Segment Labeling Method for ML-based AFIB Detection

Dimitri Dojchinovski
Innovation Dooel

Skopje, North Macedonia
dimitri.dojchinovski@innovation.com.mk

Marjan Gusev
St. Cyril and Methodius University
Computer Science and Engineering
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

***Abstract*—Atrial Fibrillation is one of the most common and mortal types of heart rhythm problems - arrhythmias. Therefore, early and accurate detection is important in detecting heart diseases and prescribing appropriate treatment therapy. Developing a technology of this kind is of pivotal importance and a challenging problem for noninvasive tools for patient monitoring and analysis.**

Electrocardiography provides comprehensive information that can be efficiently used in the management of the patients heart health. Detecting and classifying episodes of the different types of heart diseases is a subject of continuous research and immediately with new technological advances. Machine learning methods emerged as frequently used technology recently and become acknowledged for their relevance and results in this field.

Developing an effective model for detecting and classifying Atrial Fibrillation in ECG recordings requires the right data and adequate feature engineering. For this purpose we propose two methods, majority and pure segment labeling method used in the performed segmentation for feature engineering using the most popular ECG database and by integrating them in three machine learning algorithms, Support Vector Machines, Decision Trees and Random Fores.

The research concluded that the majority method trained on the Random Forest algorithm gives the highest results in the defined research space.

***Index Terms*—Atrial fibrillation, Machine learning, ECG.**

I. Introduction

Detecting and classifying episodes of the different types of heart diseases is a subject of continuous research and immediately with new technological advances. Machine learning (ML) methods emerged as frequently used technology recently and become acknowledged for their relevance and results in this field.

Atrial Fibrillation (AFib) is one of the most common and mortal types of heart rhythm problems - arrhythmias. Therefore, early and accurate detection is important in detecting heart diseases and prescribing appropriate treatment therapy. Developing a technology of this kind is of pivotal importance and a challenging problem for noninvasive tools for patient monitoring and analysis.

The dangerous consequences from AFib, are described in countless research papers. Wolf et al. [1] concluded that AFib, by itself, has a huge contribution in developing a heart attack. Other cardiovascular abnormalities diminish with age, but AFib prevail in older humans. Heeringa et al. [2] concluded that patients in age between 55 and 75 years, the risk of developing AFib is 23.8% for males and 22.2% for females.

Corresponding author: Dimitri Dojchinovski
email: dimitri.dojcinovski@innovation.com.mk

The main goal in this research is choosing the right method for labeling the featured engineered segments from the LTAFDB ECG database used to develop a model for detecting and classifying AFib episodes in long term ECG recording based on binary classification by observing the anomalies in the RR intervals.

Three types of machine learning algorithms are implemented with in the search for the optimal model:

- Support Vector Machines - SVM [3],
- Decision Tree - DT [4],
- Random Forest - RF [5].

which are broadly used in the domain of detecting arrhythmias [6] [7] [8] [9].

The Long Term AF (LTAFDB) [10] ECG database from Physionet was used for training, validating and testing as one of the most used in many researches for decades in the field of cardiovascular diseases.

The analyzed ECG database features a unique set of ECG recordings upon which a set of feature engineering methods were applied:

- calculating RR intervals and labeling them,
- feature extraction by grouping subsequent RR intervals in segments with a certain length,
- labeling the generating segments.

When labeling the segments, two methods are proposed for assigning a rhythm annotation to the segment, majority and pure method. Due to the unique nature of each of the used ECG databases it is of great importance to pick the best method for the given problem.

The performance metrics of Sensitivity, Positive Predictive Value and F-score are used for evaluation of the algorithms alongside Improvement Factor (IF) and Duration Method [11].

Section II presents the basic background in the field of cardiovascular medicine required to understand the AFib diseases. Feature engineering process and segment labeling method are analyzed in Section III and the evaluation methodology in

Section IV. Section V presents the obtained results discussed in Section VI. Final remarks and future work are presented in Section VII.

