Revista Gestão & Tecnologia
e-ISSN: 2177-6652
revistagt@fpl.edu.br
http://revistagt.fpl.edu.br/

AUTOMATED DETECTION OF ANOMALIES IN ELECTROCARDIOGRAMS USING EMPIRICAL MODE DECOMPOSITION

DETECÇÃO AUTOMATIZADA DE ANOMALIAS EM ELETROCARDIOGRAMAS USANDO DECOMPOSIÇÃO EM MODO EMPÍRICO

DETECCIÓN AUTOMATIZADA DE ANOMALÍAS EN ELECTROCARDIOGRAMAS MEDIANTE DESCOMPOSICIÓN EN MODO EMPÍRICO

Hygor Santiago
Engenheiro mecânico formado pela UFSJ
hsantiagolara@gmail.com
http://orcid.org/0000-0002-4835-5498

Milton Dias
Pós-Doutorado no Structural Dynamics Research Laboratory, da Universidade de Cincinnati, EUA. Professor Associado I da Universidade Estadual de Campinas
milton@unicamp.br

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 16.02.2021
Aprovado em 14.03.2022

ABSTRACT

Study objective: Develop an algorithm for detection and classification of heart arrhythmia in electrocardiograms.

Methodology/approach: Different ways of using the collected data are discussed, starting from the simplest one, which is the beat counter, to the more complex ones, where the complete signals present in an electrocardiogram is used. Different Machine Learning techniques were also used: K-Nearest Neighbors, Logistic Regression, Support Vector Machines and Extra Trees. The beat counter approach considers the time difference between each cardiac cycle and can be collected by a simple smart watch or an oximeter. For the complete classification of the anomalies, two other signal processing techniques were considered: the Fourier Transform and the Empirical Mode Decomposition.

Originality/Relevance: It is the first paper to use the Empirical Mode Decomposition combined with Machine Learning techniques for the classification and detection of cardiac anomalies.

Main results: The beat counter is not efficient enough to distinguish between all classes of existing anomalies, even among those studied in this work, but it presents good results for binary distinction between normal and abnormal heart beat. For the complete classification of the anomalies, the Empirical Mode Decomposition presented the best results. It is even better than the time-frequency analysis technique used in other papers on electrocardiogram classification.

Theoretical/methodological contributions: This paper presents a new application for Empirical Mode Decomposition and how it can be combined with classification techniques.

Keywords: Electrocardiograph, Empirical Mode Decomposition, Machine Learning, Classification, Cardiac Anomaly Detector, Arrhythmia Detection.

RESUMO

Objetivo do estudo: Desenvolver um algoritmo para detecção e classificação de arritmias cardíacas em eletrocardiogramas.

Metodologia/abordagem: São discutidas diferentes formas de utilização dos dados coletados, desde a mais simples, que é o contador de batimentos, até as mais complexas, onde são utilizados os sinais completos presentes em um eletrocardiograma. Diferentes técnicas de *Machine Learning* também foram utilizadas: *K-Nearest Neighbors*, Regressão Logística, Máquinas de Vetores de Suporte e *Extra Trees*. A abordagem do contador de batimentos considera a diferença de tempo entre cada ciclo cardíaco e pode ser coletada por um simples *smart watch* ou um oxímetro. Para a classificação completa das anomalias, foram consideradas duas outras técnicas de processamento de sinais: a Transformada de Fourier e a Decomposição Empírica por Modos.

Originalidade/Relevância: É o primeiro artigo a utilizar a Decomposição Empírica por Modos combinada com técnicas de *Machine Learning* para classificação e detecção de anomalias cardíacas.

Principais resultados: O contador de batimentos não é eficiente o suficiente para distinguir entre todas as classes de anomalias existentes, mesmo entre as estudadas neste trabalho, mas apresenta bons resultados para distinção binária entre batimentos cardíacos normais e anormais. Para a classificação completa das anomalias, a Decomposição Empírica por Modos

apresentou os melhores resultados. É ainda melhor do que a técnica de análise tempo-frequência usada em outros trabalhos sobre classificação de eletrocardiograma.
Contribuições teórico-metodológicas: Este artigo apresenta uma nova aplicação para Decomposição Empírica por Modos e como ela pode ser combinada com técnicas de classificação.

