A Comprehensive Study of Complexity and Performance of Automatic Detection of Atrial Fibrillation: Classification of Long ECG Recordings Based on the PhysioNet Computing in Cardiology Challenge 2017

Denis Kleyko , Evgeny Osipov and Urban Wiklund

1 Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden
2 Department of Radiation Sciences, Biomedical Engineering, Umeå University, Umeå, Sweden

E-mail: urban.wiklund@umu.se

Keywords: atrial fibrillation detection, computational complexity, feature selection, 2017 PhysioNet/CinC Challenge

Supplementary material for this article is available online

Abstract

Objective: The 2017 PhysioNet/CinC Challenge focused on automatic classification of atrial fibrillation (AF) in short ECGs. This study aimed to evaluate the use of the data and results from the challenge for detection of AF in longer ECGs, taken from three other PhysioNet datasets. 

Approach: The used data-driven models were based on features extracted from ECG recordings, calculated according to three solutions from the challenge. A Random Forest classifier was trained with the data from the challenge. The performance was evaluated on all non-overlapping 30 s segments in all recordings from three MIT-BIH datasets. Fifty-six models were trained using different feature sets, both before and after applying three feature reduction techniques.

Main Results: Based on rhythm annotations, the AF proportion was 0.00 in the MIT-BIH Normal Sinus Rhythm (N = 46083 segments), 0.10 in the MIT-BIH Arrhythmia (N = 2880), and 0.41 in the MIT-BIH Atrial Fibrillation (N = 28104) dataset. For the best performing model, the corresponding detected proportions of AF were 0.01, 0.10 and 0.38 when using the 15 best performing features.

Significance: The results obtained on the MIT-BIH datasets indicate that the training data and solutions from the 2017 Physionet/Cinc Challenge can be useful tools for developing robust AF detectors also in longer ECG recordings, even when using a low number of carefully selected features. The use of feature selection allows significantly reducing the number of features while preserving the classification performance, which can be important when building low-complexity AF classifiers on ECG devices with constrained computational and energy resources.

1. Introduction

Atrial fibrillation (AF) is a supraventricular tachyarrhythmia which is characterized by a very irregular and fast heart rate. AF is one of the most common form of heart rhythm disorders, where the occurrence among the general population is 1%–2%. The frequency of AF is increasing with age and is expected to continue to grow (Lloyd-Jones et al 2004). AF has a severe influence on health conditions and could cause stroke, congestive heart failure, and even increase the risk of death. Moreover, AF often can be undetected since it does not always cause symptoms and the irregular heart rate can also be difficult to differentiate from normal sinus rhythm. Therefore, developing methods for automatic detection of AF is an important task.

However, automatic detection of AF in single-lead ECG recordings is a complex problem, where state-of-the-art performance is typically achieved by machine learning methods. The 2017 PhysioNet/CinC Challenge (CinC2017) was devoted to this problem (Clifford et al 2017), and several of the best-performing solutions have been published along with the source code. The training dataset that was used in the competition has already become a popular source for the...
development and evaluation of new classifiers of AF in short ECG recordings. The publicly available training dataset consists of 8528 ECG recordings of 9-30 s duration, recorded with a resource-constrained device. The recordings were classified using both a voting procedure between different automatic algorithms and manual verification of many but not all annotations (Clifford et al 2017). The following four rhythm classes were defined: AF; Normal, normal sinus rhythm; Other, other abnormal rhythm including both recordings with frequent extra systolic beats and low beat-to-beat variability in heart rate; and Noise, very noisy recordings where the ECG could not be analyzed. However, there seems to be a substantial overlap between the four classes that were defined, in particular between the Normal and Other classes, but partly also between the AF and Other classes (Parva-neh et al 2018, Christov et al 2018). Still the best solutions to the challenge obtained an overall F1 score of 0.89 on the training dataset.

One problem from a computational point of view, is that many of the proposed solutions were based on deriving a large number of features (e.g., several hundreds). This is rather natural since the main objective of the challenge was to achieve as high classification performance as possible. At the same time, potential clinical applications are nearly real-time in their nature. As such they require substantially more light weight solutions. Another strong reason for reducing computational complexity is data privacy, which calls for solutions to be deployed directly on resource-constrained devices that make the recordings. For example, in (Christov et al 2018) the top 15 features were ranked using a statistical method. Feature elimination has also been applied in other studies. In a preliminary study (Abdukalikova et al 2018), we used the Recursive Feature Elimination method to reduce the number of features to 15 using the feature set proposed in (Andreotti et al 2017).

The aim of the present study is twofold. Firstly, it presents a comprehensive analysis of the computational complexity of solutions for automatic classification of AF in ECG recordings. Secondly, it presents the results of applying solutions trained with the CinC2017 dataset to other datasets for AF detection. To the best of our knowledge, this has never been done before. The main approach used in the study is to train a classifier using the CinC2017 dataset, and then evaluate how it performs for detection of AF in recordings from three other Physionet datasets: 1) The MIT-BIH Normal Sinus Rhythm database; 2) The MIT-BIH Atrial Fibrillation database; and 3) The MIT-BIH Arrhythmia database. We evaluate the performance of one popular classifier that was used in the challenge: Random Forest. This classifier is trained using different sets of feature variables from three of the solutions proposed for the challenge (where the source code is available online (Datta et al 2017, Zabihi et al 2017, Andreotti et al 2017)). The computational complexity problem is addressed by applying several feature selection methods, which allow reducing the total number of features being used.

The article is organized as follows: section 2 describes the materials used during the training and evaluation of the studied models, as well as introduces the methods used for processing the materials and obtaining the models. Results are presented in section 3. The findings of this study are discussed and placed in the context of other related work in section 4. The article is concluded in section 5.
2. Methods and Materials

2.1. Overview of the study

Figure 1 illustrates a workflow of the materials and methods being used and their interrelations. As in any data-driven study there are two phases: training and evaluation. Note that in this study, the training and testing data consist of different datasets: the CinC2017 dataset is used for training, while three MIT-BIH datasets are used for evaluation. This is done on purpose since the goal of the study was to investigate the applicability of models trained on the short recordings from the training dataset for detecting AF in long recordings.

All data-driven models considered in this study are based on features extracted from ECG recordings. In particular, four feature sets were used. Three sets of features were extracted according to solutions proposed for the 2017 PhysioNet/CinC Challenge (see subsection 2.3 for details). The fourth set aggregated all features extracted by the three solutions into a single set. Each set has been subject to one of the feature selection processes as described below. The features were used to train a classification model using a Random Forest classifier. The cross-validation process estimated the performance of a trained model on the CinC2017 dataset. The recordings in the MIT-BIH datasets include annotations, such as the start and end of periods with AF. In order to obtain the detection performance of the trained models, these annotations were compared to the predicted class labels issued by the trained models.

2.2. Materials

The study material in this retrospective study consisted of ECG recordings in adult subjects, including healthy controls, patients with different types of arrhythmia, and very noisy recordings. In total, we used four different datasets from the PhysioNet database (Goldberger et al 2000).

Albeit some of the evaluation datasets included multi-channel ECG recordings, the ECG analysis considered only single (first)-lead recordings. The subsections below introduce each dataset.

