Data Extraction Method Combined with Machine Learning Techniques for the Detection of Premature Ventricular Contractions in Real-Time

L.C. Sodré[1], B.G. Dutra [1], A.S. Silveira[1], I.M. Mizara[1]

[1]Universidade Federal do Pará, Faculdade de Engenharia Elétrica e Biomédica, Instituto de Tecnologia, Belém, Brasil.

Abstract— **Currently, diagnostics in the medical field are being automated. Thus, reducing errors of interpretation in diagnoses. This article proposes a recognition method to identify premature ventricular contraction in real time, soon enabling the minimization of damages resulting from arrhythmia. The proposed method consists of two main modules: data extraction module by way of Recursive Least Squares (RLS), guaranteeing data extraction in real time, and the classification module, its inputs being the parameters from the RLS algorithm. In the resource extraction module, autoregressive modeling (AR) is used to extract characteristics. In the classifier module, Support Vector Machine and Multilayer Perceptron are examined. The classifiers' performance was assessed by standard metrics. The proposed algorithm showed high precision and few false negatives.**

Keywords— **Classification, Real Time, Arrhythmia, Electrocardiogram, Recursive Least Squares.**

I. INTRODUCTION

Premature Ventricular Contraction (PVC) is a premature heartbeat that occurs before the expected time for the next systole [1] and originates in the ventricular region below the bifurcation of the His bundle [2]. The PVCs are the most common arrhythmias and occur not only in heart disease cases but also in healthy people [3]. These arrhythmias can be identified on the 12-lead electrocardiogram (ECG) using a signal with an enlarged QRS, with a duration time higher than 0.12s and with branch block morphology. However, there are other premature beats, originating before the bifurcation of the trunk of the bundle of His, which have a QRS with a duration timeless than 0.12s, and which are classified as PVC [4].

The PVCs are observed since approximately 600 BC, by ancient Chinese medicine, and reported in the study by physician Pien Chio [4]. According to Chio, the pulse with intermittent failure was not an impeding condition for the individual to have a healthy life. However, if there was a failure every ten beats, it could be a symptom of cardiac pathology. Also, it is advisable that any PVC that was documented on a routine electrocardiogram (ECG), especially those interpreted as “R over T”, are potential signs of short-term mortality [5]. Therefore, it is known that PVCs are directly related to various diseases such as dilated cardiomyopathy, ventricular hypertrophy, coronary diseases, Brugada syndrome, and a premature ventricular contraction that can trigger malignant arrhythmias; therefore, it is necessary to monitoring [4].

The ECG is one of the fastest and most accurate ways to diagnose cardiac arrhythmias. Although many cardiologists interpret the signal with high precision, it is known that errors can occur. Also, the specialist (usually a cardiologist) performed the analysis, which can be quite personal, causing some variations in the diagnosis among these professionals [6].

In view of this, the academic community has endeavored to develop ways to automate the classification of heartbeats using elaborate techniques, but the proposed applications have tended to advanced pre-processing, signal segmentation and high-level classifiers [6,7,8,9], thereby making the real-time application unfeasible. the hospital environment needs it. This article aims to use a data extraction method combined with machine learning techniques for the detection of PVCs in real-time, enabling monitoring and effective intervention, if necessary.

II. METHODOLOGY

For the development of this study, ECG signals from patients with arrhythmia from the MIT-BIH database were used. However, for better data quality, the signals were initially pre-processed to remove noise from the signals. Then, the real-time data are extracted from the autoregressive model that has the parameters calculated by the Recursive Least Squares (RLS) method. The RLS is used because it offers parameters needed for online classification, uses little memory and can see the dynamics change in the model.

Subsequently, based on the parameters calculated by the MQR algorithm, the Support Vector Machines (SVM) and Artificial Neural Network (RNA) classifiers in the Multilayer Perceptron (MLP) architecture performed the heartbeat labeling. The classification performance analysis was based on the confusion matrix, Receiver Operation Characteristic Curve (ROC curve), and the Area Under the ROC Curve (AUC), thus enabling the comparison between this study and other articles in the literature that use different techniques for detecting arrhythmia.