Palavras-chave: Eletrocardiógrafo, Decomposição Empírica por Modos, Aprendizado de Máquina, Classificação, Detector de Anomalias Cardíacas, Detecção de Arritmias.

RESUMEN
Objetivo del estudio: Desarrollar un algoritmo para la detección y clasificación de arritmias cardíacas en electrocardiogramas.

Metodología/enfoque: Se discuten diferentes formas de utilizar los datos recopilados, desde la más simple, que es el contador de latidos, hasta las más complejas, donde se utilizan las señales completas presentes en un electrocardiograma. También se utilizaron diferentes técnicas de *Machine Learning*: *K-Nearest Neighbors*, *Logistic Regression*, *Support Vector Machines* y *Extra Trees*. El enfoque del contador de latidos considera la diferencia de tiempo entre cada ciclo cardíaco y puede recopilarse con un simple reloj inteligente o un oxímetro. Para la clasificación completa de las anomalías se consideraron otras dos técnicas de procesamiento de señales: la Transformada de Fourier y la Descomposición de Modo Empírico.

Originalidad/Relevancia: Es el primer artículo que utiliza la Descomposición de Modo Empírico combinada con técnicas de Aprendizaje Automático para la clasificación y detección de anomalías cardíacas.

Principales resultados: El contador de latidos no es lo suficientemente eficiente para distinguir entre todas las clases de anomalías existentes, incluso entre las estudiadas en este trabajo, pero presenta buenos resultados para la distinción binaria entre latidos cardíacos normales y anormales. Para la clasificación completa de las anomalías, la Descomposición de Modo Empírico presentó los mejores resultados. Es incluso mejor que la técnica de análisis de tiempo-frecuencia comúnmente utilizada en otros artículos sobre clasificación de electrocardiogramas.

Aportes teórico-metodológicos: Este artículo presenta una nueva aplicación para la Descomposición de Modo Empírico y cómo se puede combinar con técnicas de clasificación.

Palabras clave: Electrocardiógrafo, descomposición modal empírica, aprendizaje automático, clasificación, detector de anomalías cardíacas, detección de arritmias.

1. INTRODUCTION

The circulatory system is responsible for supplying the organs with oxygen and eliminating the produced carbon dioxide. This is
done by the blood circulating throughout the body making this very important exchange. It needs to travel to the most distant cells at the ends of the body and return to the lungs where breathing replaces carbon dioxide with oxygen. In this cycle, the main component is the heart, which is a driving pump of the entire circulatory system.

Cardiovascular diseases have been the leading cause of mortality since the 1960s, accounting for a substantial burden of disease in Brazil, Rezende, *et al* (2016). Chronic non-communicable diseases constitute the main group of causes of death worldwide, being responsible for premature deaths, loss of quality of life, in addition to adverse economic and social impacts. These diseases are responsible for about 70% of deaths, exceeding more than 38 million deaths per year and significantly outnumbering deaths from external causes and infectious diseases, according to the World Health Organization. Of the approximately 45% of deaths from non-communicable chronic diseases in the world, more than 17 million are caused by cardiovascular diseases. The same happens in Brazil, where 72% of deaths result from non-communicable chronic diseases, of which 30% are due to cardiovascular diseases, 16% are due to neoplasms and 6% are related to respiratory diseases, Purcell, et al (2020).

In Brazil, cardiovascular diseases were responsible for most of the direct expenses, for substantial costs with hospitalization and indirect costs due to reduced productivity or absence from work, Allen, *et at* (2017). Cardiovascular diseases and their complications resulted in an expenditure of US$ 4.18 billion in the Brazilian economy between 2006 and 2015, according to the SUS (Public Health System). As for clinical hospitalizations, heart failure led admissions, with 2,862,739 hospitalizations (131 per 100,000 inhabitants), with 1,149,602 interventional cardiovascular surgical procedures. In addition to the financial impact, it is essential to emphasize the human cost of these diseases, since surgical procedures are very invasive. Many complications and deaths can be avoided with proper treatment, but this depends on the diagnosis still at an early stage, which is a challenge specially in the poorest regions of Brazil.