2.2.1. CinC2017 dataset

The 2017 Physionet/CinC Challenge dataset3 was used to train models evaluated in this study. In the sequel, this is referred to as the CinC2017 dataset. The goal of the challenge was to tackle one of the main limitations of the previous AF detection studies. Namely, the fact that the previous results have been reported for binary classifications (AF versus normal sinus rhythm) and small-sized datasets comprising of long high-quality ECG recordings. The challenge attempted to address these limitations by developing solutions for AF detection using a dataset with four different types of short single-lead ECG recordings. The duration of recordings varied between 9 s and 61 s. The sampling frequency was 300 Hz. Note that the use of short single-lead ECGs makes it harder to detect AF, since usually ECG signals are long and recorded with several leads. When compared to the previous studies, e.g., (Mohebbi and Ghassami 2008) and (Park et al 2009), an important advantage of the CinC2017 dataset is the very large number of included recordings, consisting of 8528 recordings for training and 3658 recordings for testing (not disclosed to the public) (Clifford et al 2017). No information is available regarding the age and gender of the subjects from whom the recordings originated. The distribution of the recordings in the training dataset into the four defined classes were: Normal ($N = 5076$), Other ($N = 2415$), AF ($N = 758$), and Noise ($N = 279$).

The evaluation of the best performing solutions (Clifford et al 2017) demonstrated that it was possible to achieve an $F_1$ score of 0.89 and 0.83 on the training and testing sets respectively. Most of the solutions to the challenge were formed by extracting a large number of features and then using them to train conventional machine learning methods (including those described in subsection 2.3).

2.2.2. The MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia dataset (MIT-BIH-ARR) (Moody and Mark 2001) is a standard test material for evaluation of arrhythmia detectors. It consists of 48 ambulatory two-channel ECG recordings sampled at 360Hz. The duration of each recording is 30 min. The dataset was collected from 47 subjects. The subjects were 25 men aged 32 to 89 years, and 22 women aged 23 to 89 years. It is worth noting that this dataset includes both recordings with different degree of arrhythmias and recordings with normal sinus rhythm, as well as 4 recordings with a pace-maker. Each recording includes two types of a reference annotation made by experts: rhythm annotations and beat annotations.

2.2.3. The MIT-BIH Normal Sinus Rhythm Database

MIT-BIH Normal Sinus Rhythm dataset (MIT-BIH-NSR) (Goldberger et al 2000) consists of 18 ambulatory long-term ECG recordings sampled at 128Hz. The duration of each recording is about 24 h. The dataset was collected from healthy adult subjects without significant arrhythmias. The subjects were 5 men aged 26 to 45 years, and 13 women aged 20 to 50 years. Both rhythm and beat annotations are available also for this dataset.

2.2.4. The MIT-BIH Atrial Fibrillation Database

MIT-BIH Atrial Fibrillation dataset (MIT-BIH-AF) (Moody and Mark 1983) consists of 23 ambulatory long-term two-channel ECG recordings sampled at 250Hz. The duration of each recording is about 10 h.
The dataset was collected from subjects with AF. However, their age and gender were not available. This dataset only includes manually prepared rhythm annotation files, which indicate one of four rhythms: atrial fibrillation, atrial flutter, AV junctional rhythm or all other rhythms.

2.2.5. Processing of MIT-BIH datasets recordings
The length of the ECG recordings in MIT-BIH datasets varied from 30 min to 24 h. However, the calculation of features and the optimization of classifiers in the solutions to the CinC2017 Challenge were based on short recordings of 30 s duration. Therefore, the MIT-BIH recordings were also divided into 30 s non-overlapping segments, where each segment was treated as a short independent recording. The corresponding features were then calculated and all segments were classified into one of the four rhythm classes. The proportion of segments in different classes was determined for both all and for each individual recording in the datasets.

To assess the performance of feature-based predictions, we had to define the “ground truth” for the MIT-BIH datasets. However, to the best of our knowledge, the detailed criteria for defining different classes in the CinC2017 dataset have not been published. Therefore, we first labeled the MIT-BIH datasets based on the rhythm annotations that were available for all three datasets. These annotations were given as the time when a change in rhythm occurred, as well as the type of the new rhythm, e.g., a change from normal sinus rhythm to AF. The following five rhythm classes were defined: (a) normal sinus rhythm; (b) AF (including atrial flutter); (c) other rhythm (all other types of arrhythmias); (d) noise; and (e) paced rhythm. The duration of noise segments was based on the provided signal quality indicator for the first ECG channel. See www.physionet.org for more information regarding the annotation of rhythm classes. The total duration of each type of rhythm was determined, as well the proportion of the total time. In addition, individual segments were labeled as ‘normal’ only if they contained 30 s with normal sinus rhythm. Otherwise, they were labelled as one of the other four rhythm classes.

We also used the annotations of individual beats when defining the “ground truth” for the MIT-BIH datasets. In this study, we defined nine additional segments in the MIT-BIH-ARR database as “Other” since the rhythm was annotated as normal but it included more than five arrhythmic beats of ventricular or supraventricular origin. Note that the rhythm was only annotated as normal or atrial fibrillation/flutter (except in a few recordings) and no annotations of beat types were available for the MIT-BIH-AF dataset, which had the consequence that the proportion of the Other class was underestimated for this dataset.

2.3. Feature sets derivation
This subsection describes the feature sets that were extracted using the Matlab source code from three solutions from the Challenge. It is worth noting that heart beats in ECG recordings were detected using the different detection algorithms that were provided by each solution. Moreover, all features were z-scored using the corresponding parameters (mean and standard deviation) obtained for the CinC2017 dataset.

2.3.1. Andreotti feature set
The first feature set consisted of 171 different features from filtered and segmented ECG recordings (Andreotti et al 2017)\(^4\), where the number of segments depends on the length of the recording. Features for all segments belonging to the same recording were summarized using 16 different measures (e.g., mean, standard deviation, etc.). Thus, the total number of features per recording could be as large as 2736. However, in this study we only present results based on the mean values across all segments of a recording (i.e., 171 features per recording), since there was virtually no difference in classification performance. The features were extracted using heart rate variability (HRV) metrics, signal quality metrics, and also morphological ones. In addition to time domain, frequency domain, and non-linear HRV metrics, metrics based on clustering of beats on Poincare plots were also used.

2.3.2. Zabihi feature set
The second feature set was also based on hand-crafted features (Zabihi et al 2017)\(^5\). The preprocessing part included baseline wander removal and denoising of a recording. Next, 491 different features were extracted from a recording. These features were ranked based on their importance selecting only the 150 highest-ranked features. Here we used only the set of ranked features (i.e., 150 features per recording). The used features were extracted from the time, frequency and time-frequency domains, and from phase space reconstruction of ECG recordings.

2.3.3. Datta feature set
The third feature set consisted of 188 features that were extracted according to the solution presented in (Datta et al 2017)\(^6\). The preprocessing part included locating and discarding noisy portions of a recording using its spectrogram, and filtering the resultant recording. The extracted features belong to the following domains: HRV metrics, morphological metrics, prior art AF metrics, frequency-domain metrics, statistical metrics, and abnormality metrics.

\[^4\] Available online: https://github.com/fernandoandreotti/cinc-challenge2017.
\[^5\] Available online: https://physionet.org/challenge/2017/sources/morteza-zabihi-208.zip.
\[^6\] Available online: https://physionet.org/challenge/2017/sources/shreyasi-datta-209.zip.
2.3.4. Aggregated feature set
The aggregated set included all features from the three considered solutions to the challenge and, thus, it consisted of 509 features.