II. Background

Electrocardiography (electrocardiogram – ECG) is a graphic method for measuring the electrical activity of the heart by tracing the electric current generated by the heart muscle during a heartbeat and it provides information of the current condition of the heart.

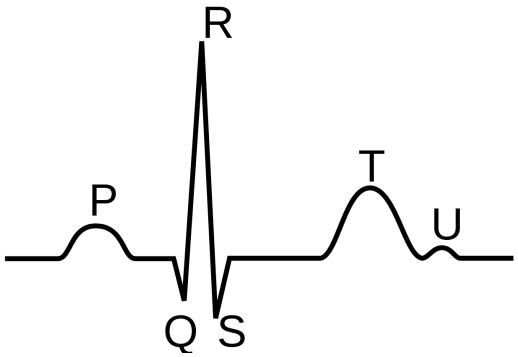
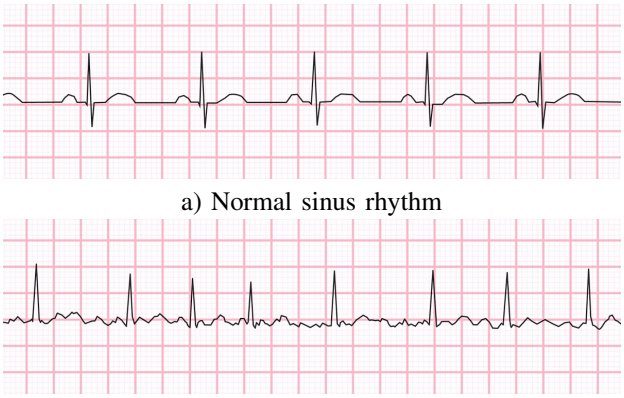


Fig. 1. QRS complex

The QRS complex is the most visible spike on the ECG line (Figure 1). Is made up of three main waves, which indicate the changing direction of the electrical impulse as it passes through the heart. The largest is the R wave (R peak), which represents the electrical impulse as it passes through the main portion of the ventricular walls. For detecting AFib there are two indices, the absence of non synchronized appearance of the P wave and irregularities in the heart rhythm. The P wave is very hard to detect, for that reason the more convenient method, irregularities in the heart rhythm, is beeing used. The time calculated by subtracting two consecutive R waves is labeled as RR interval. Irregularity in these intervals are considered as one of the most important indicators for AFib detection. Figure 2 pinpoints the difference between the intervals with irregularities in the heart rhythm and normal sinus rhythm.

Arrhythmia occurs as a variation from the normal heartbeat, usually as a result from irregularities within the specialized cardiac muscle cells that control the signals sent to the heart muscles (conduction system). There are several types of arrhythmias, but the most sever can appear in the form of AFib episode.

The main challenge in diagnosing heart disease using ECG is that the normal ECG may still differ for each person and sometimes one disease has dissimilar signs on different patient's ECG signals. Also, two distinct diseases may have approximately identical effects on normal ECG signals. These challenges complicate the heart disease diagnose. So adequate feature engineering is important for pattern recognition.



a) Normal sinus rhythm

b) AFib

Fig. 2. Analyzed ECG rhythm episodes

III. Feature engineering

In this section we describe how and with what features we use the ECG databases in the ML algorithms. The procedure is realized according to the following activities:

- RR intervals are calculated for the annotation records,
- the calculated RR intervals are labeled according to the type of arrhythmia of the second R beat for each RR interval,
- Using the sliding window technique we generate segments of multiple consecutive RR intervals, and label whether the corresponding interval belongs to the AFib rhythm episode.

The recordings in the ECG databases are transformed into data containing calculated RR intervals of each adjacent R beat, annotated with a rhythm annotation of the second of the two adjacent R heartbeats.

The labeled RR intervals are used to generated segments of successive RR intervals with the sliding window technique where each segment represents a sample as part of a research set used with machine learning algorithms.

