Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 20, Iss. 2, pp. 34-46, 2022
© Brazilian Society on Computational Intelligence

# INTELLIGENT DETECTION OF ARRHYTHMIA EPISODES IN DIALYSIS PATIENTS

**Sergio Pinto Gomes Junior , João Baptista de Oliveira e Souza Filho**
Electrical Engineering Program (PEE/COPPE) - Federal University of Rio de Janeiro (UFRJ)
sergio.junior@coppe.ufrj.br, jbfilho@poli.ufrj.br

**Felipe da Rocha Henriques**
Instrumentation and Applied Optics Program (PPGIO) and
Computer Science Program (PPCIC) - Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ)
felipe.henriques@cefet-rj.br

**Michel Pompeu Tcheou**
Electronics Engineering Program (PEL) - Rio de Janeiro State University (UERJ)
mtcheou@uerj.br

**Abstract –** This work discusses the design of an automatic detector of arrhythmia episodes in patients submitted to dialysis. The system aims to operate on portable devices in real-time, allowing a faster response of healthcare workers to possible intercurrence episodes. The detection is based on processing short windows of samples extracted from the electrocardiogram signal around the R-wave peak in raw format. A comprehensive study evaluating several classification techniques and class-imbalance strategies is conducted based on the MIT-BIH Arrhythmia Database. Besides, a new procedure for tuning the sample window length based on an experimental feature importance cumulative distribution is proposed. Results show that a Random Forest classifier, trained with minority class oversampling, is cost-effective regarding complexity and computational cost, achieving an accuracy of 98.7% for windows sizes as small as 105 samples.

**Keywords –** Dialysis, arrhythmia, beat classifier, clinical decision support

## 1. INTRODUCTION

Kidneys perform some crucial functions in the human body, such as removing salt and fluid excess, controlling blood pressure, and helping in maintaining the balance of substances like sodium, potassium, urea, and creatinine. In more severe disease presentations, a person with kidney failure may require a dialysis procedure, where a machine performs the sick kidney role [1]. This procedure typically lasts for four hours and should be repeated at least three times a week [2]. Dialysis is an increasingly safer procedure due to technological innovations. However, in approximately 30% of the sessions, some complications may occur, such as hypotension, hypertension, dialysis imbalance syndrome, hypoxemia, and cardiac arrhythmias, with the latter being more frequent [3].

According to the Brazilian Society of Cardiology, cardiac arrhythmias are a result of disturbances in the heart electrical system [4], leading to situations where the beat rate becomes fast (tachycardia), slow (bradycardia) or irregular [2] [3]. Diagnoses provided from healthcare professionals are conducted by analyzing a non-intrusive exam called the electrocardiogram (ECG) [5]. The interpretation of ECG signals is a pattern recognition task, motivating the broad research interest for mathematical models to support this task. Many works exploit the standard Machine Learning (ML) framework of feature extraction followed by classification. In these cases, the process of feature extraction may consider transformations like Wavelet [6] [7] [8] and Principal Components Analysis (PCA) [9]. Alternatively, the direct classification of raw ECG signal excerpts is also common [10] [11]. Regarding classification techniques, Support Vector Machines (SVM) [12] [13], Decision Trees (DT), Random Forest (RF) [14] [15] and Neural Networks (NN) [16] [17] [8] are the most common approaches. Considering features extracted explicitly or not, the classification task can be assumed as multi-category [18] or simply binary [19]. The first alternative includes, besides the regular ones, categories of anomalies like Left and Right Bundle Branch Block, Premature Ventricular Contraction, and Paced beats [18]. The second one only encompasses normal and abnormal categories [19]. The abnormal category integrates ventricular contraction, premature supraventricular beats, and a fusion between ventricular cases. The most common intercurrences in patients undergoing dialysis are premature atrial (PAC) and ventricular (PVC) contractions, detaining a prevalence of 40.30% and 59.70%, respectively [20].

These factors motivated us to propose an automatic arrhythmia detection system to support healthcare professionals when dealing with patients under dialysis, similar to [19], but assuming a classification in only normal and abnormal cases. In this instance, the abnormal situations comprises only PAC and PVC episodes. The proposed system aims promptly reporting arrhythmia episodes in real-time to allow quick and effective medical countermeasures whenever necessary.

The main contributions of our work are as follows:

- A new feature selection procedure is proposed based on an experimental cumulative feature distribution for setting the length of raw ECG signal windows targeting an accurate and low computational complex signal classification.
- An accurate and low-complex model for arrhythmia detection is produced, adequate for low-cost and battery-powered embedded devices, based on a short ECG raw signal window.

The paper structure is outlined as follows. Section 2 briefly describes the basic ECG signal structure. Section 3 presents the proposed system, while Section 4 describes the data used for system evaluation. Section 5 covers experiments to define the most cost-effective classification technique for system design based on a rigorous hypothesis testing, class-imbalance mitigation approaches, and a new process for tuning the sample window length with basis on a cumulative distribution of some feature importance metric; in our case, the average impurity decrease in the Random Forest model. Section 6 discuss our results based on the state of the art. Final remarks are exposed in Section 7.

## 2. ECG SIGNALS

ECG test produces a time-varying signal reflecting the electrical current flow responsible for the contraction and relaxation of cardiac muscle fibers. This signal corresponds to a measure of the electrical potential difference between two electrodes placed on the body of the patient. A normal ECG cycle exhibits the electrical depolarization and repolarization of both the atrium and the ventricle, and it is comprised within a period between consecutive heartbeats [21]. These events are associated with the peaks and valleys of a typical ECG waveform, illustrated in Figure 1. In short, the following elements can be observed in a regular beat [21]:

- P-wave: refers to a region of low voltage, reflecting the atrial depolarization;
- QRS Complex: refers to the ventricle depolarization. It is followed by the atrium repolarization, which is not evident since it is masked by the high amplitude of the QRS complex.
- T-wave: refers to the ventricular repolarization, preparing the heart for a new cycle.