2.4. Random Forest classification
Random Forest is considered as a powerful classification technique, which demonstrates high classification performance for real-life feature-based problems (Fernandez-Delgado et al 2014). For example, it was used by two of the considered solutions: (Zabih et al 2017) and (Andreottti et al 2017). We also relied on Random Forest in this study. Random Forest is an ensemble classifier. It makes a prediction by combining outcomes of several classifiers, which are trained independently on the training dataset. Individual classifiers are trained using a decision tree approach. During the experiments, the classificationEnsemble method in Matlab with 30 decision trees (otherwise default settings) was used to train Random Forest models. For a detailed description of the technique please see, for example, chapter 18.10 in (Russell and Norvig 2010).

2.5. Complexity reduction via feature variables selection
All feature sets presented in the previous subsection included a large number of features. However, extracting a large number of features is impractical due to computational and time constraints, in particular for real-time detection of AF with systems deployed on resource-constrained devices. Therefore, in order to lower the computational burden, in this study we considered three feature selection methods for reducing the number of features. Thus, for new recordings it would only be necessary to extract features chosen after the selection process.

2.5.1. Recursive Feature Elimination
The Recursive Feature Elimination (RFE) method (Guyon et al 2002) is a greedy optimization technique used to find the subset of best performing features. It repeatedly builds classification models, keeps the best feature subsets and puts the worst ones aside and afterwards computes the accuracy. This process is repeated until all features are exhausted. Next, RFE evaluates and ranks all the features based on the order of their elimination. Finally, it provides the indices of the best performing features, which form the subset of the pre-specified size.

It is worth noting that the features forming the best performing subset by means of their ranking are not necessarily individually the most important. These features perform well only in combination with the other features in the corresponding subset. During the experiments, the RFE function from Python scikit-learn machine learning library was used to implement the method.

Table 1. Confusion matrix for the 2017 PhysioNet/CinC Challenge.

|          | AF  | Normal | Other | Noise | Total |
|----------|-----|--------|-------|-------|-------|
| Actual   |     |        |       |       |       |
| Normal   |     |        |       |       |       |
| Other    |     |        |       |       |       |
| Noise    |     |        |       |       |       |
| Total    |     |        |       |       |       |

2.5.2. Neighborhood Component Analysis
The Neighborhood Component Analysis (NCA) method (Yang et al 2012) is a non-parametric method for selecting features maximizing the accuracy of predictions. It features some similarity to the nearest neighbor classification and, therefore, the neighborhood term is used. The NCA method calculates a weight for each feature, where the weigh value determines the significance of the corresponding feature for classification. Thus, the weights of less relevant features would be closer to zero and it is possible to form a subset of the desired size by choosing the features with the largest weights. During the experiments, the fscnca method in Matlab was used to implement the NCA method.

2.5.3. Statistical approach
The statistical approach included the calculation of $p$-values for each feature using Kruskal-Wallis and multiple comparison tests. The $p$-values were calculated as the sum of all scores after the multiple comparison test. It allows selecting features where classes differ. Then the desired number of features could be chosen among the features with the lowest $p$-values. During the experiments, kruskalwallis and multcompare methods in Matlab were used to implement the statistical approach.

2.6. Performance metrics
The performance of a trained classification model was evaluated based on the confusion matrix, which is a table contrasting the ‘ground truth’ against the prediction results. Table 1 presents the confusion matrix for the classes used in the CinC2017. Bold entries in the main diagonal of the table denote number of correct predictions, while off-diagonal entries denote all possible misclassifications. In principle, the classification performance of a model is characterized by the confusion matrix. However, in the case of several classes it becomes convenient to have a single numeric metric for the comparison of different models. In the challenge, the mean $F_1$ score was used as the overall performance metric of a model, which is based on the individual $F_1$ scores for each class (rhythm):
Table 2. Classification of the CinC challenge data into 4 classes. Averaged over 10 independent runs. Numbers are Accuracy/F1 Score.

| Complexity | Solution | Andreotti | Zabihi | Datta | Aggregated |
|------------|----------|-----------|--------|-------|------------|
|            |          | 0.828/0.740 | 0.857/0.780 | 0.835/0.740 | 0.863/ 0.787 |
| Moderate   | RFE, 15  | 0.824/0.743 | 0.846/0.767 | 0.829/0.737 | 0.860/0.791 |
|            | NCA, 15  | 0.805/0.701 | 0.814/0.719 | 0.810/0.708 | 0.850/0.779 |
|            | STAT, 15 | 0.783/0.692 | 0.812/0.726 | 0.774/0.682 | 0.747/0.675 |
| Low        | RFE, 5   | 0.765/0.619 | 0.809/0.706 | 0.795/0.668 | 0.817/0.718 |
|            | NCA, 5   | 0.762/0.669 | 0.762/0.671 | 0.740/0.589 | 0.777/0.684 |
|            | STAT, 5  | 0.641/0.511 | 0.763/0.643 | 0.631/0.501 | 0.668/0.590 |

- AF rhythm: $F_{1a} = \frac{2Aa}{\Sigma A + \Sigma a}$.
- Normal rhythm: $F_{1n} = \frac{2Nn}{\Sigma N + \Sigma n}$.
- Other rhythm: $F_{1o} = \frac{2Oo}{\Sigma O + \Sigma o}$.
- Noisy: $F_{1p} = \frac{2Op}{\Sigma O + \Sigma p}$.

The mean $F_1$ score is given by: $F_1 = \frac{F_{1a} + F_{1n} + F_{1o} + F_{1p}}{4}$.

Note that $F_1$ score of an ideal classifier equals 1, therefore, a model with the higher $F_1$ score is preferable.

Besides mean $F_1$ score, we also considered accuracy as another performance metric. The accuracy (denoted as $A$) was calculated as:

$$A = \frac{Aa + Nn + Oo + Op}{\Sigma A + \Sigma N + \Sigma O + \Sigma P}.$$

When reporting the results for the MIT-BIH datasets we used the overall proportions of predictions for each class, as well as bar charts for contrasting the annotations and predictions in individual recordings.

3. Results

This section first presents the cross-validation results obtained for data-driven models after training with the CinC2017 dataset; using the different features sets before and after applying the feature selection methods. Then we present the evaluation results for the ECG recordings from the MIT-BIH datasets.

3.1. Cross-validation performance on the CinC2017 dataset

Table 2 presents the cross-validation performance of 28 different model configurations trained with Random Forest. Each of the four feature sets was used in seven configurations: using all features and using the 5 and 15 best performing features for each of the three feature selection methods (RFE, NCA and statistical approach, STAT). Please see the Supplementary materials for details of the chosen features. All results reported below were obtained using 5-fold cross validation on the CinC2017 training dataset.