A. Data

The ECG signals used in this article were obtained from the MIT-BIH arrhythmia database available online and for free. The database was registered by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979. It contains 48 records obtained from 47 different patients and has two leads for the upper and lower ECG signals for all records. Records last for about 30 minutes with a sampling frequency of 360Hz and include two leads: the adapted lead of member II and one of the modified leads V1, V2, V4, or V5. At least two cardiologists labeled heartbeat [9].

Thus, the signal intervals of patients 106,114, 205 and 221 of the databases were used in this article because they had PVCs of different morphologies and frequencies. The data were divided by 90% for training and 10% for validation and, as the focus is the online detection of PVCs and Normal Beat, other types of arrhythmias present were removed.

B. Pre-Processing

From the information in the MIT-BIH database, it was possible to visualize, in most of the signals, unfavorable artfacts in the signal, for example low and high-frequency noises and the baseline drift.

Thus, in order to exclude such noises, models of analog low-pass and high-pass filters were used, both transformed into the discrete domain by way of the Tustin Transform, enabling digital signal processing in real-time, thus maintaining the characteristics of the standard ECG signal and correcting, in part, the flaws in signal acquisition.

C. Data Extraction

Data is extracted from the filtered signals using autoregressive models that have parameters optimized by the RLS identification algorithm. The transfer function of the AR model contains only one constant in the numerator and one polynomial in the denominator. The AR model in the discrete domain is represented in Equation 1.

$$y(k)=\sum_{t=1}^{na} a(t)y(k-t)+u(k) \qquad (1)$$

Where a(t) represents the model parameters and na is the model order [10]. In this study, the portion *u(k)*, which can represent both noise and input from the system, was discarded.

The identification by way of RLS allows the estimated model of the system to be perfected at each sampling period, realizing the parametric variations in the process [11]. Therefore, it is important to view the recursive estimation procedure in terms of a parallel model, as shown in Figure 1, where *y* is the output of the system, θ is the vector of parameters of the real system and θ_{est} is the vector of parameters estimated by the RLS algorithm, therefore, each sample generates a new error signal, *e(k)*, which is the output signal minus the signal estimated by the parameters, y_{est}. Thus, the error signal of each sample goes through the MQR adaptation mechanism, resulting in the optimization of the parameters of the estimated model.

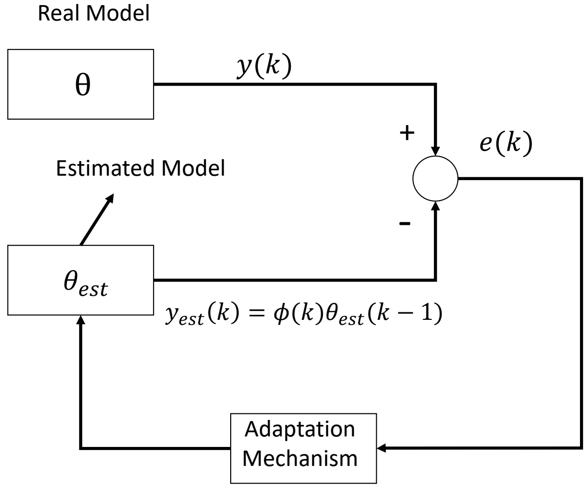

Fig. 1 MQR diagram used

The parameters are updated by the gain vector of the estimator RLS, *L*, the values present in this vector are calculated according to Equation 2, where the forgetfulness factor is represented λ, with the function of giving more importance to new acquisitions to the detriment of old ones, making the estimated model more sensitive to new samples, and *P* is the covariance matrix. In addition, important elements such as the vector of regressors, ϕ, that are the AR model variables (old output values).

$$L(k)=\frac{(P(k-1)\phi(k)}{(\lambda+\phi^T(k)P(k-1)\phi(k))} \qquad (2)$$

With the updated *L* vector gains, the estimated parameters can assume values that most represent the actual functioning of the process, as shown in Equation 3.

$$\theta_{est}(k)=\theta_{est}(k-1)+L(k)e(k) \qquad (3)$$

Finally, the Covariance Matrix is used for the next sampling period following the estimation process optimized for each acquisition. The update of matrix *P* is represented in Equation 4.

$$P(k) = \frac{([I_{na \times na} - L\phi^T(k)]P(k-1))}{\lambda} \quad (4)$$

Thus, every 0.002778 seconds (patient sampling time), new parameters are formed. The parameters calculated for the second order model, the system output (ECG value) and the determinant of the matrix P obtained in that sample will be the 4 inputs of the classifiers that enabled the online classification.

