Selecting an Optimal Training Dataset for Machine Learning based Atrial Fibrillation Detection

Dimitri Dojchinovski[1][0000−0002−8925−284X] and Marjan Gusev[2][0000−0003−0351−9783]

[1] Innovation Dooel, 1000 Skopje, North Macedonia
dimitri.dojcinovski@innovation.com.mk
[2] Sts Cyril and Methodius University, 1000 Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

Abstract. The application of Machine Learning, in recent times, has excelled with positive outcome in many fields, including the medical field, such as handling cardiovascular problems. In this paper, we aim at developing a machine learning algorithm for detecting Atrial Fibrillation, as one of the most common and mortal types of heart rhythm problems - arrhythmias. Especially we address the research question of which dataset to be used in the learning process to reveal optimal results. The experiments are conducted using the following algorithms: Support Vector Machines, Decision Trees and Random Forest training validating and testing on specific selection of the three most popular publicly available electrocardiogram databases that contain episodes of Atrial Fibrillation. The research concluded that the best results are obtained by the Random Forest algorithm trained on LTAFDB selected by the 80-10-10 rule for training, validation and testing.

Keywords: Atrial fibrillation · Machine learning · ECG · Physionet

1 Introduction

The most common cardiovascular rhythm characterized with the highest mortal rates is the irregular arrhythmia Atrial Fibrillation (AFib), developing to heart attack [17]. The risk of developing AFib is 23.8% for males and 22.2% for females for patients in age between 55 and 75 years [9]. Recently, a lot of research was conducted using Machine Learning (ML) methods and revealed promising results analyzing an electrocardiogram (ECG) [4], [5], [15], [1], [11].

The main goal in this research is choosing the right ECG database and training method to develop an algorithm for detecting and classifying AFib episodes in long-term ECG recordings based on binary classification by observing the anomalies in the RR intervals.

Physionet contains at least three publicly available ECG databases [7] with AFib episodes, including Long Term AF (LTAFDB) [16], MIT-BIH Arrhythmia (AFDB) [12], MIT-BIH Atrial Fibrillation (MITDB) [13].

Recently, we have reported a method to choose the best segment labeling method [6] to be the majority method compared to the pure method for segment labeling.

The experiments in this research are based on implementation of the following ML algorithms Support Vector Machines (SVM) [2], Decision Tree (DT) [3], Random Forest (RF) [10].

The research problem addressed in this paper is selecting the database which will reveal the best results when the algorithm is being tested with the other databases. In addition, we address the problem of some irregular heart rhythms which follow a regular pattern, such as bigeminy, trigeminy, etc., while the AFib represents a heart rhythm which is irregular and occurs without any predefined pattern.

Therefore, the problem of selecting a proper training dataset is not so trivial. Many researchers used the same dataset for training, validation and testing, which does not cover other types of arrhythmia, and, unfortunately, can not be confirmed in real-time situations.

This was the main motivation in this paper, selection of a proper testing dataset that covers different real-time type of arrhythmia, and then selecting a training dataset. We have selected the MITDB as a testing dataset, since it covers a lot different types of arrhythmia, and selection of a training dataset was not coached by better results on a similar database, which mostly contains AFib episodes.

To evaluate the test results of developed algorithms and different experiments on different ECG databases and training methods, we use the following performance metrics Sensitivity (SEN), Specificity (SPC), Positive Predictive Value (PPV) and F1 - score. The Duration Method [8] was used instead of methods based on beats or number of episodes. In addition, to compare different approaches, we use an Improvement Factor (IF).

Section 3 presents the basic knowledge in the field of cardiovascular diseases. In Section 4 we define the feature engineering. Section 2 describes the methods and ECG databases used in the research, and Section 5 the obtained results, discussed in Section 6. Finally, the conclusion and future directions are presented in Section 7.

2 Training Datasets

Publicly available ECG database from Physionet [7] have been implemented for decades in the medical field, and research for AFib includes the LTAFDB, AFDB and MITDB databases. Each of the bases contains a different number of ECG recordings of different patients recorded with modern Holter apparatus and analyzed and labeled with annotations by certified and qualified medical professionals.