We observe that when considering all features, the Zabihi feature set demonstrated the highest performance among the individual sets. However, the aggregated set performed slightly better. This is also true for the feature selections methods (except STAT), since for each configuration the subset selected from the aggregated set demonstrated the highest performance. Moreover, for a given number of features, the RFE method showed a better performance than both the NCA method and the statistical approach. With respect to the number of selected features, it is clear that when choosing only 5 features both accuracy and $F_1$ score degraded significantly. At the same time, already 15 features chosen by the RFE method performed almost as good as all features for the corresponding feature set. In particular, the highest reduction in accuracy was observed for Zabihi feature set but the decrease was only 1.3 percent. Importantly, 15 features selected by the RFE method from the aggregated dataset performed better than any model trained on all features from the individual sets.

Since Table 2 presents only aggregated performance metrics, it is worth considering the corresponding confusion matrices. Due to space limitations, only two configurations with high performance are shown: the aggregated set with all features (Table 3), and 15 features from the aggregated set chosen by the RFE method (Table 4), respectively.

The common characteristic in both tables was that the CinC2017 dataset appeared to have a significant overlap between the Normal and Other classes. This characteristic of the dataset has been also observed in other studies (Parvaneh et al. 2018, Christov et al. 2018). Since we have chosen to put the highest priority on AF detection, the large overlap motivated us to merge the Normal and Other classes, and then study the performance also in the case of three classes.

Table 5 presents the corresponding cross-validation performance for the case of three classes: AF,
In the case of four classes, the improvements in the performance of the feature elimination methods were similar to the ones obtained in the case of three classes. The 15 features chosen by the RFE and NCA methods performed consistently better for all four feature sets when classifying into three classes. This is in contrast to the case of four classes, where all features gave slightly higher mean \( F_1 \) scores than after feature selection. The differences are marginal but still this indicates that many of the introduced features appeared to have a negative effect of the classification performance.

### 3.2 Evaluation on the MIT-BIH datasets

This subsection evaluates the models selected in the previous subsection (all features and 15 RFE features) using the MIT-BIH datasets. The total number of segments was 2880, 28104, and 46083 for the MIT-BIH-ARR, MIT-BIH-AF, and MIT-BIH-NSR datasets, respectively.

| Complexity | Solution   | Andreotti | Zabihi | Datta | Aggregated |
|------------|------------|-----------|--------|-------|------------|
| High       | All features | 0.943/0.769 | 0.950/0.794 | 0.944/0.763 | 0.951/0.803 |
| Moderate   | RFE, 15    | 0.945/0.787 | 0.952/0.813 | 0.945/0.778 | 0.955/0.829 |
|            | NCA, 15    | 0.942/0.779 | 0.950/0.807 | 0.941/0.747 | 0.950/0.811 |
|            | STAT, 15   | 0.762/0.669 | 0.762/0.671 | 0.740/0.731 | 0.777/0.639 |
| Low        | RFE, 5     | 0.939/0.761 | 0.942/0.781 | 0.930/0.733 | 0.944/0.804 |
|            | NCA, 5     | 0.901/0.641 | 0.941/0.765 | 0.932/0.740 | 0.900/0.761 |
|            | STAT, 5    | 0.876/0.502 | 0.874/0.478 | 0.920/0.701 | 0.876/0.504 |

Normal + Other, and Noise. Combining the Normal and Other classes had a large effect on the accuracy since the main source of misclassifications was mitigated. At the same time, the improvements in the mean \( F_1 \) score were less notable. For example, for 15 features from the aggregated set chosen by the RFE method the accuracy increased from 0.860 to 0.955, whereas the mean \( F_1 \) score changed from 0.791 to 0.829, i.e., the improvements were 11.0 and 4.8 percent, respectively. This is because there was still an overlap between the classes, such as between AF and Noise or between AF and Normal + Other. Therefore, individual \( F_1 \) scores of AF and Noise classes did not increase dramatically. The observed results with respect to the performance of the feature elimination methods were similar to the ones obtained in the case of four classes. The 15 features chosen by the RFE and NCA methods performed consistently better for all four feature sets when classifying into three classes.
3.2.1. Proportions of predictions per dataset

The classification of the MIT-BIH datasets is presented as the proportion of segments that were classified into each class for the case of four classes (Table 6) and three classes (Table 7), respectively. While the tables do not yet contrast the predictions with the “ground truth”, it is still possible to make insightful observations.

Table 6. Classification of the MIT-BIH datasets into four classes (AF/Normal/Other/Noise), Averaged over 10 independent runs. Numbers are proportions for each class. All features and 15 features selected by the RFE.

| Solution | Andreotti | Zabihi | Datta | Aggregated |
|----------|-----------|--------|-------|------------|
| ARR/All  | 0.06/0.32/0.62/0.01 | 0.06/0.37/0.54/0.03 | 0.11/0.39/0.50/0.01 | 0.08/0.31/0.60/0.01 |
| AF/All   | 0.21/0.24/0.55/0.01 | 0.25/0.33/0.34/0.08 | 0.36/0.34/0.29/0.01 | 0.31/0.28/0.40/0.01 |
| NSR/All  | 0.00/0.42/0.58/0.00 | 0.00/0.62/0.38/0.00 | 0.00/0.75/0.25/0.00 | 0.00/0.61/0.39/0.00 |
| ARR/RFE  | 0.05/0.38/0.56/0.01 | 0.09/0.42/0.45/0.05 | 0.12/0.44/0.44/0.00 | 0.09/0.53/0.37/0.01 |
| AF/RFE   | 0.25/0.21/0.52/0.01 | 0.27/0.22/0.42/0.08 | 0.38/0.37/0.24/0.01 | 0.33/0.35/0.29/0.03 |
| NSR/RFE  | 0.00/0.74/0.26/0.00 | 0.01/0.57/0.42/0.00 | 0.01/0.90/0.09/0.00 | 0.00/0.84/0.16/0.00 |

Table 7. Classification of the MIT-BIH datasets into three classes (AF/Normal and Other/Noise). Averaged over 10 independent runs. Numbers are proportions for each class. All features and 15 features selected by the RFE.

| Solution | Andreotti | Zabihi | Datta | Aggregated |
|----------|-----------|--------|-------|------------|
| ARR/All  | 0.07/0.93/0.00 | 0.06/0.92/0.02 | 0.09/0.90/0.01 | 0.07/0.92/0.01 |
| AF/All   | 0.18/0.81/0.01 | 0.25/0.71/0.04 | 0.35/0.64/0.01 | 0.29/0.69/0.01 |
| NSR/All  | 0.01/0.99/0.00 | 0.00/1.00/0.00 | 0.00/1.00/0.00 | 0.00/1.00/0.00 |
| ARR/RFE  | 0.04/0.96/0.00 | 0.08/0.91/0.01 | 0.10/0.89/0.01 | 0.07/0.92/0.01 |
| AF/RFE   | 0.21/0.79/0.01 | 0.28/0.68/0.04 | 0.35/0.63/0.02 | 0.25/0.74/0.01 |
| NSR/RFE  | 0.00/1.00/0.00 | 0.00/1.00/0.00 | 0.00/1.00/0.00 | 0.00/1.00/0.00 |

3.2.1. Proportions of predictions per dataset

The classification of the MIT-BIH datasets is presented as the proportion of segments that were classified into each class for the case of four classes (Table 6) and three classes (Table 7), respectively. While the tables do not yet contrast the predictions with the “ground truth”, it is still possible to make insightful observations.