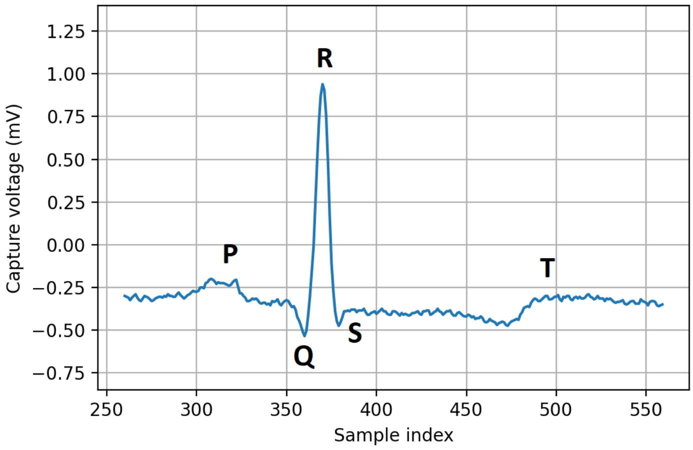

Figure 1: Example of an ECG signal excerpt (sample window) of a regular heartbeat with typical wave patterns labeled. Adapted from [22].

Arrhythmia episodes are typically identified by analyzing the R-R intervals (i.e., between peaks of successive QRS complexes) [23] [24]; P-P intervals; P, T or QRS signal duration, and the interval between P and R or Q and T waves. Morphological features from P, T and QRS waves can also be considered [25].

## 3. ARRHYTHMIA DETECTION SYSTEM OVERVIEW

Figure 2 illustrates the system proposed in this work. It comprises an auxiliary module connected to the electrocardiogram equipment, responsible for receiving the ECG signal, extracting the sample windows, classifying this window, and finally alerting the healthcare team in case of any arrhythmia occurrence. The signal segmentation/windowing procedure is based on the widely-used approach adopted by Pan and Tompkins [26], which considers windows centered on the R-peaks with two seconds duration.

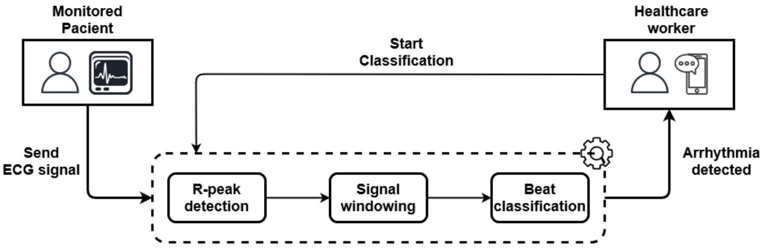

Figure 2: Proposed arrhythmia detection system.

For illustration purposes, Figures 3 and 4 depict ECG signals with PAC and PVC arrhythmia episodes, respectively. Roughly in PAC, the beat occurs earlier than expected, and the QRS complex is normally narrow. The aberrant PAC is an exception since its QRS complex is wider than usual [27]. In turn, for PVC, we have a broader QRS complex and the absence of the P wave [28].

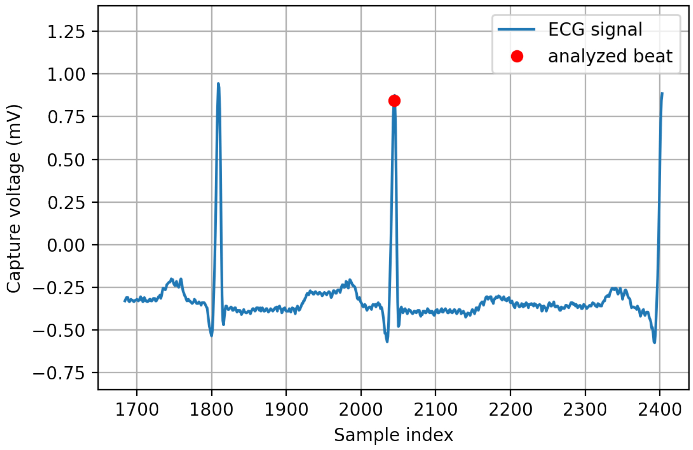

Figure 3: Illustrative ECG signal except of a premature atrial contraction (PAC) arrhythmia episode.

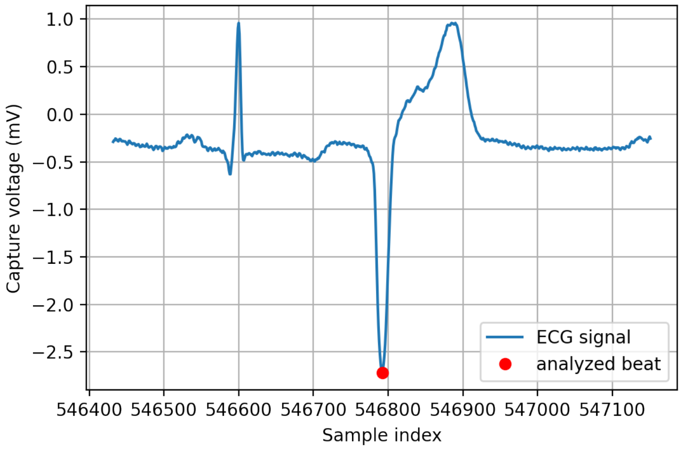

Figure 4: Illustrative ECG signal except of a premature ventricular contraction (PVC) arrhythmia.

## 4. DATASET DESCRIPTION

In our experiments, we have considered signals from the MIT-BIH Arrhythmia Database [22], available on PhysioNet [29], which includes 48 excerpts of ECG recordings from 47 subjects lasting for 30 minutes. Twenty-three excerpts were randomly extracted from a set of 4000 ECG recordings with 24-hours duration. The remaining part was selected by database authors aiming at including uncommon arrhythmias to improve database diversity. Two physicians labeled data as normal or as representative of several arrhythmia modalities. ECG signals were analyzed independently, so disagreements were further settled in a specific meeting between both.

A subset of the database restricted to normal and PAC/PVC arrhythmia cases (assumed as abnormal) was produced, containing 44 excerpts. Subsequently, we produced symmetric sample windows centered on the R-peak, containing 360 (left)$+1$ (R-peak)$+$ 360 (right) $= 721$ samples. As the recordings have been digitized at 360 samples per second, the sample windows last for two seconds. Finally, all windows have their samples normalized to zero mean and unit standard deviation. Figure 5 exhibits a regular heartbeat sample window for the sake of illustration. Table 1 summarizes the final number of sample windows for each class.

