Bioinformatics, University Medicine Greifswald, Germany
[2] Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Germany

Abstract

Background: *This contribution relates to the PhysioNet/CinC Challenge 2017 on classification of atrial fibrillation from short single lead ECG recordings. The aim is to assign an ECG to one of these classes: normal sinus rhythm, atrial fibrillation, an alternative rhythm, or too noisy.*

Methods: *We trained a convolutional neural network using waveforms of the QRS complex, P waves, T waves, noise and inter-beat time series of labeled data from PhysioNet in order to derive an accurate detection of the characteristic components of normal and arrhythmic heart beats. To identify rhythm patterns, a noise estimation function was used in combination with heart rate and analysis of RR, RT and PR intervals. We analyzed the cross-correlation of heart beat shapes by clustering the resulting correlation matrix in order to distinguish normal and abnormal heart beats that might cause rhythm changes. We examined the feature importance and used a random forest algorithm to generate a decision tree to predict the rhythm class. The classification performance was evaluated using F_1 scores.*

Results: *The convolutional neural network was able to correctly identify more than 99% of all R peaks in the QT database, whereas the detection of P and T waves reached a true positive rate of 91% and 81% respectively. The classification performance in 8,528 records of the training data set was F_1=0.94. An overall score of 0.81 was achieved when applying the algorithm to the hidden test set of the challenge.*

1. Introduction

Analyzing the heart rate variability, which is a physiological phenomenon of heart beat variation over time, is used to determine autonomic activity of a heart. Disorders in the regular heart rate as a result of disturbances in the electrical system of the heart are called arrhythmia. Expert cardiologists can identify such a physiological variation of the heart rate by analyzing the ECG leads (electrocardiogram) and thereby diagnose different cardiac disorders.

Several machine learning algorithms were proposed (e.g. [1]) to classify ECG samples to arrhythmia classes, based on the features extracted from the ECG. The QRS-complex, P and T waves are the most important features to extract from the ECG as they have specific characteristic waveforms and are dominating the amplitude. A normal healthy heart rhythm can be identified by a specific order: P wave, QRS-complex and T wave, which appear at defined and regular time intervals. Characteristic for atrial fibrillation are irregular RR intervals, no distinct P waves and usually variable intervals between two atrial activations at $>300\,\mathrm{bpm}$ [2]. ECG signals can capture deflections because of the anatomical difference of the atria and the ventricles, their sequential activation, depolarization, and repolarization. Therefore, it is important to correctly annotate R, P and T peaks in order to perform correct prediction of cardiovascular diseases using supervised learning algorithms. However, annotating P and T peaks is difficult especially in the presence of noise.

This motivates us to develop a convolutional neural network (CNN) to annotate such biomedical signals. Based on these annotations derived from short ECG recordings, our goal is to classify the hidden test set provided by the PhysioNet/CinC Challenge 2017 [3]. The aim of the challenge is to assign the ECGs to one of these classes: normal sinus rhythm (N), atrial fibrillation (A), an alternative rhythm (O), or too noisy to classify ($\sim$).

2. Methods

In order to achieve our aim, first a CNN was used to annotate ECGs, followed by feature extraction. These features were used as input variables in random forests to classify unlabeled ECG records.

The task for the CNN is to analyze an input ECG signal $X=[x_1,\ldots,x_n]$ of length n to generate an output annotation sequence $a=[a_1,\ldots,a_n]$, where a could be either R, P, T, $\sim$, O or an unknown type. Therefore, labeled datasets were used as input layers for the CNN for training purposes. R, P, T waves, and interbeat segments