Recall that for all feature sets the results for the CinC2017 dataset from all features and 15 features chosen by the RFE are comparable. In the case of MIT-BIH datasets, however, we could observe notable differences in favour of the RFE method, especially when considering the case of four classes. The MIT-BIH-NSR dataset is intuitive in the sense that it is not expected to include any signs of AF or arrhythmia. Thus, most of the segments in each recording should be classified as Normal. However in the MIT-BIH-NSR dataset, the proportions of segments classified as Other varied between 0.25 to 0.58 when using all features. On the other hand, features chosen by the RFE resulted in significantly lower proportion of Other for three of the feature sets (except Zabihi). For example, in the case of the aggregated set it decreased from 0.39 for all features to only 0.16 for 15 RFE features. It is also worth mentioning that the Datta features after RFE showed a low proportion of Other (0.09), while the other two individual feature sets had higher numbers: 0.26 for Andreotti and 0.42 for Zabihi. At the same time, the Zabihi and Datta features after RFE demonstrated non-zero proportion of AF predictions (0.01), which indicates that these subsets of features could potentially be biased towards higher false positive rates when detecting AF.

As an estimate of the “ground truth” of the proportion of AF in the MIT-BIH datasets, we calculated the total duration of sequences with AF (based on the rhythm annotations), which was divided with the total duration of all recordings. For the MIT-BIH-AF database, the proportion of time with AF was 0.41, 0.59 was marked as normal sinus rhythm, and a few short sequences was marked as noise. For the MIT-BIH-ARR database, 0.10 was marked as AF, 0.71 was normal sinus rhythm, 0.10 was other non-sinus rhythms, 0.08 was paced rhythm, and 0.01 was noise.

In the case of the MIT-BIH-AF dataset, there was a large variation in the predicted AF proportion. The highest proportion and best agreement with the annotations of AF was observed for the Datta feature set with four classes, where the predicted proportion of AF was 0.38 using all features, and 0.35 using RFE features. The lowest proportion 0.18 was predicted by all Andreotti features. For the MIT-BIH-ARR dataset, the best agreement with the annotated proportion of AF was also obtained when using the Datta feature set, although several other feature sets presented with nearly the same predicted AF proportion. Note that for each feature set, the AF proportions in each dataset were similar when using all features and RFE subsets, as well as for three and four classes, which allows...
concluding that differences in predicted AF proportions should be attributed to a particular feature set.

It is hard to make strong statements regarding the classification of the other rhythms in the two datasets with arrhythmias. Nevertheless, there are several potentially insightful observations. As shown in Table 6, the predicted proportion of the Other rhythm was rather high and showed a large variation, both when comparing feature sets and when comparing all features versus the RFE features. For the MIT-BIH-AF data set, the proportion of the Other rhythm varied between 0.24 to 0.55 and for the MIT-BIH-ARR data set the variation was in the range 0.37–0.62. The proportion of Noise was relatively low in both the MIT-BIH-AF and MIT-BIH-ARR datasets. The highest proportion was found for the models based on the Zabihi feature set (0.04 for three classes, and 0.08 for four classes), whereas the majority of the other models demonstrated less than 0.01 proportion, which was in agreement with the annotated time with noise for these data sets.

Finally, when considering the predictions made for the MIT-BIH-ARR dataset in the case of three classes, there was a large agreement in the predicted proportions of different rhythms between the models using all features versus those based on RFE features. The predicted proportion of the Other rhythm was lowest for Datta features, and highest for Andreotti features.

3.2.2. Proportions of predictions per individual recording

Figures 2–7 contrast the proportions of the annotations in individual recordings of the MIT-BIH datasets against the predictions by each feature set for the case of 15 RFE features. Please note that the legends of the figures include Paced type, which is a part of the annotations for four recordings in the MIT-BIH-ARR dataset. It is being kept for consistency reasons in two other datasets.

Figures 2 and 3 present the results for the MIT-BIH-NSR dataset for the case of four and three classes, respectively. Similar to the results in Table 6, Figure 2 demonstrates that the largest proportion of Other was predicted by Zabihi feature set while the lowest one by Datta feature set. At the same time, the largest proportion of AF false positives (sixth recording) was also attributed to Datta feature set. The aggregated feature set was the second lowest when it comes to the proportion of Other. Notably, the predictions of Other were not distributed uniformly, for example, for the aggregated feature set there were two recordings (5 and 16)
Figure 4. Predictions of the MIT-BIH-AF dataset into four classes, with the best 15 features chosen by the RFE. Bars show the proportion of different rhythms in each recording. Left-to-right order: annotations, Andreotti feature set predictions, Zabihi feature set predictions, Datta feature set predictions, the aggregated feature set predictions.

Figure 5. Predictions of the MIT-BIH-AF dataset into three classes, with the best 15 features chosen by the RFE. Bars show the proportion of different rhythms in each recording. Left-to-right order: annotations, Andreotti feature set predictions, Zabihi feature set predictions, Datta feature set predictions, the aggregated feature set predictions.

Figure 6. Predictions of the MIT-BIH-ARR dataset into four classes, with the best 15 features chosen by the RFE. Bars show the proportion of different rhythms in each recording. Left-to-right order: annotations, Andreotti feature set predictions, Zabihi feature set predictions, Datta feature set predictions, the aggregated feature set predictions.
with more than half of the segments being classified as Other.

Merging Normal and Other classes simplified the patterns observed in Figure 3 since most of the segments were predicted as the new class. When it comes to Noise prediction, only the second recording included a relatively large proportion, and all models were able to predict a similar amount of Noise. It is also worth noting that the small (but notable) amount of AF false positives in the predictions by Zabihi feature set was concentrated to a single (tenth) recording.

Figures 4 and 5 present the results for the MIT-BIH-AF dataset for the case of four and three classes, respectively. The situation with the Other predictions was similar to Figure 2 with the difference that the lowest average proportion of Other was 0.24 for Datta feature set (cf 0.09 for the MIT-BIH-NSR by the same feature set). The predictions of AF are very similar for both figures. Among the considered features set, the one from Datta was the most consistent one with the annotations. It only highly underestimated (more that 0.10) AF in two recordings (7 and 11). For example, for the other feature sets there were seven, nine, and seven such recordings for Andretti, Zahibi, and aggregated, respectively. The situation was similar with respect to false positives. The highest overestimation for each feature set (in order of their appearance) was 0.03 (first recording), 0.11 (nineteenth recording), 0.07 (thirteenth recording), and 0.05 (thirteenth recording), respectively.

Figures 6 and 7 present the results for the MIT-BIH-ARR dataset for the case of four and three classes, respectively. It is easier to see the situation with AF for the case of three classes in Figure 7. Similar to the MIT-BIH-AF dataset, Datta feature set predictions were the most consistent ones. In particular, the total mismatch of AF predictions (sum of overestimation and underestimation) on all 48 recordings for each feature set (in order of their appearance) was 3.41, 1.44, 1.08, and 1.34, respectively. For the case of four classes the corresponding scores were: 2.46, 1.89, 1.44, and 1.06, respectively, thus, the aggregated feature set was more consistent. The predictions of Other were much less accurate compared to AF. The following mismatch scores were observed for the considered datasets: 19.33, 15.04, 16.60, and 13.55, respectively, which also indicates that for the case of four classes the aggregated feature set performed best.