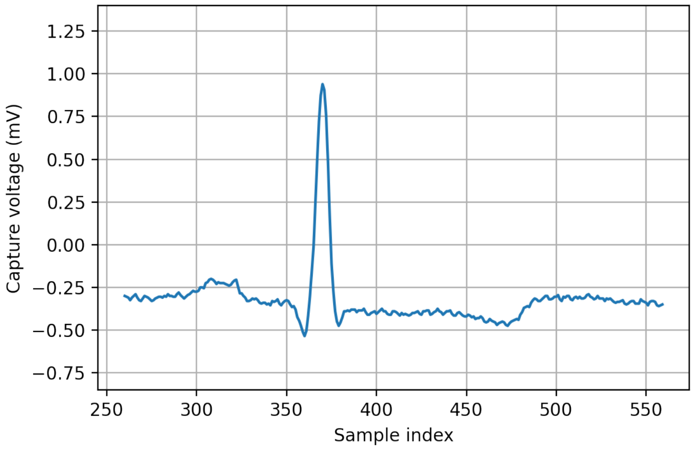

Figure 5: A regular beat sample window example.

Table 1: Number of dataset windows for each class

| Class | Windows |
|---|---|
| Normal beat | 55820 |
| Arrhythmia (PAC + PVC) | 7979 |

## 5. MACHINE LEARNING MODELS

The machine learning classifiers exploited in this work are briefly described in the following.

1. **Decision Trees (DTs)** [30] are simple non-parametric Machine Learning models based on a sequence of binary partitions from the data set, performed considering a set of variables identified as the most relevant for class prediction. However, DTs are unstable and prone to overfit classifiers since a minor change in the training set may significantly modify the tree structure.

2. **AdaBoost** [31] is an ensemble model that combines many simple classifiers, usually implemented by shallow trees, to produce a highly discriminative classification system. The model is additive, thus at each algorithm iteration, a new tree is added to the ensemble. Typically, the new classifiers focus on the instances misclassified in the previous algorithm steps. A drawback of AdaBoost refers to its sensitivity to noisy data and outliers.

3. **Random Forests (RF)** [32] represents an ensemble technique wherein multiple decision trees are generated over bootstrap samples of the data set, however, only considering a small and random subset of the original features. Usually, the final decision is taken by a majority vote over each tree outcome. An additional level of randomness may be added to RF by considering a new random feature subselection at each node split. A significant advantage of RF is conjugating a good performance in many practical problems with an intrinsic feature selection, as based on DTs. An inconvenience resides in a more challenging model interpretation than the standard decision trees.

4. **Support Vector Machine (SVM)** [33] is a binary classifier that generates a hyperplane positioned such that the margins of separation between the instances from both classes are maximized. To address non-linearly separable datasets, SVM may exploit the kernel trick, according to which the input is intrinsically mapped into a high dimensional space aiming to turn data classes linear separable. SVM often shows a high performance but has a complex and computing demanding training procedure.

5. **K-Nearest Neighbors (KNN)** [30] is a non-parametric model having as a premise that elements from a same class have similar features, thus are closer in the data space. The dynamics of KNN is simple. For a given instance, the algorithm identifies the $K$ nearest training set instances, assigning to it a class label defined by majority voting over the labels of these neighbors. As typically implemented using Euclidean distance, KNN may face difficulties when dealing with high-dimensional data.

6. **Gradient Boosting (GB)** [34], similar to Adaboost, exploits the boosting approach but uses an alternative criterion for fusing the models integrating the ensemble based on the gradient of a user-defined loss function. Although it could be applied to any classification method, DTs are the most common base classifier adopted with GB. In this case, GB disposes of an intrinsic feature selection, despite involving a much higher computational cost than RFs.

7. **Logistic Regression** [30] is a simple but powerful method that infers the probability of one event taking place, based on modelling the logarithm of its odds ratio as a linear combination of some set of explicative variables assumed as being independent.

8. **Multilayer Perceptron (MLP)** [35] is a fully connected feedforward network composed of one or more layers of perceptrons. Due to its universal approximation theorem, it is a helpful approach when dealing with complex non-linear problems. An MLP drawback is the high number of model hyperparameters and the difficulties in interpreting the decision rule implemented by the network.

# 6. EXPERIMENTS AND RESULTS

This section summarizes all the experiments performed in this work. First, we cover the classifier and class-imbalance experiments. Then, we describe and report the main findings of the proposed process of window shortening.

## 6.1. CLASSIFIER DEFINITION

In the evaluation process, we have considered eight classification techniques: Support Vector Machine (SVM) [33], Logistic Regression [30], Random Forest [32], Multilayer Perceptron neural network (MLP) [35], K-Nearest Neighbors (KNN) [30], Gradient Boosting [34], Decision Tree [30], and Adaboost [31]. We carried out a grid search procedure to optimally tune the hyperparameters of each technique. The figure of merit was the Area Under the Curve (AUC) [36], due to its robustness to imbalanced data [37]. Experiments considered a stratified ten-fold cross-validation [36], i.e, exploited folds wherein the proportion of arrhythmia (positive cases) and normal beat (negative cases) instances in each one was kept the same as in the original data (stratification). Models had their parameters randomly initialized.

### 6.1.1 GRID SEARCH

We have performed a grid search procedure considering the following set of hyperparameters for each classification model: for SVM, one has the complexity regularization parameter $C$ and the kernel type; for Random Forest, one has the number and the depth of each tree; for Gradient Boosting and for Adaboost, one has the number of trees and the learning rate; for KNN, one has the number of neighbors and the weighting model; for DT, one has the index to evaluate the quality of a candidate node split (Criterion) and the criterion to conduct this split (Splitter); for MLP, one has the number of hidden neurons and the learning rate (only single hidden layer networks with hyperbolic tangent neurons, except to the logistic output neuron, were considered); for Logistic Regression, one has the weight pruning penalty and the use or not of weights in the cost function to mitigate possible class-imbalance effects. For optimizing the NN model, we adopted the Stochastic Gradient Descent (SGD) [35] and the Cross-Entropy cost function.