 ISSN: 2325-887X DOI:10.22489/CinC.2017.347-176

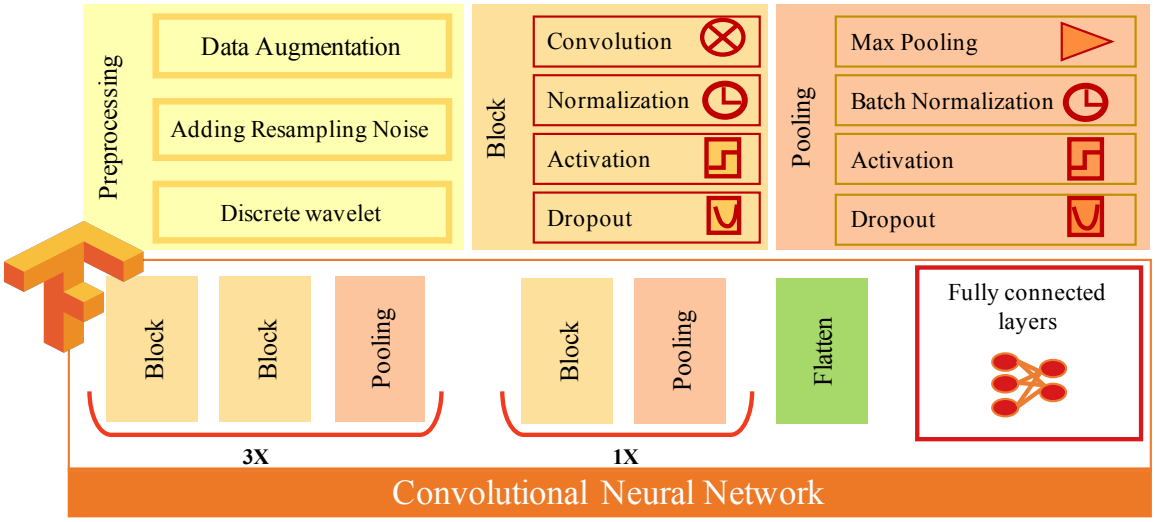

Figure 1. Architecture of the convolutional neural network using the open-source software library TensorFlow, which was used to learn features (R, P, T wave) of ECGs. In the convolution block, we used the hyperbolic tangent as an activation function of the neurons followed by PReLU activation (parametric rectified linear unit). The output of the neurons from the convolution block was combined using maximum pooling in the pooling section.

were taken from the QT database [4]. We included realistic noisy segments, which were generated by applying the WFDB function `nst` [5] to clean records at different and very low signal-to-noise ratios (see [6]). Additionally, ECG records of extrasystoles and other arrhythmic beats O were extracted from the MIT-BIH arrhythmia database [7]. The next step in our workflow was to perform data augmentation. Data augmentation was performed with shift of ± 17 ms window size to generate data replicates. Furthermore, synthetic white noise was included in order to build a robust model. In total about 12,000,000 characteristic waveforms served as training examples. The architecture of the CNN is illustrated in Figure 1.

For training purposes of the classification task, the challenge organizers provided 8,528 single lead ECG recordings with a record length up to 60 s. These recordings were collected using AliveCor devices and were labeled by a single expert. Finally, the CNN model was used to annotate the unlabeled peaks of each single ECG record and 174 features were extracted.

Feature extraction From the annotated records, the absolute value, percentiles, and interquartile range were computed for the RR, the RT, and the PR intervals. Further, these features were normalized to their relative intervals, defined as successive differences divided by their mean (according to [8]). The absolute counts and percentage of extrasystoles with and without compensatory pause, doublets, and triplets were also added as features. In order to identify extra beats in the annotated records, we used the relative RR intervals and classification rules based on relations of successive intervals as proposed in [9]. We defined the complexity (entropy) of RR intervals by computing the standard deviation of the shortened relative RR intervals, from which we removed detected extrasystoles. Furthermore, we defined the entropy on higher grades, by considering a lag when computing relative RR intervals. Additional features were generated by adjusting interval data by heart rate that was estimated by the 25% trimmed mean of RR intervals of the records. In order to use shape information, we computed the maximum cross correlation for each pair of heart beat waveforms. After that, we conducted k-Means and hierarchical clustering (average linkage, euclidean metric) on the basis of the correlation matrix and extracted basic cluster characteristics like the silhouette score and distance information.

Random forest Once the features are extracted, they are fed to a random forest that is implemented in R version 3.4.1 [10] to classify the four rhythms (N, A, O, $\sim$). We applied 10 repeats of 10-fold cross validation to the training dataset with 174 features and evaluated the performance based on F_1 scores, defined below in equation 1. Also, the hyperparameters are optimized based on the best F_1 score. Moreover, the variable importance is generated from the final model.

Evaluation criteria The classification performance was evaluated for the training data set and the hidden test set of the competition [3]. The confusion matrix was build and the F_1 score is computed by