Since the individual segments for each MIT-BIH dataset also were annotated, the classification performance is also presented as the corresponding confusion matrices. Based on the results for classification into three classes from 10 independent runs, the highest overall mean accuracy for all datasets was obtained for the Datta feature set, where the accuracy (SD) was 0.95 ± 0.002 and the mean $F_1$ score was 0.66 ± 0.006. The corresponding values for the Andretti feature set were 0.90 ± 0.005 and 0.59 ± 0.011, while they were 0.92 ± 0.011 and 0.62 ± 0.021 for the Zabihi features, and 0.91 ± 0.013 and 0.60 ± 0.026 the aggregated feature set. The mean sensitivity for detecting AF varied between models: 0.50 ± 0.04 for Andretti; 0.67 ± 0.07 for Zabihi; 0.83 ± 0.01 for Datta; and 0.58 ± 0.08 for the aggregated feature set, whereas the corresponding mean specificity was high for all models (between 0.997 and 0.999). Table 8 presents the matrices for the case of three classes for 15 RFE features from Datta feature set. As shown in the table, the observations
made for individual recordings also hold for the classification performance obtained for the confusion matrices. In the MIT-BIH-NSR and MIT-BIH-ARR datasets there were a small number of segments that were annotated as Normal or Other but were classified as AF. In the MIT-BIH-NSR dataset, segments in four recordings were classified as AF (1–3 segments in three recordings, and 25 segments in one recording). There were 11 recordings in the MIT-BIH-ARR dataset where false positive detections of AF were found, but 10 of these recordings had annotated AF segments in other segments. AF was not detected in six of the recordings in the MIT-BIH-ARR dataset, but the annotated proportion of AF was less than 0.02 in all of them (corresponding to approximately 30 s of the total time). Finally, AF was detected in all recordings in the MIT-BIH-AF dataset, although two recordings had less than 5 segments where AF was detected.

Due to the uncertainty in the labeling of the Other class, the classification into four classes resulted in lower overall accuracy. Again, the Datta feature set presented with the best performance: mean accuracy 0.81 ± 0.005, whereas the accuracy was 0.67 ± 0.05 for Andreotti, 0.53 ± 0.02 for Zahibi, and 0.75 ± 0.009 for the aggregated feature sets. The mean sensitivity for detecting AF was highest for Datta (0.89 ± 0.01), followed by the aggregated feature set (0.75 ± 0.05), Zahibi (0.61 ± 0.07) and Andreotti (0.56 ± 0.04). The mean specificity for detecting AF was between 0.991 and 0.998 for the different models. The corresponding confusion matrices are shown in Table 9. Only the confusion matrices for the Datta feature set are shown as it also demonstrated the lowest average mismatch score per recording when detecting AF in the MIT-BIH-AF and MIT-BIH-ARR datasets. The corresponding tables for the four classes case for the other feature sets are available in the Supplementary materials (Tables S.4–S.6).

4. Discussion

This study has focused on two aspects when using the results of the CinC2017: the effects of reducing the computational complexity of the solutions, and the performance of classification models trained on the CinC2017 dataset for AF detection in other datasets. For this purpose, we used an AF classifier based on feature sets proposed by three solutions, as well as their aggregation, and evaluated its performance using three MIT-BIH datasets from Physionet. Since the solutions were developed based on short ECGs, all recordings were divided into 30 s non-overlapping segments. Another reason for this segmentation, is that AF can be either permanent or occur in bursts, where the latter would be more difficult to detect if features were calculated based on the complete long recording.

Overall, we observed that it was possible to correctly detect AF in the MIT-BIH datasets using models trained with the CinC2017 dataset. At the same time, some of the feature sets performed better than others. In particular, we found that the Datta feature set presented with the best agreement with the annotated sequences of AF.
has become a popular source of data for studying AF and gained a lot of interest in the research community, and its public availability. The CinC2017 dataset has already been used by many researchers.

As mentioned in the introduction to this study, due to computational constraints, the subset of 15 features was selected by the Recursive Feature Elimination method. This reduction in the number of features also improved the complexity of a model. The subsets of 15 features were less prone to classify the segments with Normal rhythm into the Other class.

Due to a large overlap between the Normal and Other classes, we also considered the modified problem formulation where the Normal and Other classes were merged in a single class, thus, forming a classification problem with three classes. Also for case, we observed that both models trained with all features and the ones trained with the RFE selected features had qualitatively comparable performance on the MIT-BIH datasets (cf Table 7).

### 4.1. The CinC2017 dataset

As mentioned in the introduction to this study, due to its public availability, the CinC2017 dataset has already gained a lot of interest in the research community and has become a popular source of data for studying AF detection, e.g., (Hannun et al. 2019, Athif et al. 2018, Smisek et al. 2018), and (Parvaneh et al. 2018).

Most of the works using the CinC2017 dataset have focused on maximizing the performance of different solutions such as classification accuracy and mean F1 score. In this study, we did not aim to develop any new type of classifier that possibly could have improved the classification of the CinC2017 dataset itself. Instead, we used the Random Forest classifier and focused on how the selection of features affected the cross-validation performance. In addition to the three previously suggested feature sets, we also combined all features into an aggregated set with 509 features. The aggregated feature set demonstrated the largest cross-validation accuracy and highest mean F1 score, even though the relative improvements compared to the best individual feature set (Zabibi et al.) were marginal—0.7 and 0.9 percent, respectively. It is, however, worth mentioning that besides the current main focus on achieving high classification performance, it would also be important to be able to interpret and analyze the results of a solution. A reduction of the number of features without degrading the performance is one way of achieving higher interpretability of the classifier. For example, in this study, it was possible to achieve 99.6 percent of the cross-validation accuracy of the aggregated feature set with only 15 features (cf 509 in the aggregated feature set).

| MIT-BIH-NSR | Predicted | Actual |
|-------------|-----------|--------|
| AF          | Normal    | Other  |
| AF          | 0.0 ± 0.0 | 0.0 ± 0.0 |
| Normal      | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| Other       | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| Noise       | 4.6 ± 1.8 | 1061.4 ± 10.9 |

| MIT-BIH-ARR | Predicted | Actual |
|-------------|-----------|--------|
| AF          | Normal    | Other  |
| AF          | 236.8 ± 3.8 | 0.8 ± 0.4 | 36.5 ± 3.7 | 0.9 ± 0.7 |
| Normal      | 16.4 ± 2.3  | 1082.9 ± 12.3 | 654.5 ± 12.6 | 0.2 ± 0.4 |
| Other       | 49.0 ± 2.6  | 138.2 ± 19.0 | 497.0 ± 17.8 | 3.8 ± 1.3 |
| Noise       | 38.6 ± 3.0  | 59.5 ± 3.7 | 61.7 ± 3.8 | 3.2 ± 1.5 |

| MIT-BIH-AF  | Predicted | Actual |
|-------------|-----------|--------|
| AF          | Normal    | Other  |
| AF          | 10 358.1 ± 103.7 | 100.1 ± 8.0 | 941.9 ± 68.1 | 258.9 ± 70.6 |
| Normal      | 236.5 ± 34.7 | 10 427.2 ± 131.2 | 5 669.1 ± 139.4 | 96.2 ± 8.7 |
| Other       | 1.4 ± 1.1   | 2.4 ± 0.8 | 12.1 ± 1.2 | 0.1 ± 0.3 |
| Noise       | 0.0 ± 0.0   | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |

| Confusion matrices (four classes) of the test datasets for 15 RFE features from Datta feature set. Averaged over 10 independent runs. | |
4.2. Selection of feature variables

One important aim when reducing the number of calculated features is to preserve the classification performance. There is, however, another important reason for decreasing the complexity of the detection algorithms. This is when both the recording and calculations are performed using local processing on the same resource-constrained device. In more general terms, this process is known as edge computing (as opposed to cloud computing).