Brazil is among the richest countries in the world, but also among the most discrepant. In large urban centers, health care has many and diverse professionals, but in other regions, the lack of professionals to serve the population is extremely worrying. In São Paulo, 24.8%

of deaths are caused by cardiovascular diseases and in Minas Gerais, 25.4%. In Amazonas, 36.6% die due to heart problems and in Pará, 35.1%, according to SUS. The discrepancy between the indices is due to the lack of human and financial resources for health in the poorest regions of Brazil. It is extremely difficult to find professionals to work in these remote regions, further harming the service to the population. In this way, it is important to develop new technologies that help health professionals, allowing to maintain the quality of care but with a smaller contingent of personnel. Another increasingly widespread practice is telemedicine, which allows initial care at a distance and refers the patient to the most appropriate professional, when necessary.

In this sense, this work seeks to develop an algorithm for the detection of anomalous heart diseases, enabling a quick and accurate diagnosis. This algorithm can be easily implemented in software that can be used on portable equipment by professionals not specialized in patient triage for later referral to the appropriate professional. It can also be also be very helpful in telemedicine, facilitating the diagnosis, sending and storing of information.

The development of the methodology used 3 different signal processing techniques - Beat Counter (BC), Fast Fourier Transform (FFT) and Empirical Decomposition by Modes (EMD) - and 4 Machine Learning techniques - K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machines (SVM), and Extra Trees (ET). Signal processing techniques extract inputs from machine learning models that are compared to find the best combination of accuracy and sensitivity. The models are tested on the classification of electrocardiograms with normal sinus rhythm and 16 different anomalies.

2. RESEARCH BACKGROUND

Electrocardiograms have a lot of measurement noise and anatomical variations and involuntary patient movements can affect the analyses. The measurement system itself is very complex and, moreover, the magnitude to be measured is very small and it is often lower than the noise. In this way, the use of filters and signal amplifiers is very important. However, some level of noise is always found in the signal provided by the equipment. The problem can be solved through mathematical transformations, eliminating inherent errors and producing

clear signals that are better analyzed by human experts or automated systems, as in Das, *et al* (2013).

Plawiak (2018) worked with the Fourier Transform. The author initially calculated the transform for each sample and, using Genetic Algorithm, selected the most significant frequencies for classification. Other works validate the use of this technique, such as Singhal, *et al* (2020) when using Fourier signal decomposition.

Works have been developed in pattern recognition in electrocardiography. This was done by Jain, *et al* (2021) through adaptive filters with results around 99% correct. At each occurrence of a QRS complex occurs a heart systole and determines the end and the beginning of a circulatory cycle. These works present efficient ways to determine the heart rate and the prediction of certain anomalies. Such methods cannot be considered fully autonomous as they cannot diagnose all possibilities of anomalies, but they are very efficient in filtering data to be stored in a Holter.

Yazdani, *et al* (2016) were able, through Wavelet transforms, to accurately identify not only the QRS complex, but also the P and T waves with 99% accuracy. In Zidelmal, *et al* (2013) the same methodology was applied to eliminate noise and identify the electrocardiogram waves. This information fed a Machine Learning system for anomaly prediction. In the comparison between Artificial Neural Networks and Support Vector Machine, the former had 94.4% and the latter, 98.9% of success.

Anomalies can be found in the electrocardiographic signal. Martins, *et al* (2013) diagnosed occurrences of Atrial Fibrillation in electrocardiograms using the Naive Bayes method, with 99.33% accuracy, and with the Recurrent Neural Networks, with 99.77% accuracy, but with a significantly higher computational cost. Acharya, *et al* (2017) did the same using Logistic Regression and Naive Bayes, with 99.5% accuracy.

Differentiating between different types of anomalies is a significantly more complex problem than those described so far. Analyzing the results of several electrocardiograms Acharya, *et al* (2017) were able to identify four distinct groups: normal, Atrial Fibrillation, Atrial Flutter and Ventricular Fibrillation. Using Convolutional Neural Networks, the authors obtained, respectively, 98.40%, 95.32%, 97.37% and 98.70% of occurrences of these events. These are very interesting results that already demonstrate the feasibility of Machine Learning and Deep Learning for the identification and classification of anomalies in electrocardiograms.

