33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# How the Choice of Samples for Building Arrhythmia Classifiers Impact Their Performances

Eduardo Luz and David Menotti

*Abstract*— **Arrhythmia (*i.e.*, irregular cardiac beat) classification in electrocardiogram (ECG) signals is an important issue for heart disease diagnosis due to the non-invasive nature of the ECG exam. In this paper, we analyze and criticize the results of some arrhythmia classification methods presented in the literature in terms of how the samples are chosen for training/testing the classifier and the impact this choice has on their performance (*i.e.*, accuracy/sensitivity/specificity). From our implementation, we also report new accuracies for these methods, establishing a new state-of-the-art method, in terms of results.**

## I. INTRODUCTION

The electrocardiogram (ECG) is the most widely used non-invasive technique in heart disease diagnoses. It can be described as a record of the electrical phenomena originated from cardiac activity. Fig. 1 shows a schematic record of a normal heartbeat. The ECG is frequently used to detect cardiac rhythm abnormalities, otherwise known as, arrhythmias. Arrhythmias can be defined in two ways: as a unique irregular cardiac beat or as a set of irregular beats. Arrhythmias can be rare and harmless, but may also result in serious cardiac issues.

There are several methods proposed in the literature for the purpose of automatic arrhythmia classification in ECG signals and a complete system for such an aim can be divided into four subsequent categories (as shown in Fig. 2): preprocessing, segmentation, feature extraction, and classification. This study focuses on the last step. The most widely used database for evaluation of the accuracy/sensitivity/specificity (from now on performance) of arrhythmia classification systems is the MIT-BIH Arrhythmia Database [1]. This database was the first available for such a purpose and it has gone through several improvements over the years to encompass the broadest possible range of waveforms [2]. The Association for the Advancement of Medical Instrumentation (AAMI) also recommends the use of the MIT-BIH Arrhythmia Database for performance evaluation of arrhythmia systems. The AAMI has developed a standard for testing and reporting performance results of algorithms aiming at arrhythmia classification (ANSI/AAMI EC57:1998/(R)2008). According to [3], [4] few researchers have used the AAMI recommendations and standards, leading to clinically unreliable results since several methods in the literature are favored by a biased dataset (*i.e.*, where heartbeats from the same patient are used for both training and testing the classifiers, which makes a fair comparison among methods difficult).

Eduardo Luz and David Menotti are with the Department of Computing, Universidade Federal de Ouro Preto, 35400-000 Ouro Preto (MG), Brazil {eduluz,menottid}@gmail.com

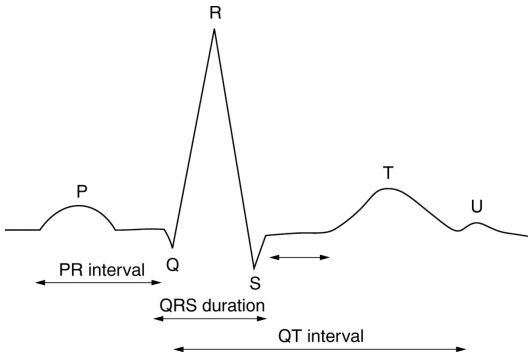


Fig. 1. A normal heartbeat ECG signal

Nevertheless, several researchers have decided not to follow the AAMI recommendations [5], [6], [7], [8], [9], reporting performance in average near to 100% as shown in Table I. The main aim of this work is to re-implement/reproduce those methods and to perform new experiments following the AAMI recommendations in order to analyze the impact of this change on the performance of those methods.

The remainder of this work is organized as follows. The methods used in our analysis are described in Section II and experiments in Section III. A discussion on the results reported for those studies is shown in Section IV. And, finally, conclusions are pointed out in Section V.