Table 2 summarizes the hyperparameters, their associated range, and the best configuration (the optimal model) identified for each method. In turn, Table 3 depicts the average (Avg) and standard deviation (Std) values for AUC, recall, precision and accuracy metrics for each optimal model inferred over the ten testing sets. One may readily note that SVM, Random Forest, and Gradient Boosting models achieved the best AUC and precision values. Regarding accuracy and recall, the values obtained are among the best in most cases. This finding leads us to consider the SVM, Random Forest, and Gradient Boosting models as strong candidates for our system design.

### 6.1.2 STATISTICAL ANALYSIS

Targeting to identify the most cost-effective classification technique for the devised system, we performed a rigorous statistical testing, considering the best setup for each method identified in Subsection 6.1.1 and restricting to AUC metrics. The Friedman test [38] was employed to identify if the differences in performance of the methodse were statistically significant in an overall sense for a significance level $\alpha = 5\%$. The following steps summarize this test:

1. Build a matrix $\mathbf{P} \in \mathbb{R}^{n \times k}$ with entries $P_{ij}$ given by the AUC values obtained with the $j$th method when evaluated with the $i$th testing set. The dimensions $n$ and $p$ correspond to the number of testing sets and methods under comparison, respectively.

Table 2: Summary of models, hyperparameters, ranges, and the best cases identified in the experiments.

| Classifier | Hyperparameters | Values | Best Configuration |
|---|---|---|---|
| SVM | C<br>Kernel | {0.1, 1, 1.5}<br>{sigmoid, rbf} | C: 1.5<br>Kernel: rbf |
| Random Forest | Depth<br>Trees | {20, 25, 30}<br>{80, 90, 100} | Depth: 30<br>Trees: 100 |
| Gradient Boosting | Trees<br>Learning rate | {80, 90, 100}<br>{0.001, 0.01, 0.1, 1} | Trees: 100<br>Learning rate: 0.1 |
| Adaboost | Trees<br>Learning rate | {80, 90, 100}<br>{0.001, 0.01, 0.1, 1} | Trees: 100<br>Learning rate: 1 |
| KNN | Neighbors<br>Weighting model | {4, 5, 6}<br>{uniform, distance} | Neighbors: 6<br>Weights: distance |
| Decision Tree | Criterion<br>Splitter | {gini, entropy}<br>{best, random} | Criterion: entropy<br>Splitter: random |
| MLP | Learning rate<br>Hidden neurons | {0.001,0.01,0.1}<br>{30, 35, 40} | Learning rate: 0.001<br>Hidden neurons: 30 |
| Logistic Regression | Penalty<br>Class weight | {l1, l2}<br>{balanced, None} | Penalty: l2<br>Class weight: Balanced |

Table 3: Average and standard deviation of some performance metrics derived over the ten testing folds for each classification technique evaluated (see text).

| Classifier | Avg AUC | Std AUC | Avg Recall | Std Recall | Avg Precision | Std Precision | Avg Accuracy | Std Accuracy |
|---|---|---|---|---|---|---|---|---|
| SVM | 0.9932 | 0.0012 | 0.9209 | 0.010 | 0.9911 | 0.0028 | 0.9891 | 0.0015 |
| Random Forest | 0.9923 | 0.0011 | 0.9123 | 0.0072 | 0.9899 | 0.0032 | 0.9878 | 0.0008 |
| Gradient Boosting | 0.9904 | 0.0013 | 0.8661 | 0.0117 | 0.9809 | 0.0031 | 0.9811 | 0.0015 |
| Adaboost | 0.9831 | 0.0024 | 0.8266 | 0.0101 | 0.9332 | 0.0077 | 0.9709 | 0.0015 |
| KNN | 0.9774 | 0.0040 | 0.9095 | 0.0136 | 0.9766 | 0.0055 | 0.9860 | 0.0017 |
| Decision Tree | 0.9562 | 0.0025 | 0.9261 | 0.0087 | 0.9170 | 0.0045 | 0.9802 | 0.0012 |
| MLP | 0.9543 | 0.0563 | 0.8476 | 0.1424 | 0.8344 | 0.2356 | 0.9279 | 0.1086 |
| Logistic Regression | 0.9484 | 0.0045 | 0.6543 | 0.0122 | 0.8579 | 0.0119 | 0.9432 | 0.0028 |

2. Compute the ranking matrix $\mathbf{R}$ with elements $R_{ij}$ by sorting each column of $\mathbf{P}$ in a decreasing order, filling the corresponding columns of $\mathbf{R}$ $(R_{\cdot j})$ with the set of sorting indexes identified in the first process.

3. Compute the following quantities:

   - The average rank of each classifier (i.e., over $\mathbf{R}$ columns):

$$\overline{R}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} R_{ij}, \quad 1 \leq j \leq k \tag{1}$$

   - The overall mean rank:

$$\overline{R} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} R_{ij} \tag{2}$$

   - The sum of squares total:

$$SS_{Total} = k\sum_{j=1}^{n}(\overline{R}_{\cdot j} - \overline{R})^2 \tag{3}$$

   - The sum of squares error:

$$SS_{Error} = \frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(\overline{R}_{ij} - \overline{R})^2 \tag{4}$$

4. Finally, one must produce the following test statistic:

$$\chi_F^2 = \frac{SS_{Total}}{SS_{Error}}. \tag{5}$$

Note that the critical value $\chi_c$ is the smallest empirical value $\chi_c$ such that $P(\chi_F^2 > \chi_c) \leq \alpha$ [39] and the $p$-value is defined as $p = P(\chi_F^2 > \chi_c)$. Therefore, the null hypothesis must be rejected, i.e. the observed differences would be assumed as statistically significant whenever $p \leq \alpha$ or, equivalently, $\chi_F^2 > \chi_c$. Critical values associated with the triplets $n$, $k$, and $\alpha$ can be found in F distribution tables or approximated by a $\chi^2$ distribution with $k-1$ degrees of freedom when $kN \geq 30$ [39]. In our experiments, $\chi_F^2 = 56.97$ and $p = 3.2 \times 10^{-10}$, indicating that the methods performed differently.