An annotation file is accompanied for each ECG record in the database with annotations for each registered heart beat: timestamp, sample id, beat type,

Table 1. Training datasets Physionet ECG databases

ECG DB	records	avg. length	# samples	sampling	episodes
LTAFDB	83	24h	8.903.169	128 Hz	9
AFDB	25	10h	1.221.532	250 Hz	4
MITDB	48	30min	109.483	360 Hz	11

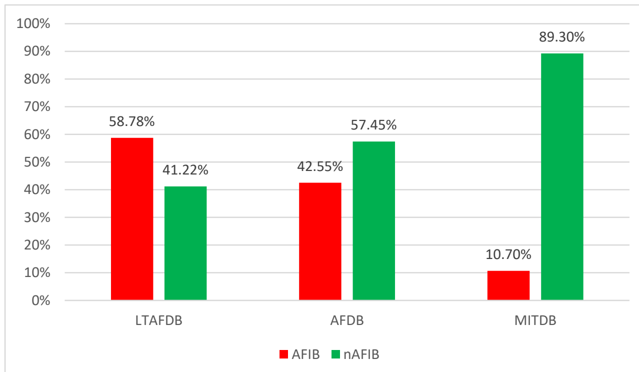

Fig. 1. Overview of class ratio in each ECG databases

rhythm annotation, and other features. The timestamp of each heart beat in the recordings is relative to the beginning of the ECG recording.

19 beat annotations and 22 non-beat annotations are recognized according to the description of Physionet [7]. The type of heart rate will play a role in determining the features needed to train the algorithms.

Since the goal is to develop a binary classifier, the rhythm annotations are divided into a positive class AFib and a negative class nAFib where all other rhythm annotations are, including normal and other abnormal heart rhythm annotations.

LTAFDB includes 83 recordings from 83 individual patients each 24 hours long with a total of 8.903.169 annotated beats sampled at 128 Hz with the following nine rhythm annotations: *(AB*, *(AFib*, *(B*, *(IVR*, *(N*, *(SBR*, *(SVTA*, *(T* and *(VT*.

AFDB includes 25 recordings from 25 individual patients each 10 hours long with a total of 1.221.532 annotated beats sampled at 250 Hz. These beat samples are annotated by four rhythm annotations: *(AFib*, *(AFL*, *(NOD* and *(N*.

MITDB includes 48 recordings from 48 individual patients each half hours long with a total of 109.483 annotated beats sampled at 360 Hz. All this beat samples are annotated by eleven rhythm annotations: *(AFib*, *(AFL*, *(B*, *(IVR*, *(N*, *(NOD*, *(P*, *(SBR*, *(SVTA*, *(T* and *(VT*.

Various rhythm episodes may influence the quality of the AFib classifier, so the selection of a proper training dataset is very important to avoid problems in real-time situations.

3 Feature Engineering

Solving a problem related with cardiovascular problems of any kind requires introduction to the functionality of the heart and how is represented for more descriptive analysis.

Heartbeats are part of an electrical system (Electrical conduction heart system) that regulates the contraction of heart muscles, where they differ from one another. A heartbeat can be recognized from a standard ECG contains by identification of a P-wave (a wave with small amplitude), QRS complex (the most exposed wave succeeding the P-wave) and T-wave (the wave succeeding the QRS complex, with higher amplitude than P-wave but not exposed as the QRS complex).

R-peak is the highest amplitude and the central point of the QRS complex. The rhythm in the ECG is determined by the occurrence of R-peaks, and RR distances between neighboring R-peaks. Irregularity is then detected by the analysis of the RR sequence.

Rhythm is determined by the patterns of the RR segments and can be recognized as normal sinus rhythm (NSR), or one of the following arrhythmia:

- Atrial bigeminy (AB)
- Atrial fibrillation (AFib)
- Atrial flutter (AFL)
- Ventricular bigeminy (B)
- 2° heart block (BII)
- Idioventricular rhythm (IVR)
- Nodal (A-V junctional) rhythm (NOD)
- Paced rhythm (P)
- Pre-excitation (WPW) / (PREX)
- Sinus bradycardia (SBR)
- Supraventricular tachyarrhythmia (SVTA)
- Ventricular trigeminy (T)
- Ventricular flutter (VFL)
- Ventricular tachycardia (VT)

AFib is usually diagnosed by the absence of P-waves and irregular rhythm that does not follow any particular pattern.