Each of the segments belongs to one of two classes, AFib (positive class, denoted by 1) or nAFib (negative class denoted by 0). The negative class includes samples of other arrhythmias that appear in ECG data bases, and the positive class consists of samples with Atrial Fibrillation arrhythmia. In the formed segments, the sliding window can encompass samples of both classes, with the segment class depending on the number of samples of both classes. Two methods are introduced for this problem:

- labeling the segments with majority vote (*majority method*),
- labeling pure segments (*pure method*).

For example, if we are creating a segment with sliding window size 5,the created segment has 5 features: f1, f2, f3, f4, f5, where each feature represents the length of the RR intervals of the 5 consecutive samples covered by the sliding window.

Since the sliding window in each iteration contains a certain number of samples, the class of the newly added segment is

EKG samples						method M							method P					
Sample #	Type	Rhythm		RR		f1	f2	f3	f4	f5	class		f1	f2	f3	f4	f5	class
83664	N	(N		#		#	#	#	#	#	#		#	#	#	#	#	#
83774	N	(N		110		110	#	#	#	#	#		110	#	#	#	#	#
83885	N	(N		111		111	110	#	#	#	#		111	110	#	#	#	#
83993	N	(N		108		108	111	110	#	#	#		108	111	110	#	#	#
84103	N	(N		110		110	108	111	110	#	#		110	108	111	110	#	#
84214	N	(N		111		111	110	108	111	110	0		111	110	108	111	110	0
84324	N	(N		110		110	111	110	108	111	0		110	111	110	108	111	0
84396	N	(AFIB		72		72	110	111	110	108	0		72	110	111	110	108	#
84484	N	(AFIB		88		88	72	110	111	110	0		88	72	110	111	110	#
84548	N	(AFIB	→	64	→	64	88	72	110	111	1	v	64	88	72	110	111	#
84643	N	(AFIB		95		95	64	88	72	110	1		95	64	88	72	110	#
84783	N	(AFIB		140		140	95	64	88	72	1		140	95	64	88	72	1
84863	N	(AFIB		80		80	140	95	64	88	1		80	140	95	64	88	1
84974	N	(N		111		111	80	140	95	64	1		111	80	140	95	64	#
85082	N	(N		108		108	111	80	140	95	1		108	111	80	140	95	#
85192	N	(N		110		110	108	111	80	140	0		110	108	111	80	140	#
85303	N	(N		111		111	110	108	111	80	0		111	110	108	111	80	#
85413	N	(N		110		110	111	110	108	111	0		110	111	110	108	111	0
85522	N	(N		109		109	110	111	110	108	0		109	110	111	110	108	0
85630	N	(N		108		108	109	110	111	110	0		108	109	110	111	110	0
85740	N	(N		110		110	108	109	110	111	0		110	108	109	110	111	0
85848	N	(N		108		108	110	108	109	110	0		108	110	108	109	110	0
85958	N	(N		110		110	108	110	108	109	0		110	108	110	108	109	0

Fig. 3. Forming of segment with 5 segment length

determined by the class of the dominant rhythm annotation that most of the RR intervals that make the segment belong to.

Method P (pure segments) annotates the segment with one of the classes if all the samples that make up the segment belong to one of the classes. With this method, the segments that contain samples of both classes are not added to the dataset for research. This kind of segments every time the rhythm annotations change, which means they take a very small part of the total number of segments.

Method M (majority segments) annotates the segment with one of the classes depending of the dominant rhythm annotation that most of the RR intervals that make the segment belong to. From the example with five segment length (Figure 3), the segments that contain three or more samples of the same class, it labels the whole segment with that class.

The occurrence of this segments is the reason we use odd numbers in the segment lengths, so that we can easily determine the class of the segment with the majority method (Figure 3).

An example of applying the majority method is presented in Figure 3. Six segments are assigned to the AFib class (labelled with 1), and 15 segments with the normal class (labeled with 0). Note that three segments can not be determined (labeled with #) due to missing data to apply the majority vote rule in the segment.