D. Classifiers

In this study, two classifiers were used, Support Vector Machines (SVM) and Artificial Neural Network (ANN).

Support Vector Machines is a supervised machine learning method and the most popular classifier for diagnosing arrhythmias on ECGs [12], with different types of kernel functions (linear, radial base, or polynomial functions) [13]. The SVM is based on the statistical learning theory developed by Vapnik to obtain good classifiers [14]. It was defined for problems of two classes and extended to several classes' problems, solving several of them in topologies one against one or one against all. Two classes (A and B) can be distinguished mathematically by solving Equation 4 and labeling positive results as A and negative results as B.

$$F(x) = \omega \cdot h(x) + b \quad (5)$$

Where x is the input vector, *h* is a function that allows the classification nonlinearly with different kernel types, ω is the classifier weight vector and *F(x)* is the predicted output. The training phase aims to find weights for the function that maximizes the distance of each class. In this article, he used the binary detection format, as it is a classification between two standards, the regular beats and the beats derived from PVCs.

The use of ANN is increasing in the task of classifying and grouping arrhythmias since they have a high performance in this task [6]. ANNs are computational techniques that present a model based on intelligent organisms' neural structure and acquire knowledge through experience [15]. They have useful applications where the rules for solving the problem are hardly known or difficult to formalize, and when it is necessary to solve the problem quickly [16]. There are several application domains, such as shape recognition, signal processing, vision, speech, forecasting, modeling, decision aid, robotics, and extraction of rules [17].

The neuron is the functioning unit of a neural network that processes each value of its inputs and produces an output related to a transfer function called an activation function. A neuron is connected to other neurons by synapses, in which each synapse has a related weight. In this article, Multilayer Perceptron (MLP) was used, which is a neural network with one or more hidden layers with an undetermined number of neurons. The hidden layer has this name because it is not possible to predict the desired output in the intermediate layers. To training the MLP network, the commonly used algorithm is the backpropagation (Backpropagation).The MPL used has a hidden layer and 20 neurons. This configuration presented the best result for the data set.

The input of the classifiers are the two parameters of the AR model by way of RLS, the system output (ECG value) and the determinant of the matrix P. In possession of these inputs, SVM and MPL classified the heartbeat as normal or PVC during its occurrence.

E. Performance Appraisers

After training and validating both classifiers, MPL and SVM, necessary data are created to calculate their accuracy. However, other indexes are needed to evaluate each classifier's performance, then ROC curve and Confusion matrix were used.

The confusion matrix analyzes the classification performance using four items: the true positive (TP), when the classifier labels positive and gets it right; the real negative (TN), when it labels negative and gets it right; the false positive (FP), when it classifies positive and makes mistakes; finally the false negative (FN) when it labels as not belonging to the class and misses.

The ROC graph is based on the rate of true positives (TPR), represented in Equation 5, and the false-positive rate (FPR), represented in Equation 6. For the construction of the ROC graph, FPR is plotted in the ordinate axis (x-axis) and TFR on the abscissa axis (y-axis) [18]. With the ROC Curve, it is possible to calculate a scalar index called AUC, which is integral, so the closer to the value 1, the better and the classification, thus giving more objectivity for the classifier's analysis. One way of interpreting AUC is as the probability that the model will rank a random positive example higher than a random negative example.

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

III. Results

In this study, two classifiers were designed to function in real-time, which label two types of a heartbeat: regular beat and premature ventricular contraction. The classifiers use the

RLS algorithm parameters to label each sample, the ECG signal resulting from the filtering present in the pre-processing, and thus, allowing the classification of the beat when it occurs.