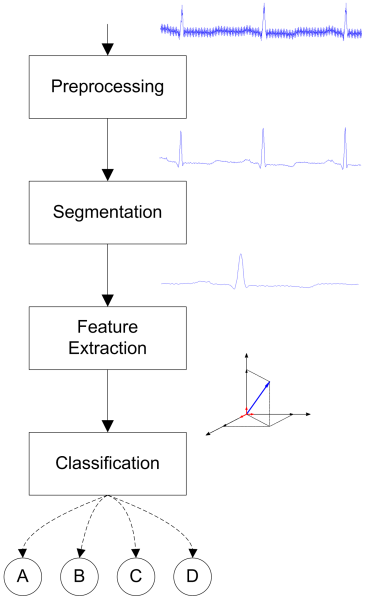


Fig. 2. A diagram of a classification system of arrhythmia

TABLE I

CLASSIFICATION PERFORMANCE OF METHODS USING RANDOM SELECTION OF SAMPLES (HEARTBEATS) - BIASED SELECTION

| Method | Accuracy | Sensitivities (%) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | L | R | A | V | P | a | ! | F | x | j | f | E | J | e |
| Ye *et al.* [5] | 99.91 | 99.95 | 100 | 99.99 | 99.65 | 99.26 | 100 | 92.86 | 100 | 99.73 | 100 | 100 | 100 | 100 | 97.06 | 100 |
| Yu & Chen [6] | 99.65 | 99.97 | 99.33 | 99.54 | 99.76 | 99.04 | 100 | - | - | - | - | - | - | - | - | - |
| Yu & Chou [7] | 98.71 | 99.65 | 96.25 | 99.15 | 98.40 | 98.45 | 99.37 | - | 90.12 | - | - | - | - | 91.53 | - | - |
| Güler & Übeyli [8] | 96.94 | 97.78 | - | - | 97.78 | 95.56 | - | - | - | - | - | - | - | - | - | - |
| Song *et al.* [9] | 99.35 | 99.65 | - | - | 88.29 | 92.15 | - | - | 99.75 | - | - | - | - | - | - | - |

## II. METHODS

In this section, we describe six methods so that we may further analyze their performances. One of them, in our consideration, is a state-of-the-art method, since its authors have followed the AAMI recommendations [3]. In the remaining five methods, we focus on analysis since their authors did not follow the AAMI recommendations [5], [7], [6], [8], [9]. These last methods report performance in average near to 100% as shown in Table I.

In [3], a system to classify ECG heartbeats into five types of heartbeats (classes), which are the ones recommended by the AAMI standards, using single and multiple ECG leads is introduced. Several feature sets are proposed based on the combinations of ECG morphology, heartbeat intervals, and RR-intervals. All configurations are tested with a statistic classifier model (linear discriminant) using supervised learning. Approximately 50,000 heartbeats from the MIT-BIH arrhythmia database were used for the supervised learning model and over 50,000 different heartbeats for testing. A new approach for arrhythmia classification based on morphological and dynamic features is proposed in [5]. Coefficients of wavelet transform (WT) and independent component analysis (ICA) are extracted and used as morphological features. The authors define RR-interval information as dynamic features. A Support Vector Machine (SVM) is used for the classification of heartbeats into 15 types of heartbeats and the authors use two leads for final classifier decisions. A total of 110,076 heartbeats from the MIT-BIH arrhythmia database are used in that work, in which 85,945 heartbeats are exclusively for testing.

In [9], the authors also use WT to extract 17 features from a 400ms window centered at R peak of the QRS complex, and SVM to classify arrhythmias in ECG signal sampled from the MIT-BIH arrhythmia database. Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) are employed to reduce feature dimensions, resulting in higher classification performances. The system is designed to classify six types of heartbeats, in which 4,135 heartbeats are selected for training and another 85,630 heartbeats for testing. All heartbeats are extracted from the MIT-BIH arrhythmia database and the MIT-BIH malignant ventricular arrhythmia database. A method based on ICA and probabilistic neural network (PNN) is proposed in [7]. ICA is used to decompose the ECG signal into coefficients. These coefficients, along with RR-interval, are used as features for the PNN to classify eight types of heartbeats from 9,800 heartbeats randomly selected from the MIT-BIH arrhythmia database. A similar work is proposed in [6] to classify 6 types of heartbeats using PNN as a classifier and WT coefficients to build the feature vector, along with RR-interval information.