Figure 3 presents also the result of applying the pure method. Only two segments are assigned to the AFib class (labeled with 1) and 9 segments to the nAfib class (labeled with 0). Note 13 segments can not be determined since the segments are with mixed rhythm episodes.

After extracting the features, the data generated from each ECG data base is divided by a ratio of 80/10/10, where 80% of the data is for training, 10% for validation and 10% for testing. 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. During the training procedure we used 3-fold cross validation.

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.

IV. Evaluation methods

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.

To evaluate the performances of the trained models, in the validation and testing phase the models process the data for validation or testing in the same input shape as the training data used so far, but only the segment class in the feature set is not exposed to the models, they have to predict it. Once the models have predicted the classes of the appropriate segments, the results are processed so that the data returns to its original state as ECG recordings labeled with the newly predicted classes. The predicted results are then compared with the original reference data set for validation and testing and the model performance are calculated. Namely, the segments are subdivided into the constituent RR intervals assigned to the corresponding class, her one RR interval may belong to the segments of the two classes, thus applying the majority method that assigns the class with the most RR interval in the segments.

Sensitivity (*SEN*), also known as hit rate or true positive rate (TPR), measures the proportion of correctly identified positive cases (sequences correctly classified as AFib) in regards to the actual number of positive cases.

Positive Predictivity Value (*PPV*), also known as precision, is the proportion of positive results that are true positives (sequences correctly classified as AFib) in regards to the total number of positive results

A statistical measure of a test's accuracy that combines SEN and PPV is known as F1 score. The F1 score, also called F score or F Measure, is calculated by the harmonic mean of the precision and sensitivity.

Due to the large research dimension and the large number of ML algorithms with different feature engineering and segment labeling during training, when presenting and comparing

the results, an Improvement Factor IF *(Improvement Factor)* metric has been introduced which is computed by comparing the F1-score values of the performance of the analyzed and reference algorithm.

V. Results

Training was conducted on LTAFDB with RR intervals as features. The results are presented in Table 4 for each odd length of segments from 5 to 49 (23 segments) and for both majority and pure segment labeling methods.

	majority			pure		
Segment length	SVM	DT	RF	SVM	DT	RF
5	63.66%	84.94%	85.28%	69.04%	85.02%	85.32%
7	65.71%	86.11%	86.17%	49.65%	85.96%	86.22%
9	69.60%	86.50%	86.56%	69.57%	86.36%	86.40%
11	61.43%	86.84%	86.74%	65.59%	86.43%	86.72%
13	69.08%	87.04%	86.89%	60.36%	86.47%	86.90%
15	70.64%	87.22%	87.30%	64.13%	86.54%	87.14%
17	62.00%	87.19%	87.53%	60.06%	86.69%	87.54%
19	43.71%	87.08%	87.87%	68.27%	86.85%	87.76%
21	64.63%	87.37%	87.94%	65.53%	86.94%	88.05%
23	61.45%	87.11%	88.26%	64.79%	86.86%	87.78%
25	53.25%	87.47%	88.24%	64.81%	86.87%	88.16%
27	50.64%	87.69%	88.50%	56.87%	86.85%	88.11%
29	66.74%	87.72%	88.43%	57.72%	87.01%	88.05%
31	44.65%	87.71%	88.55%	65.49%	87.16%	88.37%
33	61.66%	87.82%	88.58%	65.68%	87.21%	88.44%
35	65.82%	87.82%	88.86%	65.56%	87.40%	88.57%
37	58.65%	87.95%	88.82%	65.40%	87.36%	88.75%
39	69.54%	88.00%	88.79%	65.55%	87.44%	88.58%
41	65.32%	88.42%	89.05%	65.37%	87.53%	88.68%
43	65.50%	88.49%	89.10%	65.40%	87.52%	88.99%
45	64.19%	88.30%	89.02%	65.32%	87.58%	88.86%
47	65.75%	88.31%	89.13%	65.25%	87.66%	88.82%
49	65.62%	88.40%	89.10%	65.27%	87.84%	88.90%

Fig. 4. F1 score of ML based AFib detection algorithms with majority and pure methods

VI. Discussion

Table 5 shows the improvement factor values of all tested ML algorithms and feature engineering segment labeling methods for different segment length. The average values of these results lead to a conclusion that the improvement factor is positive in DT and RF algorithms which means that the majority method is better, while for SVM it is worse.