Aiming to identify which methods stand out, Post-hoc tests are necessary to evaluate pairwise comparisons between them [36]. For this, we adopted the Nemenyi test [40] that employs the following statistic when comparing two arbitrary methods $i$ and $j$

$$q_{ij} = \frac{\overline{R}_{.i} - \overline{R}_{.j}}{\sqrt{\frac{k(k+1)}{6n}}}. \tag{6}$$

In this case, the critical values $q_\alpha$ are based on the Studentized range statistic divided by $\sqrt{2}$. Thus, we must assume that two methods performed differently whenever $q_{ij} > q_\alpha$. Alternatively, one may consider the absolute difference between the average ranks of two methods defined as $adr_{ij} = |\overline{R}_{.j_1} - \overline{R}_{.j_2}|$ with which the critical difference (CD) [40] may be computed as

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}. \tag{7}$$

In this case, the null hypothesis must be rejected whenever $adr_{ij} > CD$. Table 4 depicts a summary of Nemenyi test findings, wherein the "true" label corresponds to the methods identified as performing similarly. One may observe that the top three previous analysis methods (SVM, Random Forest, and Gradient Boosting) attained a similar performance, outperforming all alternative algorithms ($p < 0.04$). In terms of computational cost, especially in the operational phase, the Random Forest and Gradient Boosting options are often less computationally demanding than SVM, providing attractive solutions for portable devices.

Table 4: Summary of Nemenyi test outcomes (the label "true" represents methods statistically equivalent.)

| Classifier | Gradient Boosting | Random Forest | Adaboost | KNN | Decision Tree | MLP | Logistic Regression |
|---|---|---|---|---|---|---|---|
| SVM | true | true | true | false | false | false | false |
| Gradient Boosting | — | true | true | true | false | true | false |
| Random Forest | — | — | true | false | false | true | false |
| Adaboost | — | — | — | true | true | true | true |
| KNN | — | — | — | — | true | true | true |
| Decision Tree | — | — | — | — | — | true | true |
| MLP | — | — | — | — | — | — | true |

### 6.2. IMBALANCED LEARNING

Imbalanced learning denotes any process of data representation or information extraction with data representability issues. The class imbalance may constitute a severe problem when developing classifiers, primarily when the number of class instances significantly differ, which may compromise the decision boundaries inferred by the classifier [37]. In our case, the imbalance factor is $55820/7979 \approx 7$, corresponding to seven regular heartbeats to each arrhythmia instance. False negatives are especially more harmful to the health of patients than false-positive cases. Thus, the system performance with the minority class is of great importance, motivating us to evaluate strategies to mitigate possible class-imbalance effects over the classifier performance.

Majority class undersampling and minority class oversampling are the two basic class-imbalance strategies considered here. In the first case, a random sampling process reduces the number of dominant class instances to the same number observed in the minority class. In the opposite direction, the second procedure inflates the minority class by artificially generating new samples through Synthetic Minority Oversampling (SMOTE) technique [41], as we did in our case.

The undersampling and oversampling experiments were restricted to the RF model referred in Table 3 (optimal case). We evaluated reducing the original class-imbalance factor to 5, 3, and 1. Figure 6 exhibits a boxplot chart [42] summarizing the results. In this figure, oversampling and undersampling experiments are labeled as OVx and UNx, respectively, where x denotes the class-imbalance factor. The OV1 strategy (i.e., the same number of class instances for both classes) resulted in a higher median. Friedman test pointed out that methods performed differently ($\chi_F^2 = 58, 27$, $p = 1.01 \times 10^{-10}$), while Nemenyi test confirmed that $OV1$ outperformed all methods ($p < 0.007$), unless $OV3$ ($p = 0.75$).

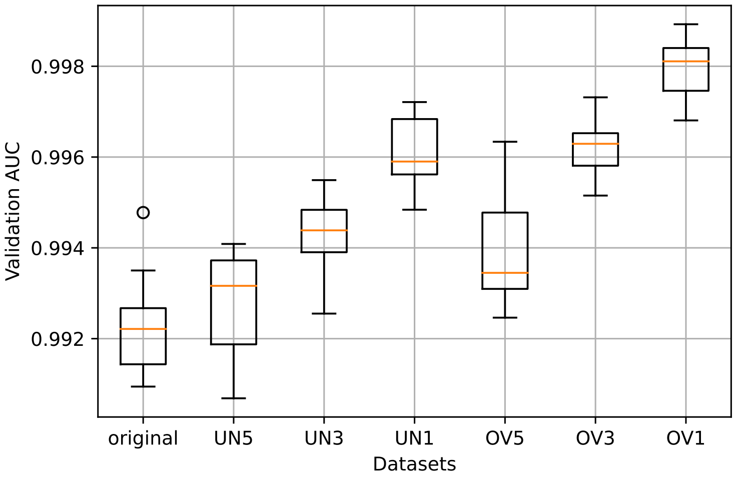

Figure 6: Boxplot chart for the class-imbalance experiments.

### 6.3. WINDOW SHORTENING

Tree-based methods, like Random Forest and Gradient Boosting, have an intrinsic feature selection process. Consequently, feature importance indexes could be promptly obtained in many machine learning packages, like the Scikit-learn [43], measuring the level of a specific input variable contribution to the classifier outcomes. Since our system considers sample windows centered in the R peak of ECG signals, this feature importance information can be handy for tuning the length of the windows. Shorter windows represent less complex and more energy-efficient models, an attractive issue in portable devices, as may result in a higher generalization (i.e., in a better performance in practical settings).

Figure 7 exhibits feature importance index values (FIIV) normalized to sum up one. Window samples closer to the R-peak (centered in zero for convenience) are the most significant for prediction, especially those situated shortly after the peak occurrence.