After obtaining the Performance Appraisers, it was possible to create the confusion matrix present in Table 1 and the precision and AUC index present in Table 2 of both classifiers.

Table 1 Matrix confusion in percentage

	MPL		SVM	
	N	PVC	N	PVC
N	93.2%	15.9%	99.3%	3.5%
PVC	6.8%	84.1%	0.7%	96.5%

Table 2 AUC and accuracy values

	ROC Area	Accuracy
MPL	0.9176	91.59%
SVM	0.9816	98.86%

For better interpretation, the results were drawn to the ROC curve, represented in Figure 2, to analyze graphically the performance of the classifiers for both classes proposed in the article.

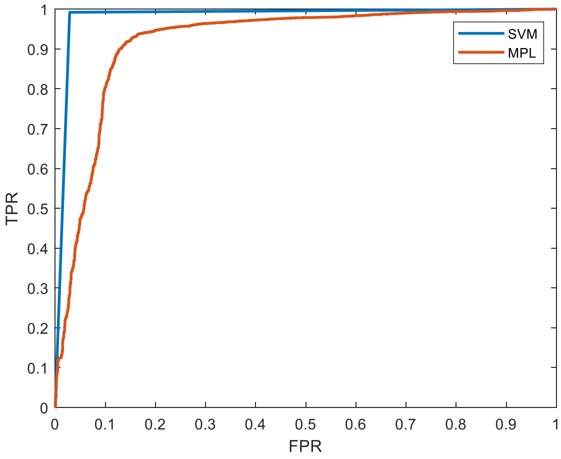

Fig. 2 ROC Curve

The Table 3 shows other studies obtained in the literature on the classification of arrhythmia. The table also shows the data extraction method used by the authors, the classifier with the best results and the accuracy of the best classifier.

Table 3 Accuracy comparison

Works	Data Extraction	Classifiers	Accuracy
Han et. al. (2010) [19]	Time Analysis	ARTMAP	96,90%
Lin et. al. (2006) [20]	Time Analysis	Clustering	90%
Rogal et. al. (2009) [6]	Wavelet	ART2	93,42%
Fernandes et. al. (2010) [7]	Heart Rate Variability	PPM-C	99,71%
Chazal et. al. (2003) [8]	ECG-Intervals	Weighted LD	89%
Alickovic et. al. (2015) [9]	AR by way of Burg method	SVM	99.93%
In this work	AR by way of RLS	SVM	98.86%

IV. Discussion

When compared, the SVM showed better results than the MPL neural network, because the indices point to better accuracy in general and the detection of the two classes (regular and PVCs). However, both techniques were successful and were able to label both classes. SVM and MPL obtained good values in the confusion matrix and consequently, in the ROC curve for both classes. However, the regular beat got more hits than the PVCs because it follows a pattern, while the PVCs are more diversified in morphology and frequency. Also, by analyzing the ROC curves, the SVM classification tends to avoid having false negatives for PVCs, more than the neural network. This point in the medical diagnostic is essential because the damage to the patient's health is more harmful when there is a false negative in the tests, as the patient no longer has the necessary medical monitoring [13]; finally, the value of the scalar AUC for SVM is 0.9816, and for MPL it is 0.9176, thus revealing the best performance of SVM about MPL.

Table 3 shows the results found in the literature. Most of the studies in Table 3 use advanced pre-processing, signal segmentation and complex data extraction methods, thus obtaining great results and a greater variety of arrhythmia classes [6,7,8,9], however, it prevents the application of the classification in real time due to the possible high computational cost of extraction methods and the classification does not use data that is obtained sample by sample of the ECG. In view of this, the proposed method uses basic preprocessing combined with a method of extracting data in real time with low computational cost, which is why RLS is widely used in industry [11], allowing the identification of the arrhythmia at the moment it occurs, thus obtaining one of the best accuracy of the works shown in Table 3.

V. Conclusion

In this study, a new automated categorization algorithm has been developed for ECG heartbeat classification that aims to be applied in hospitals, due to the combination of filtering and data extraction techniques that do not require much computational work, both procedural and in memory, and that allows the real-time labeling of the heartbeat. The two classifiers (MPL and SVM) were used to compare. Thus, the SVM presented the best results both for correctness and classification between classes obtaining the accuracy of 98.86%.