Since RF outperformed DT and SVM we conclude that the majority method for segment labeling in feature extraction process is bette than the method of pure segments. Note that the majority method includes segments containing features of both the AFib and non-AFib classes, they are one type of transition from one rhythmic episode to another making their number in the whole dataset insignificant, but good to include so that algorithms would know how to handle such situations.

VII. Conclusion

In this paper, we experiment in determining the optimal method for labeling segments with a certain length of consecutive RR intervals for training the models used for detecting

	ratio majority/pure validation		
segment length	SVM	DT	RF
5	-7.78%	-0.09%	-0.05%
7	32.34%	0.17%	-0.05%
9	0.05%	0.16%	0.18%
11	-6.34%	0.47%	0.03%
13	14.44%	0.65%	-0.01%
15	10.15%	0.79%	0.18%
17	3.24%	0.58%	-0.01%
19	-35.97%	0.26%	0.13%
21	-1.38%	0.49%	-0.12%
23	-5.15%	0.28%	0.55%
25	-17.84%	0.69%	0.09%
27	-10.95%	0.96%	0.43%
29	15.62%	0.82%	0.43%
31	-31.83%	0.62%	0.21%
33	-6.12%	0.70%	0.15%
35	0.39%	0.48%	0.32%
37	-10.32%	0.67%	0.07%
39	6.08%	0.64%	0.24%
41	-0.07%	1.01%	0.41%
43	0.15%	1.11%	0.13%
45	-1.73%	0.83%	0.18%
47	0.77%	0.75%	0.35%
49	0.53%	0.64%	0.22%
Average	-2.25%	0.60%	0.18%

Fig. 5. Improvement factor of majority versus pure method

and AFib in ECG recordings. The majority method 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] P. A. Wolf, R. D. Abbott, and W. B. Kannel, "Atrial fibrillation as an independent risk factor for stroke: the framingham study.," *Stroke*, vol. 22, no. 8, pp. 983–988, 1991.

[2] J. Heeringa, D. A. van der Kuip, A. Hofman, J. A. Kors, G. van Herpen, B. H. C. Stricker, T. Stijnen, G. Y. Lip, and J. C. Witteman, "Prevalence, incidence and lifetime risk of atrial fibrillation: the rotterdam study," *European heart journal*, vol. 27, no. 8, pp. 949–953, 2006.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

[4] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[5] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[6] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ecg signals using machine learning techniques: A survey," in *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 714–721, IEEE, 2015.

[7] R. Colloca, A. E. Johnson, L. Mainardi, and G. D. Clifford, "A support vector machine approach for reliable detection of atrial fibrillation events," in *Computing in Cardiology 2013*, pp. 1047–1050, IEEE, 2013.

[8] S. Datta, C. Puri, A. Mukherjee, R. Banerjee, A. D. Choudhury, R. Singh, A. Ukil, S. Bandyopadhyay, A. Pal, and S. Khandelwal, "Identifying normal, af and other abnormal ecg rhythms using a cascaded binary classifier," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[9] J. Hu, W. Zhao, Y. Xu, D. Jia, C. Yan, H. Wang, and T. You, "A robust detection method of atrial fibrillation," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45, pp. 1–4, IEEE, 2018.

[10] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *Europace*, vol. 9, no. 7, pp. 466–470, 2007.
[11] M. Gusev and M. Boshkovska, "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.