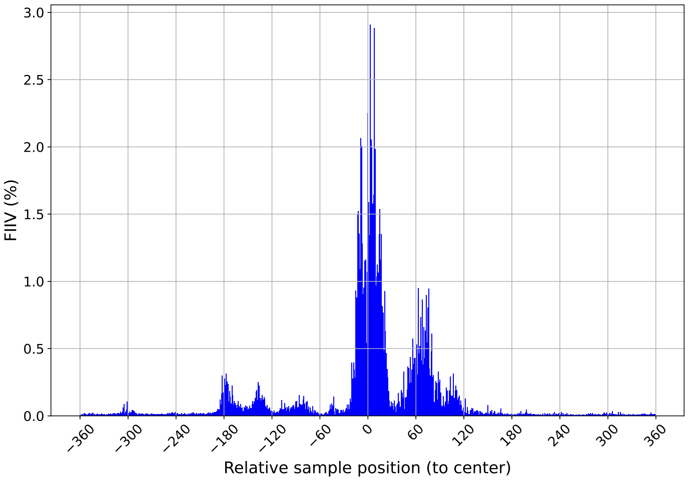

Figure 7: Normalized feature importance values as a function of its relative position in the sample window.

We must stress that these findings are in coherence with typical PAC and PVC patterns observed in the visual analysis of the ECG signal, as illustrated in Figures 3 and 4, motivating us to investigate the use of shorter windows. The experimental process of window length selection proposed here considered the following premises:

- The sample window was assumed symmetric around the R-peak for simplicity.
- Candidate windows had a size $2W + 1$, with $W$ defined by a cumulative feature importance distribution (CFID) given by

$$\mathrm{CFID(W)} = \sum_{n=-W}^{W} \bar{f}i(n), \tag{8}$$

where $\bar{fi}(n)$ is the zero-centered and normalized feature importance index, i.e., $\sum_{n=-W}^{+W} \bar{fi}(n) = 1$. Note that these feature importance indexes can be inferred by any classification algorithm with a intrinsic feature selection, in our case, the Random Forest.

- Based on a set of CFID percentages $\mathcal{P} = \{p_1, p_2, \cdots, p_L\}$ under investigation, one must compute the corresponding set of $W$ values given by $\mathcal{W} = \{W_1, W_2, \cdots, W_L\}$ such that

$$W_i = \lfloor \text{CFID(W)}^{-1} \rfloor, \tag{9}$$

where the operator $\lfloor x \rfloor$ denotes the greatest integer lower or equal to $x$.

Figure 8 exhibits the CFID curve for this experiment. One may readily note that Window sizes with a $W$ as small as 30 retained more than half of feature importance. Values of $W$ equal to 92 and 156 led to CFID values of $80\%$ and $90\%$, respectively. This result is in line with Figure 7, wherein most of the feature importance is concentrated between samples in the range of -160 to +160.

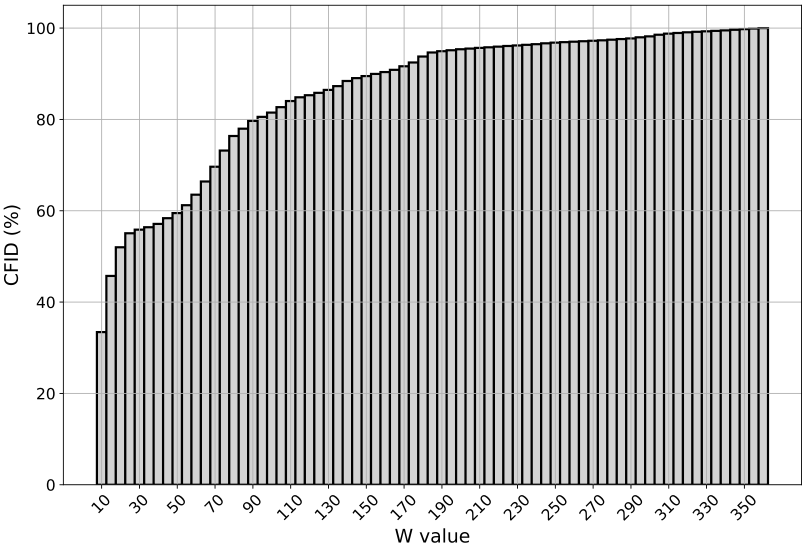

Figure 8: Cumulative feature importance distribution for the average decrease in impurity as feature importance index.

The experimental results considering CFID values from $33\%$ up to $100\%$ (the original model) are summarized in Table 5. Surprisingly, even for values of CFID as small as $33\%$, the system performed surprisingly well. Note that the gains in the average AUC had level off for CFID values greater than $55\%$. The Friedman test results ($\chi_F = 76.91$, $p = 2.04 \times 10^{-13}$) confirmed that the window size indeed affected the performance of the models. In turn, Nemenyi tests demonstrate that models with CFID greater or equal to $60\%$ ($p > 0.9$) performed similarly. Since $W = 52$ for CFID $= 60\%$, the proposed strategy allowed a reduction in the size of the window by a factor of $\frac{2\times360+1}{2\times52+1} = \frac{721}{105} \approx 7$.

Table 5: Window length experiments for different CFID values (see text).

| CFDI (%) | W | Average AUC | Std AUC |
|---|---|---|---|
| 33% | 10 | 0.9834 | $2.97 \times 10^{-3}$ |
| 45% | 15 | 0.9878 | $2.80 \times 10^{-3}$ |
| 52% | 20 | 0.9900 | $2.20 \times 10^{-3}$ |
| 55% | 25 | 0.9905 | $1.66 \times 10^{-3}$ |
| 60% | 52 | 0.9933 | $2.07 \times 10^{-3}$ |
| 70% | 70 | 0.9958 | $0.92 \times 10^{-3}$ |
| 80% | 92 | 0.9967 | $0.81 \times 10^{-3}$ |
| 90% | 156 | 0.9975 | $0.69 \times 10^{-3}$ |
| 100% | 360 | 0.9979 | $0.65 \times 10^{-3}$ |

Figure 9 shows the ROC curve [36] associated with the short-window model ($W = 52$) for the model correspondent to the fold that defines the median value in the cross-validation experiments. At a threshold of $0.8$, the false positive rate (FPR = 1 -

Specificity) is approximately zero, while for a threshold equal to $0.5$, the sensitivity and FPR are $93.1\%$ and $0.4\%$, respectively, while the specificity is $99.6\%$. Finally, for a sensitivity close to $100\%$ (threshold $= 0.05$), the FPR increases to $20\%$, reducing the specificity to $80\%$. It is also worth highlighting that this model achieved an accuracy of $98.7\%$.