Martins, *et al* (2013) diagnosed Left and Right Branch Blocks, Extra Atrial and Ventricular Systole with a system based on Support Vector Machine and Artificial Neural Networks. It is important to note that the author used principal component analysis to reduce the dimensionality of the inputs. With the Artificial Neural Networks, a hit rate of 92.77% was obtained, while with the Support Vector Machines, the hit rate was 93.48%.

The work by Plawiak (2018) is one of the most complete in the diagnosis of anomalies in electrocardiograms. The authors were able to classify ECG signals into 17 groups, the normal one and 16 anomalies. The authors compared the following methods: Support Vector Machines, Artificial Neural Networks, K-Nearest Neighbors and Probabilistic Neural Networks. Genetic Algorithm was also used to determine the best inputs for the system in order to avoid data redundancy or confusion in the algorithm.

3. THE ELECTROCARDIOGRAM - ECG

An ExG is a measure of the electrical potential between two points on the human body. When the heart is located between these points it is called an ECG, electrocardiogram.

There are 12-lead combinations of measurement points that provide information about the electrical polarization of the heart in different planes.

All leads are very important and they are all recorded by the electrocardiograph machine and used by the physician to identify many cardiac anomalies such as spots of cardiac injury or areas with abnormal behavior.

A cardiac arrhythmia is an irregularity or abnormality that occurs in the heartbeat and can be characterized by the variation of the time between the beats and/or by the change in the waveform of the electrical polarization, measured by the electrocardiogram.

In principle, any lead can be used to identify arrhythmias, as this dysfunction does not depend on the area of the heart, it occurs in the organ as a whole. The L2 lead is a voltage measurement performed between the electrodes positioned on the patient's right shoulder and left leg. In this work, only this lead is used, once it provides a cross-sectional view of the heart in this plane.

The heart is a four-chamber volumetric pump. The two upper chambers are called atria and the lower chambers are called ventricles, and they perform similar and simultaneous

functions with each other. The right side of the heart pumps blood to the lungs (small circulation) and the left side pumps blood to the rest of the body (large circulation). The cycle begins with the entry of blood through the atria. At this moment, the electrical polarization of this chamber occurs. In the electrocardiogram (L2 lead) a broad wave of low amplitude, called P wave, is observed, as is shown in Figure 1, which presents a simulated ECG signal.

When the atria are completely filled, the sinus node (the electrical center of the heart) releases a small electrical charge causing the muscles to contract, increasing pressure. At this point, the mitral and tricuspid valves are released allowing blood to flow into the ventricles. By rapidly entering the ventricles, the sinus nodes provide an even greater load so that more energy is delivered to the blood. It is an abrupt and very strong discharge, resulting in a narrow, high-amplitude peak on the ECG. This whole process is very fast and the waves mix in the electrocardiogram. This process is called the QRS complex on the ECG and is illustrated in Figure 1.

The last step of the cardiac cycle is the polarization of the ventricle. The signal is again a broad wave of low amplitude called T wave. This wake can also be seen in Figure 1.

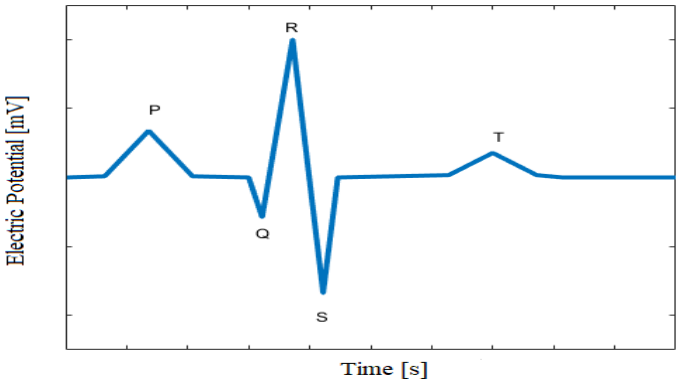

Figure 1. Simulation of an ideal ECG signal.
Source: Author.

4. MACHINE LEARNING METHODS

In the previous section, the cardiac cycle was described and the general and expected form of the ECG signal was presented. However, cardiac arrhythmias promote changes in the

morphology of these signs. To identify whether the ECG signal represents normal or abnormal behavior of the heart, machine learning methods are used, and these methods are described below.