A combined neural network model is proposed in [8] to accomplish the task of arrhythmia classification. The ECG signal is decomposed into wavelet coefficients to represent the morphology and statistical features to depict their distribution. Two neural network levels are employed; the second one receiving the outputs of first level networks as input data. Four types of heartbeats, obtained from the MIT-BIH database are classified. In that experiment, 360 beats are used for training and 360 beats for testing. Notice that the six methods described here and used in our analyses and are re-implemented as faithfully as possible to their description using Matlab[1].

## III. EXPERIMENTS AND RESULTS

The MIT-BIH arrhythmia database contains 48 half-hour recordings, sampled at 360Hz, and eighteen types of heartbeats were classified and labeled. To comply with the AAMI recommendations, only 44 recordings from the MIT-BIH arrhythmia database should be used for evaluation of arrhythmia classification methods, excluding the 4 recordings that contain paced beats. The ANSI/AAMI EC57:1998/(R)2008 standard recommends to group those heartbeats into five classes: 1) normal beat; 2) ventricular ectopic beat (VEB); 3) supraventricular ectopic beat (SVEB); 4) fusion of a VEB and a normal beat; and 5) unknown beat type, as shown in Table II). Moreover, the AAMI standards also recommend dividing the recordings into two datasets: one for training and another for testing such that heartbeats from one recording (patient) are not used simultaneously for both training and testing the classifier.

Two datasets are created, one following the AAMI recommendations (DS1) and another which does not (DS2). In order to build DS2, $50\%$ of patients' heartbeats are randomly selected for training (DS2-train) and the remaining for testing (DS2-test). The training partition (DS1-train) is composed of heartbeats from recordings 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230, whilst the test partition (DS1-test) of 100, 103, 105, 11, 113, 117, 121, 123, 200, 202, 210, 212, 213,

[1]The source code of implementations are available at http://code.google.com/p/embc-ecg-paper/.

TABLE II
MAPPING THE MIT-BIH ARRHYTHMIA TYPES TO THE AAMI CLASSES

| The AAMI heartbeat class | N | SVEB | VEB | F | Q |
|---|---|---|---|---|---|
| Description | Any heartbeat not in the S, V, F, or Q class | Supraventricular ectopic beat | Ventricular ectopic beat | Fusion beat | Unknown beat |
| MIT-BIH heartbeat types (code) | normal beat (N) | atrial premature beat (A) | premature ventricular contraction (V) | fusion of ventricular and normal beat (F) | paced beat (P) |
| | left bundle branch block beat (L) | aberrated atrial premature beat (a) | ventricular escape beat (E) | | fusion of paced and normal beat (f) |
| | right bundle branch block beat (R) | nodal (junctional) premature beat (J) | | | unclassified beat (U) |
| | atrial escape beat (e) | supraventricular premature beat (S) | | | |
| | nodal (junctional) escape beat (j) | | | | |

214, 219, 221, 222, 228, 231, 232, and 234, adding up to approximately 100,000 beats.

All methods are run with both datasets (DS1 and DS2) and the results are shown in Tables III and IV. Small differences in terms of performance of some classes are observed between results reported in this work and the ones reported by the authors of the methods. We suggest that these differences are related to some missing implementation details, *e.g.*, lack of information about the digital filters or classifier parameter values.