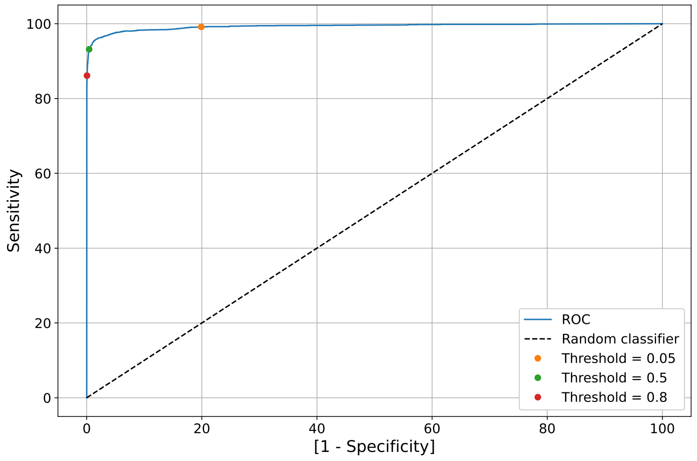

Figure 9: ROC curve of the short-window model (W=52) for the median AUC model (see text).

## 7. DISCUSSION

Several models for an automatic detection of different types of arrhythmia have been proposed in the literature. Table 6 summarizes the performance, the classification techniques, and the datasets considered by some state-of-the-art approaches in heartbeat classification of ECG signals for the MIT-BIH Arrhythmia Database. From this table one can readily perceive that the majority of these studies addressed the classification of multiple heartbeat types, in opposition to our work and [9], which both focused on a binary classification of normal and abnormal heartbeats, the latter integrating premature ventricular and atrial contraction cases. Face the differences of experimental methodologies, such as the number of classes and database partitions, the results reported by each work are not directly comparable. However, generally speaking, our model performed quite satisfactory. Indeed, as shown in Table 6, the reference [19] surpassed the accuracy of our model in approximately 0.8 percentual points. However, we must stress that it is far more complex than ours, involving a deep neural network with seven hidden layers. Therefore, our proposal is more cost-effective regarding performance and energy efficiency. Moreover, the model proposed is in more agreement with the scope of our application since targeting the operation in low-cost embedded devices such as ESP32 [44] and Raspberry Pi [45].

Table 6: State-of-the-art methods for heartbeat classification of ECG signals

| Work | Classes | Classification technique | Accuracy |
|---|---|---|---|
| Dong et al. [18] | Normal, Left and right bundle branch block beat, Premature ventricular contraction and Paced beat | RBF Neural network | 97.8% |
| Inan et al. [46] | Normal, Premature ventricular contraction and Others (Left and right bundle branch block beat, Paced beat and Atrial flutter) | MLP Neural network | 95.2% |
| Martis et al. [9] | Normal e abnormal (Premature ventricular contraction and Premature atrial contraction) | SVM | 98.4% |
| Sannino e De Pietro [19] | Normal and abnormal (Premature ventricular contraction, Supraventricular premature beat and Fusion of paced and normal beat) | Deep Neural Network | 99.5% |
| This work | Normal e abnormal (Premature ventricular contraction and Premature atrial contraction) | Random Forests | 98.8% |

## 8. CONCLUSION

The present work discussed the design of an automated system for promptly alerting healthcare professionals about arrhythmia episodes in patients under dialysis. The system is based on a raw classification of sample windows extracted from the ECG signal centered in the R-peak. A comprehensive study including multiple classification techniques and class-imbalance strategies was conducted. A procedure for tuning the sample window size based on a cumulative feature importance distribution was proposed.

Due to its inherent simplicity aligned with a high performance, Random Forest was the most cost-effective solution for system design, achieving one of the highest AUC performances, similar to SVM, the best performing method. Concerning class-imbalance, the process of oversampling with SMOTE, targeting to result in classes of equal size, significantly improved the classification efficiency. Finally, the new process of tuning the sample window length allowed a reduction in the window size by a factor of 7, without significantly compromising the model performance.

In future works, we intend to implement the system proposed in this paper in some low-cost embedded device. In addition, we plan to exploit Transfer Learning strategies and tiny Deep Learning models. Finally, we consider evaluating the proposed methodology with other ECG datasets.

## REFERENCES

[1] Brasil. “Diretrizes clínicas para o cuidado ao paciente com doença renal crônica - DRC no Sistema Único de Saúde”. 2014.

[2] F. de Souza Terra, A. M. D. D. Costa, E. T. Figueiredo, A. M. de Morais, M. D. Costa and R. D. Costa. “As principais complicações apresentadas pelos pacientes renais crônicos durante as sessões de hemodiálise”. *Revista da Sociedade Brasileira de Clínica Médica*, vol. 8, no. 3, pp. 87, 2010.

[3] M. C. M. de Castro. “Atualização em diálise: complicações agudas em hemodiálise”. *Brazilian Journal of Nephrology*, vol. 23, no. 2, pp. 108–13, 2001.

[4] J. I. Guimarães, J. C. Nicolau, C. A. Polanczyk, C. A. Pastore, J. A. Pinho and et. al. “Diretriz de interpretação de eletrocardiograma de repouso”. *Arquivos Brasileiros de Cardiologia*, vol. 80, pp. 1–18, 2003.

[5] C. J. Grupi, F. S. de Brito and A. H. Uchida. “Eletrocardiograma de Longa Duração: o Sistema Holter - Parte II”. *Journal of Cardiac Arrhythmias*, vol. 12, no. 3, pp. 134–146, 1999.

[6] E. Jayachandran, P. Joseph K, R. Acharya U *et al.*. “Analysis of myocardial infarction using discrete wavelet transform”. *Journal of medical systems*, vol. 34, no. 6, pp. 985–992, 2010.