$$F_1 = \frac{F_1(N) + F_1(A) + F_1(O)}{3}. \qquad (1)$$

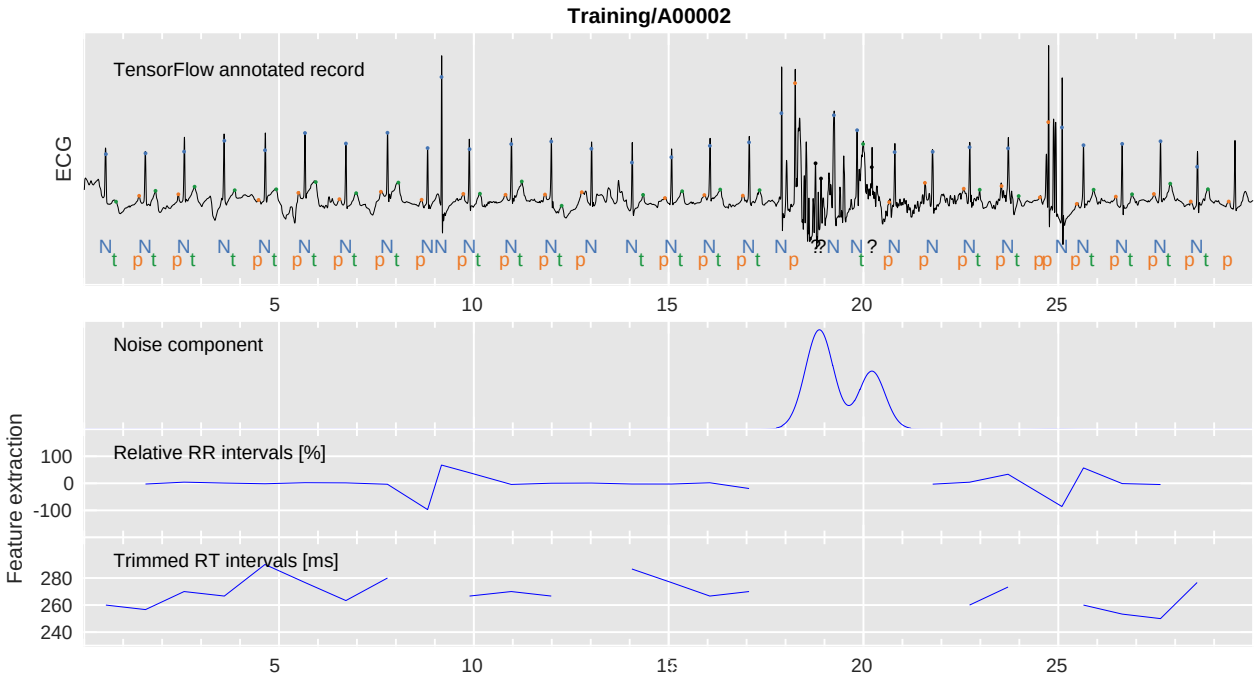

Figure 2. Features were extracted by analyzing the annotated records of the training set. A convolutional neural network was trained to annotate the ECG records using the TensorFlow. In this training data set (A00002), the noise component is elevated around the 20th second and arrhythmia are indicated by deflections in the sequences of relative RR intervals at 9th and 25th second. The RT interval time is at a constant level of about 260 ms.

The scoring system treats all classes equally and is giving the accuracy on the basis of precision and recall. Partial scores are giving by $F_1(x)$ for different types of classes $x \in \{N, A, O, \sim\}$ and are defined by true positive (TP), false positive (FP) and false negative (FN) counts:

$$F_1(x) = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \qquad (2)$$

3. Results

We evaluated the true positive rate and positive predictivity of our CNN model by comparing the resulting annotations with the reference annotations given for the QT database [4]. A strict tolerance level of 25 ms and 50 ms was set for the time difference between both annotations to count as successful (true positive). Some T wave annotations occur twice in the reference annotation file so that we have previously removed reference annotations which refer not to the T wave peaks. Table 1 reports the CNN annotation performance for both tolerance levels in 82 records. The accuracy in the detection of R peaks was very high with a true positive rate of 98.7% and a positive predictive value of 99.7%, given a tolerance of 50 ms. Also, the detection of P waves works very well, 91.0% of all P waves were correctly annotated while producing a low false negative rate of 4.7%. T wave detection was the hardest of all, the sensitivity is 80.7% only. Fortunately, the positive predictive value of 89.8% is quite high, such that the effects on the feature extraction process is not that serious, but leads to interval sequences with missing values as seen in Figure 2. Table 2 reports the overall and partial F_1 scores for the training and test set. Noisy records are the hardest of all to predict – of 284 noisy labeled records 99 were falsely classified as A and 94 were classified as O. This affects also the partial scores of A and O.

Table 1. Annotation performance for QTdb

		R	P	T
Reference counts		86892	78665	88013
Model counts		86020	75126	79047
True positive rate	25 ms	0.981	0.886	0.732
(sensitivity)	50 ms	0.987	0.910	0.807
Positive	25 ms	0.991	0.928	0.815
predictive value	50 ms	0.997	0.953	0.898

Table 2. Classification performance measured by F_1

	Overall	N	A	O	∼
Training	0.94	0.97	0.94	0.93	0.46
Test set	0.81[1]	0.90[2]	0.80[2]	0.67[2]	NA

[1] Final result on 3,658 recordings, excluding $F_1(\sim)$
[2] Challenge phase scoring based on 1,000 recordings

Feature importance The important features were identified through random forests and were further analyzed. The top 10 features are listed in Figure 3. Some of these features are specialized to distinguish a single rhythm from the other three classes. The complexity (entropy) of relative RR intervals [8] for instance, showed a strong discriminative power to separate normal rhythms from the others. Counts of extrasystoles are useful in order distinguish normal rhythms from atrial fibrillation and other rhythms. The 90% quantile of RR intervals (adjusted by heart rate) was found to be useful to discriminate atrial fibrillation from other rhythms. Noisy records were strongly associated with cluster complexity measures, as defined by the correlation matrix, which is not included in the top 10 list as it is sorted by the overall importance.