The method of feature extracting in our previous research [6] starts with calculation of RR segments and continues with labeling of the rhythm episode (arrhythmia). Segments are labeled according to the majority of the assigned arrhythmia for the majority-based rule, on contrary of the pure-method which uses only segments with single arrhythmia. The sliding window technique is used to generate all the segments of consecutive RR segments to be used in the training, validation and testing phase.

The number of successive RR intervals in the generated segments with the sliding window are in the odd number range of $[5, 7, \ldots, 49]$ in order to apply the majority rule.

4 Experimental Setup

4.1 Evaluation Metrics

Model performance is calculated using metrics: sensitivity, specificity, positive predictivity value, F1-score and improvement factor. All of these are defined using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Sensitivity (SEN) measures the proportion of correctly identified sequences of AFib in regards to the all detected positives and is calculated by (1).

$$SEN = \frac{TP}{TP + FN} \tag{1}$$

Positive Predictivity Value (PPV) or simply a precision - measures the proportion of sequences correctly classified as AFib in regards to the real number of positive results and is calculated by (2).

$$PPV = \frac{TN}{TN + FP} \tag{2}$$

The F1 score is calculated by (3) as a harmonic mean of the precision and sensitivity.

$$F1 = \frac{2 * PPV * SEN}{PPV + SEN} \tag{3}$$

All of the above mentioned methods are calculated following the duration method (Gusev et al. [8]), where the evaluation is performed by comparing the lengths and how much the episodes match the original data with that provided by the model. In a way, this method standardizes the measuring of the performances of the models when dealing with this kind of problem.

An Improvement Factor IF *(Improvement Factor)* of the algorithm A versus B compares the correpsonding F1-score values and is calculated by (4). If the ratio is positive then the result is in favor of A.

$$IF = \frac{F1_A - F1_B}{F1_B} \tag{4}$$

4.2 Training Procedure

After extracting the features, the data of each ECG database is prepared for research using the model by dividing into a train, validation and test dataset. The training dataset is used exclusively to train the machine learning algorithms. A validation set checks the performance of training algorithms to see if there is room for improvement and to serve as an indicator when optimizing algorithms. The test dataset shows the final performance of the algorithms after being optimized, such a dataset is hidden from the algorithms to the end to see how they respond to new unseen samples.

In this experiment the data was split by a ratio of 80/10/10, where 80% of the data is for training, 10% for validation and 10% for testing. This split ratio is based on the Pareto Principle (80/20 rule) [14], where in most cases 80% of effects come from 20% of causes, it statistically does come close to explaining many human, machine, and environmental phenomena. For this reason, and typically without knowledge of the source, many use the 80/20 split for training and testing.

Since each of the data in the ECG databases represents individual patients, actually dividing them with such a ratio does not amount to merging all the patients and randomly dividing them. Since the ratios of classes in each data set are different, the data are grouped in such a way that the ratios of classes in the training, validation and testing dataset are almost the same as the ratios of classes in the respective ECG database.

In this paper, we decided to use one dataset for training, validation, and testing using the 80/10/10 rule, but in addition to test the algorithm also with the other two datasets.

We use notation of L for LTAFDB, A for AFDB and M for MITDB. The conducted experiments include all combinations of tested all combinations (Training-Validation-Testing) for the ECG datasets, such as, (L-L-L) meaning the LTAFDB is used with 80/10/10 rule for the LTAFDB, etc.

5 Results

The experiment is performed in order to select the adequate ECG database. Due to the smaller number of AFib episodes in the MITDB database, we used only LTAFDB and AFDB for training with RR intervals as features, with both labeling methods majority and pure trained individually on the three proposed machine learning algorithms. The testing was conducted on the MITDB ECG database. Table 5 presents the values of the improvement factor of the validation process for models trained and validate on LTAFDB (L-L-L) and trained and validate on AFDB (A-A-A), and Table 6 represents the values of the improvement factor of the testing process on MITDB (L-L-M and A-A-M).

As we can see AFDB has a advantage above LTAFDB in this state of development. This is due to the fact that the AFDB (Fig. 3) database has relatively few different types of arrhythmias than LTAFDB (Fig. 2), where the differences in arrhythmias occur.

Therefore, we started testing the model with the other database as a more reliable result. It is noted that when testing with other databases, all the results are negative, meaning that the LTAFDB was generally generalized, that is, when AFib was detected in data outside the training dataset it went as a better ECG database. In the following experiments, the LTAFDB serves as a base for finding the optimal detection and classification specifications for AFib, more in Section 6.