4.1. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is based on the distance between the points relative to each data in hyperspace, Mailagaha Kumbure, *et al* (2020). Each data is considered as a point in hyperspace. As points, it can geometrically calculate the distance between them, thus considering the closest ones as similar. When a new data is presented to the system, the distance between all the training data the closest ones influence the classification. The prediction is given as the classification of nearest neighbors.

Such system is extremely simple computationally speaking and very fast. This distance can be measured in different ways such as Euclidean, Manhattan, Minkowski... A major factor in the quality of the algorithm is the number of neighbors (K) considered, few neighbors errors in training data and how many neighbors can leave the system fragile gross in its diagnosis, Ma and Li (2021).

4.2. Logistic Regression

Logistic Regression is a generalization of linear regression that allows the classification into three parts: a) prediction function, b) cost function and 3) optimization of the function of prediction, Fan *et al. (2020)*. Initially the Logistic Regression behaves like a linear regression where it tries to calculate a line in hyperspace that produces the best answer for any input data. This straight line is the prediction function, which is given by:

$$R = \sum_{i=1}^{n} (a_i . x_i) + b, \quad (1)$$

where *a* is a constant that multiplies the relative *x* input and then all these parcels are added to a constant *b*, *a* and *b* are initially random, Josephus *et al*., (2020). This is the concept of establishing a line in hyperspace where *R* has the desired value for any input, the sigmoid function normalizes *R* between 0 and 1. So the system tries to make the binary classification

between groups, for multiclass cases the concept is expanded to as many sigmodal functions as problem classes.*a* and *b* being initially random, Josephus et al., (2020). This is the concept of establishing a line in hyperspace, where *R* has the desired value for any input, and using the sigmoid function to normalize *R* between 0 and 1. So the system tries to do the binary classification between groups. For multiclass cases, the concept is expanded to as many sigmoidal functions as problem classes. The sigmoid function is given by:

$$S = \frac{1}{1+e^{-R}}, \quad (2)$$

The value *S* is the output of the algorithm for each input of the interaction and, after being found, it is compared to the known value of the answer *y*. This comparison constitutes the cost function *J*:

$$J = \sum_{i=1}^{m} \quad (3)$$

It can also add regularization weights with Ridge or Lasso to the cost function Xiao *et al*. (2021), this acts in a way to privilege lower values of *a* and *b* avoiding large spatiality which sometimes makes the system more efficient. The Lasso is the sum of coefficients *a* and *b* multiplied by an l that promotes the weight given to regularization. The Ridge, is similar, but each coefficient is high squared. As presented in this work, the two parts act together, starting to configure the ElasticNet regularization.

$$J' = J + l_1 . \sum_{i=1}^{n} (a| \quad |i + b) + l_2 . \sum_{i=1}^{n} (a_i + b)^2 \quad (4)$$

In the other interactions, the algorithm tries to reduce this cost function by finding the optimal parameters for *a* and *b* of the prediction function. This optimization is done by the descending gradient. Finally, the ideal values of *a* and *b* are found and the algorithm manages distinguish the classes with the best possible hit rate.

4.3. Support Vector Machines

Support Vector Machines use vector calculus to separate the different classes through a hyperplane, and two positioned hyperplanes equidistant and parallel to the main, being as close as possible to the data (for binary classification), Nakagawa *et al*. (2021). In theory this method is very similar to regression linear, but the form of calculation is quite different. A vector orthogonal to the hyperplane is taken. as reference and vectors orthogonal to parallel hyperplanes called support vectors, all inputs are taken as vectors where each variable is a component.

The classification is based on the projection of the vector corresponding to the input onto the vector orthogonal to the hyperplane, if the modulus is greater than a certain value, it belongs to one class, if it is smaller, to another. The method tries to find the separation hyperplane and the support hyperplanes always seeking to maximize the distance between the support vectors. The method used for the optimization is the Lagrange multiplier, Jiang *et al*. (2017).

4.4. Extra Trees

Unlike the previously mentioned techniques, Extra Trees does not use the descending gradient for its optimization. The best coefficients are adjusted by Tree of Decision. Other techniques are also optimized by this method such as Decision Tree and Random Forest, but Extra Trees considers greater randomness in its parameterization, making it more robust than the others, in certain cases.