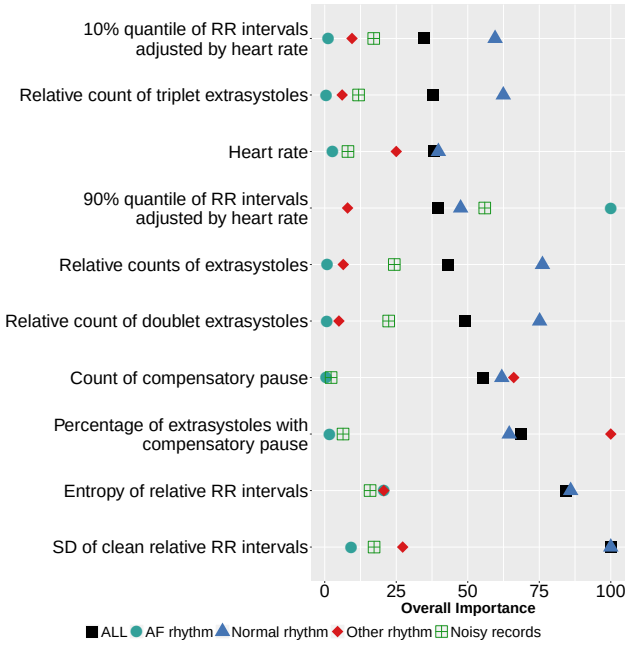

Figure 3. Features in ascending order by its overall importance and differentiated according to rhythm classes.

4. Conclusion

The combination of supervised deep learning (CNN) for the annotation of ECGs and an ensemble method (random forests) for the classification of rhythms has shown remarkable results in the training and hidden test set. Nevertheless, we see opportunities to improve and extend the quality of annotation, especially the detection of P and T waves were just moderate. A better sensitivity could be achieved with more accurate input data and an optimization of batch sizes. Weak results are observed in the classification of noisy records, for which the used features are not specific enough. We noted, the occurrence of noise prior or subsequent to atrial fibrillation could be a reason for misclassification. We believe that the combination of machine-learning techniques in pre- and postprocessing tasks and hand-crafted feature generation with human knowledge is appropriate for to classify cardiac rhythms. In the way we have solved the task, one is able to identify the causes for misclassification. In contrast to total black box systems, weaknesses can be easily identified and improvements can be made by implementing more specific or new features in order to the increase the accuracy. This is the way how we want to strengthen the trust in using modern analyzing techniques in ECG processing.

References

[1] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. In eprint arXiv:1707.01836. 2017; .

[2] Camm AJ, Kirchhof P, Lip GY, Schotten U, Savelieva I, Ernst S, Van Gelder IC, Al-Attar N, Hindricks G, Prendergast B, et al. Guidelines for the management of atrial fibrillation: the Task Force for the Management of Atrial Fibrillation of the European Society of Cardiology (ESC). European heart journal 2010;31(19):2369–429.

[3] Clifford G, Liu C, Moody B, Silva I, Li Q, Johnson A, Mark R. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge 2017. In Computing in Cardiology, volume 44. 2017; in press.

[4] Laguna P, Mark RG, Goldberg A, Moody GB. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In Computers in Cardiology. 1997; 673–676.

[5] Moody GB. WFDB applications guide. Harvard-MIT Division of Health Sciences and Technology, 10 edition, 2003.

[6] Vollmer M. Noise Resistance of Several Top-Scored Heart Beat Detectors. In Computing in Cardiology, volume 44. 2017; in press.

[7] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine 2001;20(3):45–50.

[8] Vollmer M. A Robust, Simple and Reliable Measure of Heart Rate Variability using Relative RR Intervals. In Computing in Cardiology, volume 42. 2015; 609–612.

[9] Vollmer M. Arrhythmia Classification in Long-Term Data Using Relative RR Intervals. In Computing in Cardiology, volume 44. 2017; in press.

[10] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

Address for correspondence:

Marcus Vollmer / marcus.vollmer@uni-greifswald.de
Institute of Bioinformatics / University Medicine Greifswald
Walther-Rathenau-Str. 48 / 17475 Greifswald / Germany