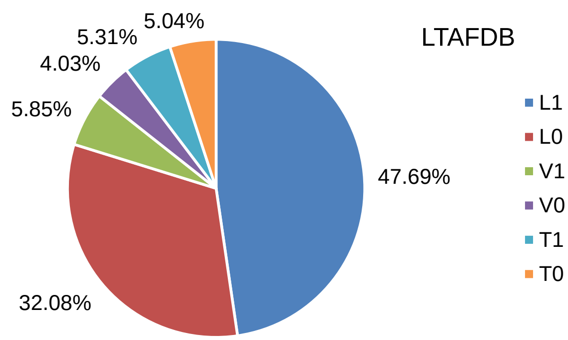

Fig. 2. LTAFDB dataset distribution, L stands for Training, V for Validation, T for Testing, 1 for class 1 (AFib), and 0 for class 0, (non-AFib)

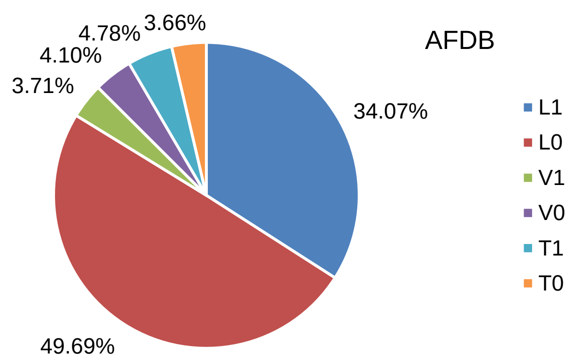

Fig. 3. AFDB dataset distribution, L stands for Training, V for Validation, T for Testing, 1 for class 1 (AFib), and 0 for class 0, (non-AFib)

6 Discussion

The results of experiment determine which ECG training database yields better results. Figure 6 shows the results of testing the trained models on the MITDB, according to the average of the improvement factor of the trained models, AFDB shows better validation results (Fig. 5), but worse testing because of the characteristics of the databases themselves, as presented in (Fig. 3), AFDB has few arrhythmias which are different from AFib and therefore not trained with examples that have characteristics similar to the ADIB with irregular rhythm and which can lead to erroneous conclusions. More precisely, LTAFDB is better for training, encompassing more types of arrhythmias, and on the other hand the number of segments used for training is larger (Fig. 1).

Table 6 shows the improvement factor values of all developed ML algorithms on LTAFDB and AFDB tested on MITDB. The average values of these results lead to a conclusion that the improvement factor is negative for all algorithms