In this study, the experimental part of feature variables selection concerned the complexity. The observed results were consistent with our preliminary results reported in (Abdukalikova et al 2018). In particular, the cross-validation results on the CinC2017 dataset demonstrated that the RFE method showed the best results among the three considered feature selection methods. Moreover, the classification performance of the subsets of 15 features chosen by the RFE method was very close to the original feature sets (the largest reduction in accuracy was 1.3 percent), while the use of the subsets of 5 features resulted in a significant performance degradation, e.g., 13 percent decrease of $F_1$ score for the NCA method relative to the results on the full aggregated feature set. Moreover, in the case of three classes 15 features selected by the RFE performed slightly better than the original feature sets (the least improvement in $F_1$ score was 2.0 percent, Datta feature set). These results indicate the principal possibility of making simple detectors showing results on a par with large models. However, we have not focused on details of calculation time of individual features (some are more complex to calculate). Therefore, an interesting direction for future work would be to include the computational aspects for each feature such as involved preprocessing steps, transformations, and its complexity. Nevertheless, a simple way to assess an immediate effect of the reduced number of features would be the time it takes to train and evaluate a model before and after feature selection. For example, the average time per one independent run for all features in the aggregated feature set on the CinC2017 dataset (Table 2) was 32.6 s while for 15 features it was only 6.0 s thus, the gain was more than five times. Intuitively, for individual feature sets the gains were smaller as they have less features.

Other related works studying feature selection are (Sadr et al 2018) and (Christov et al 2018). In (Sadr et al 2018), the authors proposed a low-complexity approach using the features extracted in several domains but from RR intervals only. While the solution was positioned as a low-complexity one, it is worth noting that the number of used features was 119. A single hidden layer neural network was used as a classifier and the study reported a cross-validation mean $F_1$ score of 0.76. We used a different classifier and obtained somewhat higher mean $F_1$ score of 0.79 for 15 RFE features selected from the aggregated feature set. Similar to our study, the authors did not perform any detailed analysis of complexity, such as studying the time required for calculating each individual feature.

Another very relevant work that focused on assessing the most important features of a particular solution was the study by (Christov et al 2018), which also participated in the challenge. The solution was based on 44 features and a Linear Discriminant Analysis was used as a classifier. The best performing features were selected using a forward stepwise selection procedure, and 15 features were selected either from HRV metrics or from beat morphological metrics. Similar to the results reported here, they demonstrated that it was possible to get nearly the same performance with the subset of best performing features as achieved with all features. Besides mentioning that the solution’s running time was low when related to the challenge’s server quota, there was no other scrutinization of the complexity. Nevertheless, the authors suggested that a low-complexity solution would have to combine HRV and morphological features.

4.3. AF detection in the MIT-BIH datasets

The considered MIT-BIH datasets are commonly used in studies devoted to AF detection. For example, all three datasets were used to evaluate an approach for AF detection using Shannon entropy and symbolic dynamics as predictive feature variables (Zhou et al 2015), where the authors reported an accuracy above 0.9. When placing these results in the context of findings reported in this study, one important difference is that we classified recordings at the level of 30s segments, while in (Zhou et al 2015) individual beats were classified. Nevertheless, similar to (Zhou et al 2015) we observed no or very few AF false positives (see, e.g., Table 8) in the recordings in the MIT-BIH-NSR dataset (but frequent Other false positives in the case of four classes). There were also fewer AF false positives in the case of three classes, though the AF false positives in the case of four classes were concentrated to a few recordings (e.g., the sixth recording for Datta feature set). For the other two datasets with AF (MIT-BIH-AF and MIT-BIH-ARR), and in the case of three classes and 15 RFE features, the Datta feature set demonstrated the lowest average mismatch score per recording, which was approximately 0.05 and 0.02, respectively. It has also demonstrated the highest accuracy (cf Table 8) amongst the considered feature sets. In the case of four classes, this feature set was also the best one for the MIT-BIH-AF dataset with 0.04 average mismatch score, while for the MIT-BIH-ARR dataset it was the second best (0.03) after the aggregated feature set (0.02). Thus, these results allow us concluding, that from the AF detection point of view: a) the Datta feature set was the most promising one; b) there was only a small difference between the four and three classes cases.
4.4. Data labeling
Another aspect, which should be studied in greater detail is incorrect labeling of ECG segments. This is possible since most of the recordings in the CinC2017 dataset were labeled using a voting procedure between different algorithms, where a manual inspection and labeling only was performed if there was a disagreement between the algorithms (Clifford et al 2017). It has been shown in (Zhu et al 2014) that the overall classification performance can be improved, and even outperform manually labeling, by combining predictions made by different models via a voting procedure. However, remaining errors in the labeling of data could be one reason why the top $F_1$ scores in the CinC Challenge was not higher than 0.83 on the testing data.

Another potential issue is the labeling of the Other class. Although we used the best solutions to the challenge with many features, a large overlap between Normal and Other classes was observed (see, e.g., Table 3), which clearly indicates that it is difficult to separate a subset of recordings in these two classes. A similar observation was made in (Christov et al 2018) where the authors considered three different HR ranges: Bradycardia (HR $<$ 50bpm), Normal HR (HR = 50–100 bpm), and Tachycardia (HR $>$ 100 bpm). Interestingly, recordings included in Other class dominated in both the Bradycardia (86 % of total) and Tachycardia (53 % of total) ranges. These findings support that a relabeling of the recordings included in Other class could be motivated, e.g., by introducing separate classes for the recordings with frequent extra systolic beats (i.e., arrhythmia) and for the recordings with low beat-to-beat variability in heart rate (which could indicate autonomic dysfunction).

Finally, errors in the data or feature variables (e.g., due to incorrect detection of beats) could also affect the quality and reliability of a solution. We have not focused on this aspect in the current study, but this is something that should be investigated in future studies.

In this study, we focused on the detection of AF. For this purpose we mainly relied on the annotations of changes in rhythm and on the characteristics of MIT-BIH datasets. This includes the onset and end of AF periods, and the fact that no or very few non-sinus beats were present in the MIT-BIH-NSR dataset. Therefore, we are confident that the labeling of segments with AF is a valid representation of the occurrence of AF in the recordings. On the other hand, we put less focus on the definition of segments as “Other rhythm”. In the CinC2017 data, “Other rhythm” was defined as all non-AF abnormal rhythms, including low HRV. In the MIT-BIH datasets, segments with low HRV became annotated as “Normal” rhythm. Moreover, the definition of the Other class was a subjective task, where we had to specify the “necessary” proportion of a non-sinus rhythm that had to be present in a 30 s segment before it was labelled as Other rhythm. This could be done for the MIT-BIH-ARR and MIT-BIH-NSR datasets but not for recordings the MIT-BIH-AF dataset, where only annotations of normal rhythm and AF but no annotations of beat types were available. Thus, the classification performance automatically was poor for detecting the Other class in the MIT-BIH-AF dataset.