The Decision Tree algorithm consists of separating data according to some characteristic where the information gain is maximum, Chen *et al*. (2021). This method can be compared to the divide-and-conquer philosophy, by separating data into groups to sorting can be done much faster and with the possibility of using less data than other algorithms, Panhalkar and Doye (2021). This separation is done in a row time until the data is satisfactorily sorted.

Extra Trees is a more sophisticated application of the decision tree algorithm. In this case, an ideal decision tree model will not be built, but several models will be built of simpler decision tree and at the end, Mohana *et al*. (2021). The decision is taken by the majority of votes. In addition, Extra Trees starts from a randomly selected database within the training data, each decision tree is assembled from this smaller group of data. It can be seen as a

randomly created forest with several decision tree individuals where the final model is less affected by erroneous data because such data are part of most individuals.

This is an ensemble method where not only one ideal model is created, but several limited models constitute the final answer, Makungwe *et al*. (2021). This type of model is more robust than previously seen as it drastically reduces overfitting. Selecting randomly the database into groups, a decision tree is created to better classify a new data from the input data that make up each group, generating so many decision trees how many groups limited to their inputs. Also, the trees are simpler because in the creation of the node all possibilities are tested within a small group of random variables by the method of gain of information which creates many trees of shape fast, but individually bad. When entering a new data in the system, it will be tested in each tree and thus many of the same or different classifications are presented. The answer most recurrent "wins" the vote and is the system's response. Extra Trees is very efficient, surpassing other well-known methods in many applications, Chen *et al.* (2021).

5. THE MIT-BHI DATABASE

The database used in this work has 1000 signals, acquired with a recording time of 10s and a sampling frequency of 360Hz. The signals, collected by MIT-BHI and publicly available for studies, have 17 classes of L2 lead electrocardiogram signals as described below:

- Normal sinus rhythm (283 samples),
- Premature atrial beat (66 samples),
- Atrial flutter (20 samples),
- Atrial fibrillation (135 samples),
- Supraventricular tachycardia (13 samples),
- Wolff-Parkinson-White (211 samples),
- Ventricular extra systole (133 samples),
- Ventricular extra systole bigeminy (55 samples),
- Ventricular extra systole trigeminy (13 samples),
- Ventricular tachycardia (10 samples),
- Idioventricular flutter (10 samples),
- Ventricular flutter (10 samples),
- fusion between ventricular and normal beat (11 samples),
- Left branch block (103 samples),
- Right branch block (62 samples),

- Blocking (10 samples) and
- Pacemaker (45 samples).

6. SIGNAL PROCESSING

6.1. Beat Counter

The heart pace should be constant and between 60 and 100 beats per minute. In some arrhythmia the heart rate is altered, losing its constancy and/or ideal pace. The first entries for the Machine Learning algorithms in this work refer to cardiac rhythm. Figure 2 shows the algorithm for the beat counter.

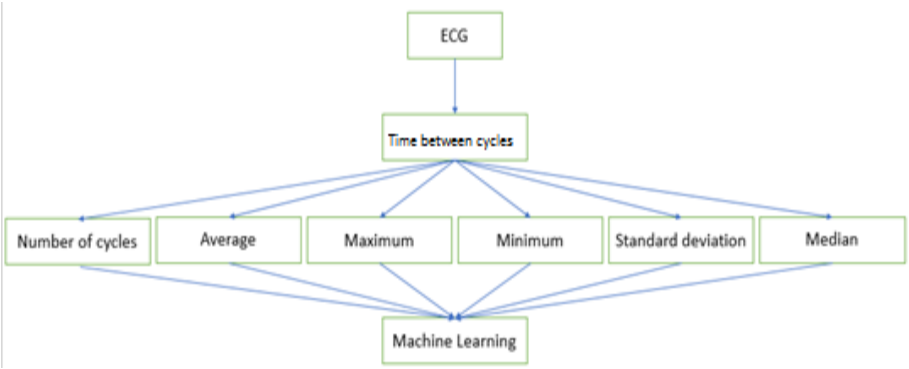

Figure 2. Algorithm for classification of ECG using Beat Counter.
Source: Author.