	ltafdb						afdb					
	majority			pure			majority			pure		
Segment length	SVM	DT	RF	SVM	DT	RF	SVM	DT	RF	SVM	DT	RF
5	63.66%	84.94%	85.28%	69.04%	85.02%	85.32%	5.73%	91.75%	91.99%	78.79%	91.67%	91.99%
7	65.71%	86.11%	86.17%	49.65%	85.96%	86.22%	12.38%	92.75%	92.92%	48.12%	92.76%	92.86%
9	69.60%	86.50%	86.56%	69.57%	86.36%	86.40%	77.94%	93.33%	93.29%	9.20%	93.09%	93.27%
11	61.43%	86.84%	86.74%	65.59%	86.43%	86.72%	18.58%	93.53%	93.45%	4.34%	93.47%	93.33%
13	69.08%	87.04%	86.89%	60.36%	86.47%	86.90%	60.07%	93.58%	93.82%	60.12%	93.69%	93.74%
15	70.64%	87.22%	87.30%	64.13%	86.54%	87.14%	19.54%	93.71%	93.55%	5.84%	93.53%	93.73%
17	62.00%	87.19%	87.53%	60.06%	86.69%	87.54%	27.01%	93.87%	93.72%	78.11%	93.64%	93.71%
19	43.71%	87.08%	87.87%	68.27%	86.85%	87.76%	34.12%	93.89%	93.67%	79.74%	93.66%	93.63%
21	64.63%	87.37%	87.94%	65.53%	86.94%	88.05%	39.22%	93.78%	93.49%	82.42%	93.81%	93.63%
23	61.45%	87.11%	88.26%	64.79%	86.86%	87.78%	65.76%	93.57%	93.36%	63.29%	93.60%	93.51%
25	53.25%	87.47%	88.24%	64.81%	86.87%	88.16%	59.83%	93.49%	93.43%	25.55%	93.73%	93.53%
27	50.64%	87.69%	88.50%	56.87%	86.85%	88.11%	70.09%	93.56%	93.30%	36.02%	93.84%	93.50%
29	66.74%	87.72%	88.43%	57.72%	87.01%	88.05%	57.81%	93.46%	92.70%	46.48%	93.72%	93.06%
31	44.65%	87.71%	88.55%	65.49%	87.16%	88.37%	73.16%	93.36%	92.61%	34.66%	93.34%	92.88%
33	61.66%	87.82%	88.58%	65.68%	87.21%	88.44%	44.55%	93.23%	92.31%	40.17%	93.38%	92.58%
35	65.82%	87.82%	88.86%	65.56%	87.40%	88.57%	42.11%	93.33%	92.04%	77.00%	93.30%	92.34%
37	58.65%	87.95%	88.82%	65.40%	87.36%	88.75%	23.32%	93.20%	91.88%	30.20%	93.08%	92.29%
39	69.54%	88.00%	88.79%	65.55%	87.44%	88.58%	41.93%	93.13%	91.89%	36.48%	93.14%	92.09%
41	65.32%	88.42%	89.05%	65.37%	87.53%	88.68%	74.19%	93.28%	91.57%	56.59%	93.30%	92.06%
43	65.50%	88.49%	89.10%	65.40%	87.52%	88.99%	35.56%	92.69%	91.48%	42.58%	93.15%	91.96%
45	64.19%	88.30%	89.02%	65.32%	87.58%	88.86%	36.98%	93.11%	91.58%	46.66%	93.15%	91.59%
47	65.75%	88.31%	89.13%	65.25%	87.66%	88.82%	72.23%	93.02%	91.33%	45.02%	93.25%	91.50%
49	65.62%	88.40%	89.10%	65.27%	87.84%	88.90%	51.03%	92.95%	91.52%	41.01%	93.22%	91.39%