## IV. DISCUSSIONS

Before starting our analysis on the classification performance, we present the three measures employed: *i.e.*, accuracy, sensitivity, and specificity. Accuracy is defined as the ratio of total beats correctly classified and the number of total beats,

$$Accuracy = \frac{beats\ correctly\ classified}{number\ of\ total\ beats}. \quad (1)$$

Sensitivity can be defined as the ratio of correctly classified beats of one class and the total beats classified as that class, including the missed classification beats, *i.e.*,

$$Sensitivity = \frac{true\ positives}{true\ positives\ +\ false\ negatives}. \quad (2)$$

Specificity stands for the ratio of correctly classified beats among all beats of a specific class, *i.e.*,

$$Specificity = \frac{true\ negatives}{true\ negatives\ +\ false\ positives}. \quad (3)$$

Sensitivity and specificity are the most important measure for our analysis, since the number of heartbeats for each class in the MIT-BIH arrhythmia database is very imbalanced and a single class (*e.g.*, the normal beats) could represent most of the total accuracy, while the sensitivity and specificity directly depend on the number of samples for each class.

Comparing the results achieved by our implementation of methods using dataset DS1 and DS2, shown in Tables III and IV, respectively, we can see a significant difference in terms of sensitivity and specificity. This observation can be extended to the accuracy figures. All values of measures present higher values when the methods do not follow the AAMI recommendations. However, the use of a dataset for training a classifier and then testing it with samples (heartbeats) from the same patients, in other words, the methods which do not follow the AAMI recommendations, helps the classifier to yield better classification results, since it is specialized in those data. It is worth pointing out that all methods employed in the analysis in this study are consistent and use advanced techniques to solve the arrhythmia classification problem.

In addition to the fact that heartbeats from the same recording used both for training and testing may favor the classifier, there is another practice that can lead to biased conclusions as well. Several methods do not use the complete data from the MIT-BIH arrhythmia database as done in [6] and [7], where only 23,200 and 9,800 heartbeats are used, respectively. In those instances, the heartbeats were randomly chosen and the classifiers may have been favored by easy-to-classify heartbeats.

Moreover, according to [4], only a few of the methods presented in the literature have, in fact, used the AAMI standards. The analysis of figures in Tables III and IV suggests that the results of several methods in the literature are unreliable and should not be taken into account clinically before a robust performance test can be performed.

There is also a lack of standards regarding classes of heartbeats to be analyzed. In some cases, the classifiers are designed to classify a specific number of classes, *e.g.*, 2, 3, 10. In other cases, the authors present the performance of methods for non-standard classes (*i.e.*, non-arrhythmia beat annotation codes), such as Ventricular Flutter Wave (!) and

TABLE III

CLASSIFICATION PERFORMANCE OF METHODS FOLLOWING THE AAMI RECOMMENDATIONS. * RESULTS OBTAINED WITH FEATURE SETS FS1 AND FS2. # ONLY LEAD DII IS USED.

| Method | Accuracy (%) | Sensitivity/Specificity (%) | | | | |
|---|---|---|---|---|---|---|
| | | N | SVEB | VEB | F | Q |
| Chazal *et al.* [3]* | 75.35 | 77.62/97.35 | 19.90/ 8.59 | 73.35/53.00 | 89.79/ 6.11 | 0.00/ 0.00 |
| Ye *et al.* [5]# | 75.15 | 80.20/78.15 | 3.16/10.30 | 50.17/48.51 | 0.00/ 0.00 | 0.00/ 0.00 |
| Yu & Chen [6] | 73.87 | 81.46/74.24 | 0.00/ 0.00 | 20.96/59.40 | 0.00/ 0.00 | 0.00/ 0.00 |
| Yu & Chou [7] | 75.21 | 78.30/79.19 | 1.81/ 5.92 | 83.89/66.42 | 0.26/ 0.08 | 0.00/ 0.00 |
| Güler & Übeyli [8] | 66.70 | 69.17/72.08 | 0.00/ 0.00 | 78.81/43.79 | 1.80/ 0.48 | 0.00/ 0.00 |
| Song *et al.* [9] | 76.29 | 77.99/83.89 | 27.00/48.34 | 80.75/38.67 | 0.00/ 0.00 | 0.00/ 0.00 |