The time interval between each QRS complex is taken and all the following information is extracted: mean value, median, variance, standard deviation, maximum and minimum values. This easily extractable information carris very important information for the classification of cardiac anomalies. This analysis does not necessarily need to be done by the electrocardiogram, as it just counts the heartbeats. An oximeter or a smart watch could do this measurement, as the shape of the electrical waves is completely ignored.

6.2. Fast Fourier Transform

A well-known method of extracting information from the signal is the Fast Fourier Transform (FFT). This transformation allows study the signal in the frequency domain. This

technique is commonly used in papers about electrocardiographic signals. As in Das, *et al* (2013), Plawiak (2018), Singhal, *et al* (2020), Martins, *et al* (2013) and Acharya, *et al* (2017).

By transforming the ECG from the time to frequency domain, the resulting waveform expresses important information about the morphology of the electrocardiographic signal. For this reason, the Fourier Transform was used as input to the Machine Learning methods.

The electrocardiographic signal has its most important components at low frequencies. In the range below 1 Hz there is predominantly information about voluntary or non-voluntary muscle movements. At frequencies above 100 Hz no significant events associated with heart behavior are observed. Thus, in this work, the signals were all filtered in the range from 1 to 100 Hz.

6.3. Empirical Mode Decomposition

The Empirical Mode Decomposition (EMD) method proposes the decomposition of non-stationary signals, derived from linear or non-linear systems, in linear series. Each decomposition is called mode and must be independent, called Intrinsic Mode Functions (IMF). All the IMF added together result in the original signal, which can reconstruct the signal, although certain modes only provide system noise information. In this way, the EMD calculation is a way to eliminate measurement errors and random signals inherent to the system, allowing a better analysis of the signals. The nature of EMD can produce wobbles with very different scales in one mode, or wobbles with similar scales in different modes. When this phenomenon is undesirable, and the scales are similar for each mode, the problem of mixing the modes occurs.

Figure 3 shows the decomposition of a signal its IMF. In the first frame the full signal is displayed with red markings for local highs and blue markings for the minimums. In the second, these points are joined and, considering the interpolation, their behavior is well defined throughout the domain. Then, the average of both is taken and the line formed can be seen in black in the third frame. It is observed how such a line can predict the general behavior of the signal, being little affected by noise or high frequency components of the original signal. With this information, the first IMF is extracted and the result can be seen in the fourth frame. Depending on the application of the EMD, the result after the extraction of the first IMF still presents certain trend and contains relevant information, so the process is

repeated until the result is just noise. The process can be interrupted when the resultant becomes monotonous, that is, when no more behavioral tendency can be obtained. In this case, the function has only fluctuations around zero and its IMF would then be a straight line with only image at the null value.

Figure 3 shows the decomposition of a signal its IMF. In the first frame the full signal is displayed with red markings for local highs and blue markings for the minimums. In the second, these points are joined and, considering the interpolation, their behavior is well defined throughout the domain. Then, the average of both is taken and the line formed can be seen in black in the third frame. It is observed how such a line can predict the general behavior of the signal, being little affected by noise or high frequency components of the original signal. With this information, the first IMF is extracted and the result can be seen in the fourth frame. Depending on the application of the EMD, the result after the extraction of the first IMF still presents certain trend and contains relevant information, so the process is repeated until the result is just noise. The process can be interrupted when the resultant becomes monotonous, that is, when no more behavioral tendency can be obtained. In this case, the function has only fluctuations around zero and its IMF would then be a straight line with only image at the null value.

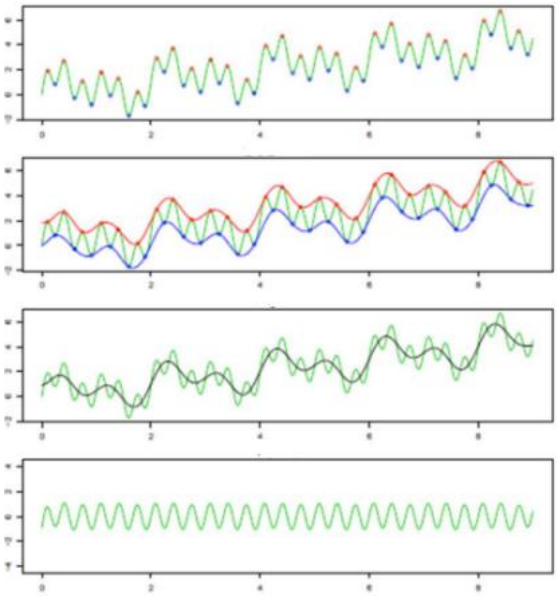

Figure 3. Intrinsic Mode Functions for a time series.
Source: Silva (2019)