Fig. 4. F1 score results with majority and pure methods

	E1.1 Improvement factor - validation						E1.2 Improvement factor - testing					
	majority			pure			majority			pure		
Segment length	SVM	DT	RF	SVM	DT	RF	SVM	DT	RF	SVM	DT	RF
5	-90.99%	8.02%	7.87%	14.13%	7.82%	7.82%	-72.29%	1.65%	1.72%	57.70%	1.60%	1.56%
7	-81.16%	7.72%	7.84%	-3.09%	7.91%	7.71%	-68.12%	0.99%	3.11%	4.84%	-2.58%	3.76%
9	11.97%	7.90%	7.77%	-86.78%	7.79%	7.94%	36.13%	-2.04%	0.12%	-59.54%	-1.43%	0.07%
11	-69.76%	7.70%	7.74%	-93.38%	8.14%	7.63%	-60.93%	-2.53%	-0.44%	-77.93%	-6.08%	0.42%
13	-13.04%	7.51%	7.98%	-0.40%	8.34%	7.87%	-6.83%	-6.42%	-2.25%	-35.73%	-5.02%	-1.17%
15	-72.33%	7.43%	7.16%	-90.89%	8.08%	7.56%	-36.34%	-9.65%	-5.05%	-76.42%	-4.95%	-4.60%
17	-56.45%	7.66%	7.07%	30.06%	8.02%	7.04%	-48.59%	-8.70%	-6.29%	3.72%	-11.19%	-5.69%
19	-21.94%	7.82%	6.60%	16.80%	7.84%	6.69%	15.20%	-9.13%	-10.20%	31.26%	-9.41%	-8.02%
21	-39.31%	7.34%	6.32%	25.77%	7.90%	6.35%	-41.05%	-9.18%	-8.63%	3.49%	-10.90%	-11.65%
23	7.01%	7.43%	5.78%	-2.31%	7.76%	6.52%	0.36%	-12.70%	-11.16%	-6.67%	-10.18%	-11.61%
25	12.37%	6.89%	5.88%	-60.58%	7.90%	6.09%	23.32%	-14.02%	-13.25%	-52.09%	-13.62%	-13.48%
27	38.40%	6.70%	5.42%	-36.66%	8.04%	6.11%	61.07%	-13.13%	-13.21%	-40.08%	-16.30%	-14.85%
29	-13.38%	6.54%	4.83%	-19.47%	7.71%	5.69%	4.05%	-15.08%	-15.00%	-30.21%	-13.77%	-14.24%
31	63.86%	6.45%	4.59%	-47.07%	7.08%	5.11%	101.99%	-15.34%	-16.77%	-46.27%	-15.02%	-17.89%
33	-27.75%	6.16%	4.21%	-38.84%	7.07%	4.68%	-30.82%	-16.59%	-18.56%	-40.31%	-17.15%	-19.78%
35	-36.03%	6.27%	3.59%	17.44%	6.75%	4.25%	-34.71%	-19.75%	-20.89%	-6.25%	-17.58%	-20.51%
37	-60.23%	5.97%	3.45%	-53.81%	6.55%	3.99%	-53.06%	-18.38%	-22.77%	-49.22%	-18.57%	-20.48%
39	-39.71%	5.82%	3.49%	-44.36%	6.52%	3.97%	-13.69%	-17.94%	-22.17%	-44.94%	-19.98%	-22.34%
41	13.57%	5.50%	2.83%	-13.42%	6.59%	3.81%	-2.72%	-19.12%	-22.82%	-13.59%	-20.68%	-20.91%
43	-45.72%	4.75%	2.67%	-34.90%	6.43%	3.35%	-37.91%	-18.24%	-22.52%	-38.38%	-20.12%	-23.48%
45	-42.40%	5.45%	2.87%	-28.57%	6.36%	3.07%	-38.30%	-18.52%	-23.11%	-31.55%	-18.18%	-23.36%
47	9.86%	5.33%	2.46%	-31.00%	6.38%	3.02%	0.30%	-18.98%	-25.28%	-34.57%	-20.15%	-24.74%
49	-22.24%	5.15%	2.72%	-37.17%	6.13%	2.81%	-15.13%	-19.46%	-23.97%	-39.05%	-22.45%	-25.74%
Average	-25.02%	6.67%	5.27%	-26.89%	7.35%	5.61%	-13.83%	-12.27%	-13.02%	-27.03%	-12.77%	-12.99%

Fig. 5. F1 score results with majority and pure methods

and segment labeling methods, which means that LTAFDB is more suitable for this kind of problem.

Selection of a proper input database was analyzed with accent on combining two datasets and shuffling the segments into a dataset used for training, valida-

ratio AFDB/LTAFDB testing						
	majority			pure		
Segment length	SVM	DT	RF	SVM	DT	RF
5	-72.29%	1.65%	1.72%	57.70%	1.60%	1.56%
7	-68.12%	0.99%	3.11%	4.84%	-2.58%	3.76%
9	36.13%	-2.04%	0.12%	-59.54%	-1.43%	0.07%
11	-60.93%	-2.53%	-0.44%	-77.93%	-6.08%	0.42%
13	-6.83%	-6.42%	-2.25%	-35.73%	-5.02%	-1.17%
15	-36.34%	-9.65%	-5.05%	-76.42%	-4.95%	-4.60%
17	-48.59%	-8.70%	-6.29%	3.72%	-11.19%	-5.69%
19	15.20%	-9.13%	-10.20%	31.26%	-9.41%	-8.02%
21	-41.05%	-9.18%	-8.63%	3.49%	-10.90%	-11.65%
23	0.36%	-12.70%	-11.16%	-6.67%	-10.18%	-11.61%
25	23.32%	-14.02%	-13.25%	-52.09%	-13.62%	-13.48%
27	61.07%	-13.13%	-13.21%	-40.08%	-16.30%	-14.85%
29	4.05%	-15.08%	-15.00%	-30.21%	-13.77%	-14.24%
31	101.99%	-15.34%	-16.77%	-46.27%	-15.02%	-17.89%
33	-30.82%	-16.59%	-18.56%	-40.31%	-17.15%	-19.78%
35	-34.71%	-19.75%	-20.89%	-6.25%	-17.58%	-20.51%
37	-53.06%	-18.38%	-22.77%	-49.22%	-18.57%	-20.48%
39	-13.69%	-17.94%	-22.17%	-44.94%	-19.98%	-22.34%
41	-2.72%	-19.12%	-22.82%	-13.59%	-20.68%	-20.91%
43	-37.91%	-18.24%	-22.52%	-38.38%	-20.12%	-23.48%
45	-38.30%	-18.52%	-23.11%	-31.55%	-18.18%	-23.36%
47	0.30%	-18.98%	-25.28%	-34.57%	-20.15%	-24.74%
49	-15.13%	-19.46%	-23.97%	-39.05%	-22.45%	-25.74%
Average	-13.83%	-12.27%	-13.02%	-27.03%	-12.77%	-12.99%