Finally, future work in this area is to use other classes of arrhythmia for identification in real time and use this method of extraction and online classification for other human signals such as electroencephalogram and electromyogram.

Interest conflicts

The authors declare that they have no conflict of interest.

Reference

1. Friemann A, (2016) Revista Diagnóstico e Tratamento, Vol. 21, Ed. 2. São Paulo.
2. Pastore CA, Pinho C; et al (2009) Diretrizes da Sociedade Brasileira de Cardiologia sobre Ánalise e Emissão de Laudos, Arq. Bras. Cardiol. vol.93 no.3 supl.2 São Paulo.
3. Grupi CJ, Lima M (2008) Extrassístoles: apresentação e classificação. In: Pastore CA, Grupi J, Moffa PJ, editores. Eletrocardiologia atual. 2a ed. São Paulo: Atheneu; p. 261-72.
4. Pimenta J, Valente N (2015) Como Conduzir Pacientes Assintomáticos com Extrassístoles Ventriculares? Sociedade de Cardiologia do Estado de São Paulo.
5. Lown B, Fakhro AM; et al (1967) The coronary care unit. New perspectives and directions. JAMA; 199(3):156-66.
6. Rogal SR, Neto A, (2009) Automatic Detection of Arrhythmias Using Wavelets and Self-Organized Artificial Neural Networks. Ninth International Conference on Intelligent Systems Design and Applications. 10.1109 / ISDA.2009.22
7. Medeiros F, Cavalcante A (2010). Classificação de Sinais Eletrocardiográficos através do Algoritmo de Compressão PPM. CBIS 2010.
8. Chazal P, Reily R, et al (2003) Automatic classification of ECG beats using waveform shape and heartbeat interval features. IEEE International Conference on Acoustics, Speech, and Signal Processing. 10.1109 / ICASSP.2003.1202346.
9. Alickovic E, Subasi A (2015), Effect of Multiscale PCA De-noising in ECG Beat Classification for Diagnosis of Cardiovascular Diseases. Circuits Systems and Signal Process. 10.1007/s00034-014-9864-8.
10. Semmlow JL (2004), Biosignal and Biomedical Image Processing: Matlab-Based Applications..
11. Coelho A, Coelho L (2016) Identificação de sistemas dinâmicos lineares.
12. Park KS, Cho BH, et al (2008). Hierarchical support vector machine-based heartbeat classification using higher order statistics and hermite basis function. 10.1109 / CIC.2008.4749019.
13. Moura KOA., Favieiro, GW, et al. (2016) Support vectors machine classification of surface electromyography for non-invasive naturally controlled hand prostheses. 10.1109 / EMBC.2016.7590819.
14. Vapnik, VN (1999) An overview of statistical learning theory. IEEE Transactions on Neural Networks.10.1109/ 72.788640.
15. .Krose B, Smagt PVD (1996) An Introduction Neural Networks.
16. Mitchell, T (1997) Machine Learning, McGraw-Hill.
17. Steiner, MTA, Soma NY, et al (2006) Using Neural Networks Rule Extraction for Credit-Risk Evaluation. International Journal Of Computer Science And Network Security, v. 6, n. 5A, p. 6-17.
18. Prati RC, Batista GE, Monard MC (2008) Curvas ROC para avaliação de classificadores. IEEE Latin America Transactions.
19. Ham F, Ham S (2010) Classification of Cardiac Arrhythmias Using Fuzzy ARTMAP. IEEE Transactions on Biomedical Engineering Volume: 43 , Issue: 4. 10.1109/10.486263.
20. Lin Z, Ge Y, Tao G (2006) Algorithm for Clustering Analysis of ECG Data. IEEE Engineering in Medicine and Biology 27th Annual Conference. 10.1109/IEMBS.2005.1615302.

Institute: Universidade Federal do Pará.
Street: Rua Augusto Corrêa.
City: Belém.
Country: Brazil.
Email: sodre209333@gmail.com