4.5. Combination of classes
In this study we only considered models where the classification has been done simultaneously. An alternative approach would be to perform cascaded classification. One of the best solutions to the challenge (Datta et al 2017) works precisely in such a way. For example, (Datta et al 2017) merged AF and Noise as well as Normal and Other and performed a binary classification as the first step. Note that our case with three classes (Normal and Other being combined) could be seen as a modification of the above scheme, but where AF and Noise classes are not merged and, thus, classification is ternary. We expected that such a modified problem formulation would significantly improve the cross-validation accuracy on the CinC2017 dataset, which was confirmed by the results in Table 5, however, mean $F_1$ scores improvements were less noticeable. In the case of the MIT-BIH datasets, the usage of three classes simplified the interpretation of the predictions, especially, for the MIT-BIH-AF and MIT-BIH-NSR datasets. It is also worth mentioning that we have performed the experiments with a cascaded model similar to (Datta et al 2017) for the case of three classes, where the model was first making a binary classification Noise versus All. At the next stage, all records classified as All were sent to another classifier separating AF class from Normal/Other class. The results were very similar to the case of a model trained to choose between three classes, therefore, we have not reported the cascaded model here.

4.6. Classifiers
This study reported the results obtained with the Random Forest classifier. It is worth mentioning that at least three out of top-ten solutions to the challenge have used it as well as solutions outside of the top-ten, e.g., (Kropf et al 2017). Moreover, most of the solutions have used ensemble classifiers such as XGBoost and AdaBoost. At the same time, when solution’s computational complexity is considered, ensemble classifiers are not optimal since they have to build and store several classification models. Therefore, in the future work it is worth exploring the details of the trade-offs between the AF detection performance and computational complexity, which could be achieved when using different classification techniques.
5. Conclusion

The CinC2017 dataset and the features extracted by the investigated solutions from the challenge showed promising results for detecting atrial fibrillation also in long-term ECG recordings. This confirms that the proposed features are generic for building atrial fibrillation detectors. The use of feature selection methods preserves the classification performance. This indicates the possibility of building low-complexity classifiers, which can be used on a resource-constrained device recording the ECG. However, it remains to investigate of how to accurately detect other abnormal heart rhythms based on the data and results from the 2017 Physionet/CinC Challenge.

ORCID iDs

Denis Kleyko @ https://orcid.org/0000-0002-6032-6155
Evgeny Osipov @ https://orcid.org/0000-0003-0069-640X
Urban Wiklund @ https://orcid.org/0000-0002-1313-0934

References

Abdulaliukova A, Kleyko D, Osipov E and Wiklund U 2018 Detection of atrial fibrillation from short ECGs: minimalistic complexity analysis for feature-based classifiers Computing in Cardiology (CinC) 45 1–4
Andreetti F, Carr O, Pimentel M A, Mahdi A and De Vos M 2017 Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG Computing in Cardiology (CinC) 44 1–4
Athif M, Yasawadene P C and Daluwatte C 2018 Detecting atrial fibrillation from short single lead ECGs using statistical and morphological features Physiol. Meas. 39 1–10
Christov I, Krasteva V, Simova I, Neycheva T and Schmid R 2018 Ranking of the most reliable beat morphology and heart rate variability features for the detection of atrial fibrillation in short single lead ECGs using statistical and morphological features Physiol. Meas. 39 1–10
Clifford G D, Liu C, Moody B, Lehman L-W H, Silva I, Li Q, Johnson A and Mark R G 2017 AF classification from a short single lead ECG recording: the hysioNet/Computing in cardiology challenge 2017 Computing in Cardiology (CinC) 44 1–4
Datta S, Puri C, Mukherjee A, Banerjee R, Choudhury A D, Singh R, Ukal A, Bandyopadhyay S, Pal A and Khandelwal S 2017 Identifying normal, AF and other abnormal ECG rhythms using a cascaded binary classifier Computing in Cardiology (CinC) 44 1–4
Fernandez-Delgado M, Cernadas E, Barro S and Amorim D 2014 Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15 3133–81 (http://jmlr.org/papers/v15/delgado14a.html)
Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng C and Stanley H 2000 PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals Circulation 101 e215–20
Guyon I, Weston J, Barnhill S and Vapnik V 2002 Gene selection for cancer classification using support vector machines Mach. Learn. 46 389–422
Hannun A, Rajpurkar P, Haghpanah M, Tison G, Bourn C, Turakhia M and Ng A 2019 Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network Nat. Med. 25 65–9
Kropf M, Hayn D and Schreier G 2017 ECG classification based on time and frequency domain features using random forests Computing in Cardiology (CinC) 44 1–4
Lloyd-Jones D M et al 2004 Lifetime risk for development of atrial fibrillation: the framingham heart study Circulation 110 1042–6
Mohebbi M and Ghassemian H 2008 Detection of atrial fibrillation episodes using SVM 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 177–80
Moody G and Mark R 1983 A new method for detecting atrial fibrillation using r-r intervals Computers in Cardiology 10 227–30
Moody G and Mark R 2001 The impact of the mit-bih arrhythmia database IEEE Eng. Med. Biol. Mag. 20 45–50
Park J, Lee S and Jeon M 2009 Atrial fibrillation detection by heart rate variability in poincare plot BioMedical Engineering OnLine 8 1–12
Parvaneh S, Rubin J, Rahman A, Conroy B and Babaeizadeh S 2018 Analyzing single-lead short ECG recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation Physiol. Meas. 39 1–8
Russell S and Norvig P 2010 Artificial Intelligence: A Modern Approach (London, England: Pearson Education, Ltd.)
Sadr N, Jayawardhana M, Pham T, Tang R, Balaei A T and de Chazal P 2018 A low-complexity algorithm for detection of atrial fibrillation using an ECG Physiol. Meas. 39 1–8
Smisek R, Hejc I, Ronzhina M, Nemcova A, Marosanova L, Kolarova J, Smital L and Vitek M 2018 Multi-stage SVM approach for cardiac arrhythmias detection in short single-lead ECG recorded by a wearable device Physiol. Meas. 39 1–14
Yang W, Wang K and Zuo W 2012 Neighborhood component feature selection for high-dimensional data Journal of Computers 7 161–8
Zabihie M, Rad A B, Kasaggelos A K, Kiranayz S, Narkilatis S and Gabbouj M 2017 Detection of atrial fibrillation in ECG handheld devices using a random forest classifier Computing in Cardiology (CinC) 44 1–4
Zhou X, Ding H, Wu W and Zhang Y 2015 A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate PLoS One 10 1–16
Zhu T, Johnson A, Behar J and Clifford G 2014 Crowd-sourced annotation of ECG signals using contextual information Ann. Biomed. Eng. 42 871–84