[7] J. A. Gutiérrez-Gnecchi, R. Morfin-Magaña, D. Lorias-Espinoza, A. del Carmen Tellez-Anguiano, E. Reyes-Archundia, A. Méndez-Patiño and R. Castañeda-Miranda. “DSP-based arrhythmia classification using wavelet transform and probabilistic neural network”. *Biomedical Signal Processing and Control*, vol. 32, pp. 44–56, 2017.

[8] N. K. Dewangan and S. P. Shukla. “ECG arrhythmia classification using discrete wavelet transform and artificial neural network”. In *2016 IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)*, pp. 1892–1896. IEEE, 2016.

[9] R. J. Martis, U. R. Acharya, A. K. Ray and C. Chakraborty. “Application of higher order cumulants to ECG signals for the cardiac health diagnosis”. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1697–1700. IEEE, 2011.

[10] Y. Ozbay and B. Karlik. “A recognition of ECG arrhythmias using artificial neural networks”. In *2001 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pp. 1680–1683. IEEE, 2001.

[11] S. S. Xu, M.-W. Mak and C.-C. Cheung. “Towards end-to-end ECG classification with raw signal extraction and deep neural networks”. *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1574–1584, 2018.

[12] S. Faziludeen and P. V. Sabiq. “ECG beat classification using wavelets and SVM”. In *2013 IEEE Conference on Information & Communication Technologies*, pp. 815–818. IEEE, 2013.

[13] N. Kohli, N. K. Verma and A. Roy. “SVM based methods for arrhythmia classification in ECG”. In *2010 International Conference on Computer and Communication Technology (ICCCT)*, pp. 486–490. IEEE, 2010.

[14] G. Pan, Z. Xin, S. Shi and D. Jin. “Arrhythmia classification based on wavelet transformation and random forests”. *Multimedia Tools and Applications*, vol. 77, no. 17, pp. 21905–21922, 2018.

[15] N. Emanet. “ECG beat classification by using discrete wavelet transform and Random Forest algorithm”. In *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, pp. 1–4. IEEE, 2009.

[16] J.-S. Wang, W.-C. Chiang, Y.-L. Hsu and Y.-T. C.Yangb. “ECG arrhythmia classification using a probabilistic neural network with a feature reduction method”. *Neurocomputing*, vol. 116, pp. 38–45, 2013.

[17] M. K. Sarkaleh and A. Shahbahrami. “Classification of ECG arrhythmias using discrete wavelet transform and neural networks”. *International Journal of Computer Science, Engineering and Applications*, vol. 2, no. 1, pp. 1, 2012.

[18] X. Dong, C. Wang and W. Si. “ECG beat classification via deterministic learning”. *Neurocomputing*, vol. 240, pp. 1–12, 2017.

[19] G. Sannino and G. D. Pietro. “A deep learning approach for ECG-based heartbeat classification for arrhythmia detection”. *Future Generation Computer Systems*, vol. 86, pp. 446–455, 2018.

[20] N. Mahmood, A. M. H. A. Giasuddin, K. A. Jhuma, M. M. Hoque and S. S. Chowdhury. “Patterns of Cardiac Arrhythmia in Haemodialysis Patients”. *Anwer Khan Modern Medical College Journal*, vol. 7, no. 1, pp. 28–33, 2016.

[21] P. E. McSharry, G. D. Clifford, L. Tarassenko and L. A. Smith. “A dynamical model for generating synthetic electrocardiogram signals”. *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289–294, 2003.

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

[23] M. Llamedo and J. P. Martínez. “Heartbeat classification using feature selection driven by database generalization criteria”. *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 616–625, 2010.

[24] P. de 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.

[25] J. Wiens and J. Guttag. “Active learning applied to patient-adaptive heartbeat classification”. volume 23, pp. 2442–2450, 2010.

[26] J. Pan and W. J. Tompkins. “A real-time QRS detection algorithm”. *IEEE Transactions on Biomedical Engineering*, , no. 3, pp. 230–236, 1985.

[27] D. Conen, M. Adam, F. Roche, J.-C. Barthelemy, D. F. Dietrich, M. Imboden, N. Künzli, A. von Eckardstein, S. Regenass, T. Hornemann, T. Rochat, J.-M. Gaspoz, N. Probst-Hensch and D. Carballo. “Premature atrial contractions in the general population: frequency and risk factors”. *Circulation*, vol. 126, no. 19, pp. 2302–2308, 2012.

[28] X. Liu, H. Du, G. Wang, S. Zhou and H. Zhang. “Automatic diagnosis of premature ventricular contraction based on Lyapunov exponents and LVQ neural network”. *Computer Methods and Programs in Biomedicine*, vol. 122, no. 1, pp. 47–55, 2015.

[29] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng and H. E. Stanley. “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals”. *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[30] T. Hastie, R. Tibshirani and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[31] T. Hastie, S. Rosset, J. Zhu and H. Zou. “Multi-class adaboost”. *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[32] L. Breiman. “Random forests”. *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf. “Support vector machines”. *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[34] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. *Annals of statistics*, pp. 1189–1232, 2001.

[35] S. O. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall, Boston, 1999.

[36] N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[37] H. He and Y. Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[38] M. Friedman. “The use of ranks to avoid the assumption of normality implicit in the analysis of variance”. *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[39] L. Martin, R. Leblanc and N. K. Toan. “Tables for the Friedman rank test”. *Canadian journal of statistics*, vol. 21, no. 1, pp. 39–43, 1993.

[40] P. B. Nemenyi. “Distribution-free multiple comparisons”. Ph.D. thesis, Princeton University, 1963.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique”. *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[42] D. F. Williamson, R. A. Parker and J. S. Kendrick. “The box plot: a simple visual method to interpret data”. *Annals of internal medicine*, vol. 110, no. 11, pp. 916–921, 1989.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] S. Spanulescu. *Esp32 Programming for the Internet of Things*. Lulu Press, Inc, 2018.

[45] W. Gay. *Raspberry Pi hardware reference*. Apress, 2014.

[46] O. T. Inan, L. Giovangrandi and G. T. A. Kovacs. “Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features”. *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2507–2515, 2006.