TABLE IV

CLASSIFICATION PERFORMANCE OF METHODS NOT FOLLOWING THE AAMI RECOMMENDATIONS. * RESULTS OBTAINED WITH COMBINED FEATURE SETS FS1 AND FS5. # ONLY LEAD DII IS USED.

| Method | Accuracy (%) | Sensitivity/Specificity (%) | | | | |
|---|---|---|---|---|---|---|
| | | N | SVEB | VEB | F | Q |
| Chazal *et al.* [3]* | 86.01 | 86.55/99.28 | 81.48/31.04 | 79.99/55.75 | 85.50/13.44 | 25.00/ 1.79 |
| Ye *et al.* [5]# | 96.53 | 98.73/96.31 | 72.35/94.54 | 82.56/97.81 | 65.59/88.55 | 95.77/99.33 |
| Yu & Chen [6] | 81.10 | 85.20/81.23 | 0.00/ 0.00 | 69.97/79.17 | 0.00/ 0.00 | 0.00/ 0.00 |
| Yu & Chou [7] | 95.39 | 96.86/97.32 | 73.75/88.40 | 92.26/94.28 | 51.00/73.38 | 94.13/80.84 |
| Güler & Übeyli [8] | 89.06 | 93.20/90.25 | 0.00/ 0.00 | 81.56/74.63 | 0.00/ 0.00 | 0.00/ 0.00 |
| Song *et al.* [9] | 98.66 | 99.50/98.94 | 86.39/94.27 | 95.83/97.42 | 73.57/90.21 | 0.00/ 0.00 |

Non-Conducted P-wave (x) [5], [7].

## V. CONCLUSIONS

In this paper, we have analyzed and criticized the results of some arrhythmia classification methods presented in the literature in terms of how the samples have been chosen for training/testing the classifier and the impact of this choice on their performance.

Researchers have been working on improvements and many of them have shown remarkable results. Nevertheless, few authors have considered the impact on the performance of the classifiers caused by the way the samples (heartbeat) were selected for building the dataset used for training and testing the classifiers.

We have discussed, described and re-implemented methods that may use heartbeats from the same patients for training and testing classifiers, which favor their results in terms of performance. We then have run experiments for the re-implemented methods using datasets that both follow and do not follow the AAMI recommendations. The resulting performances for the methods support our claims. That is, how the samples (heartbeats) are chosen for classification learning imposes a bias in the performance results. In addition, overall, the data used from the same patient should not be used for training and testing a classifier. This practice, *i.e.*, of putting data from the same patient in both sets should be avoided as already stated in [3]. Moreover, another contribution, after following our implementations, has shown we were able to establish new performance values for the studied methods.

Thus, the choice of an unbiased dataset, as recommended by the AAMI standards, should be used for arrhythmia classification methods in order to obtain reliable results. With this fact in mind, several methods in the literature can be redone using unbiased datasets. These results should be used to report new prediction values for these methods, establishing a new state-of-the-art method in terms of performance.

## REFERENCES


[1] "MIT-BIH ECG database," available at http://ecg.mit.edu/.

[2] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[3] P. Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.

[4] T. Ince, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ECG signals," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1415–1427, 2009.

[5] C. Ye, M. T. Coimbra, and B. V. K. V. Kumar, "Arrhythmia detection and classification using morphological and dynamic features of ECG signals," in *IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 1918–1921.

[6] S. Yu and Y. Chen, "Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1142–1150, 2007.

[7] S. Yu and K. Chou, "Integration of independent component analysis and neural networks for ECG beat classification," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2841–2846, 2008.

[8] I. Güler and E. D. Übeyli, "ECG beat classifier designed by combined neural network model," *Pattern Recognition*, vol. 38, no. 2, pp. 199–208, 2005.

[9] M. H. Song, J. Lee, S. P. Cho, K. J. Lee, and S. K. Yoo, "Support vector machine based arrhythmia classification using reduced features," *International Journal of Control, Automation, and Systems*, vol. 3, no. 4, pp. 509–654, 2005.