Fig. 6. Improvement factor of testing LTAFDB and AFDB on MITDB

tion and verification. However, this approach showed that this approach provides higher performance scores than the models trained just on LTAFDB, but they showed poor performance when tested on data outside of the train domain. For example, if tested on another ECG database, such as AFDB, it shows poorer results.

7 Conclusion

In this paper, we experiment in determining the adequate ECG database for training the models used for detecting and AFib in ECG recordings. LTAFDB ECG database proves to be adequate for this problem.

Future work aims at developing an optimal model for ML-based AFib detection in ECG recording, and find the optimal ML algorithm, segment length, features sets, segments labeling method and adequate ECG database.

References

1. Babaeizadeh, S., Gregg, R.E., Helfenbein, E.D., Lindauer, J.M., Zhou, S.H.: Improvements in atrial fibrillation detection for real-time monitoring. Journal of electrocardiology **42**(6), 522–526 (2009)

2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152 (1992)
3. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984)
4. Colloca, R., Johnson, A.E., Mainardi, L., Clifford, G.D.: A support vector machine approach for reliable detection of atrial fibrillation events. In: Computing in Cardiology 2013. pp. 1047–1050. IEEE (2013)
5. Datta, S., Puri, C., Mukherjee, A., Banerjee, R., Choudhury, A.D., Singh, R., Ukil, A., Bandyopadhyay, S., Pal, A., Khandelwal, S.: Identifying normal, af and other abnormal ecg rhythms using a cascaded binary classifier. In: 2017 Computing in Cardiology (CinC). pp. 1–4. IEEE (2017)
6. Dojchinovski, D., Gusev, M.: Segment labeling method for ml-based afib detection. In: Proceedings of Conference of Informatics and Information technology (CiiT). No. 40, Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engneering (2020), https://repository.ukim.mk/handle/20.500.12188/8273
7. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Mietus, J., Moody, G., Peng, C., Stanley, H.: Components of a new research resource for complex physiologic signals. PhysioBank, PhysioToolkit, and Physionet (2000)
8. Gusev, M., Boshkovska, M.: Performance evaluation of atrial fibrillation detection. In: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). pp. 342–347. IEEE (2019)
9. Heeringa, J., van der Kuip, D.A., Hofman, A., Kors, J.A., van Herpen, G., Stricker, B.H.C., Stijnen, T., Lip, G.Y., Witteman, J.C.: Prevalence, incidence and lifetime risk of atrial fibrillation: the rotterdam study. European heart journal **27**(8), 949–953 (2006)
10. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
11. Hu, J., Zhao, W., Xu, Y., Jia, D., Yan, C., Wang, H., You, T.: A robust detection method of atrial fibrillation. In: 2018 Computing in Cardiology Conference (CinC). vol. 45, pp. 1–4. IEEE (2018)
12. Moody, G.: A new method for detecting atrial fibrillation using rr intervals. Computers in Cardiology pp. 227–230 (1983)
13. Moody, G.B., Mark, R.G.: The impact of the mit-bih arrhythmia database. IEEE Engineering in Medicine and Biology Magazine **20**(3), 45–50 (2001)
14. Newman, M.E.: Power laws, pareto distributions and zipf's law. Contemporary physics **46**(5), 323–351 (2005)
15. Peimankar, A., Puthusserypady, S.: Ensemble learning for detection of short episodes of atrial fibrillation. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 66–70. IEEE (2018)
16. Petrutiu, S., Sahakian, A.V., Swiryn, S.: Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace **9**(7), 466–470 (2007)
17. Wolf, P.A., Abbott, R.D., Kannel, W.B.: Atrial fibrillation as an independent risk factor for stroke: the framingham study. Stroke **22**(8), 983–988 (1991)