In order to avoid the mixing of modes, the Ensemble Empirical Mode Decomposition (EEMD) method was proposed, where noise is inserted in the measurement before decomposition. Such methodology allows the comparison of the result with different noise characteristics. Different signals are generated composed by the sum of the original signal with several random noises, and different from each other. At the end, for each signal, the IMF is calculated, and the final response resulting from the EEMD is formed by the average of the value found for each iteration. In this way, the process is more robust and if any specific noise causes a significant change in the system response, it will be mitigated in the average calculation.

The EEMD process is similar to the EMD, where each IMF is obtained from the signal, but as already mentioned, in the former method a random noise is inserted into the signal before processing. The process occurs as many times as necessary where different noises are inserted with each sample. In the end, each sample contains its IMF and the algorithm's final answer is the average of these. EEMD achieves more regular modes by avoiding large differences in time scale like the simpler EMD. Despite being useful, EEMD generates problems such as residual noise in signal reconstruction. The addition of adaptive noise by the CEEMDAN method showed an important improvement in the EEMD, reaching negligible reconstruction errors in relation to the previous methodology. The difference between them is in the addition of noise, while in EEMD it is added to the original signal and all the IMF obtained, in CEEMDAN it is considered new noise for each IMF. In the end, several IMF are obtained at each iteration and the average of these is considered. In the CEEMDAN methodology the computational cost becomes considerably higher, which can be a big problem when dealing with large amounts of data.

The CEEMDAN method is even more random once more noise is introduced at different stages of the process. In the first step the noise is inserted in a similar way to the EEMD, the difference occurs in the second step when the IMF is obtained. Only the first IMF in the system is calculated, so new noise is added and recalculated. The process generates several values for each IMF and the solution is the average of these.

Working with the entire vector coming from each IMF is computationally infeasible. As it is a well-defined time series, it can be summarized using an autoregressive (AR) model, where only the coefficients can describe the signal. In this way, the IMF is expressed by its

regressive coefficients, which are the inputs to Machine Learning models. Figure 4 represents such algorithm.

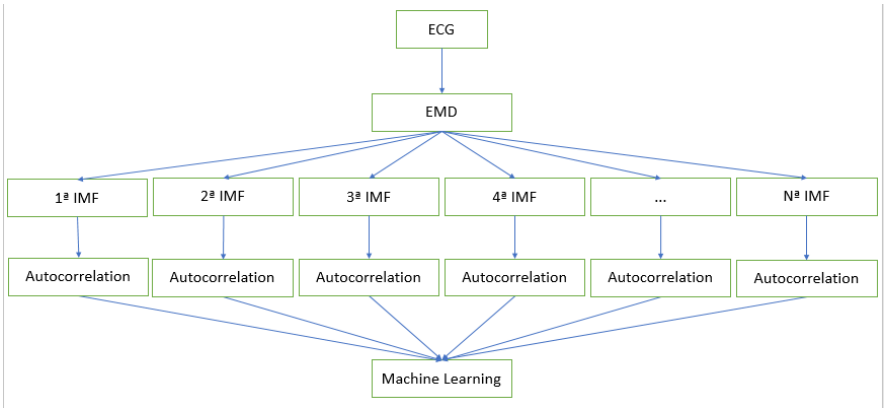

Figure 4. Algorithm for classification of ECG using EMD.
Source: Author

7. RESULTS AND DISCUSSIONS

7.1. Beat Counter

Figure 5 shows the accuracy and sensitivity results of the models for detecting anomalies using the Beat Counter. With this signal processing technique, it was not possible to precisely identify the anomaly, but only to identify whether the heart rhythm was normal or not. This analysis is very important given the ease of acquiring signals for the Beat Counter, which can be used in simpler equipment.

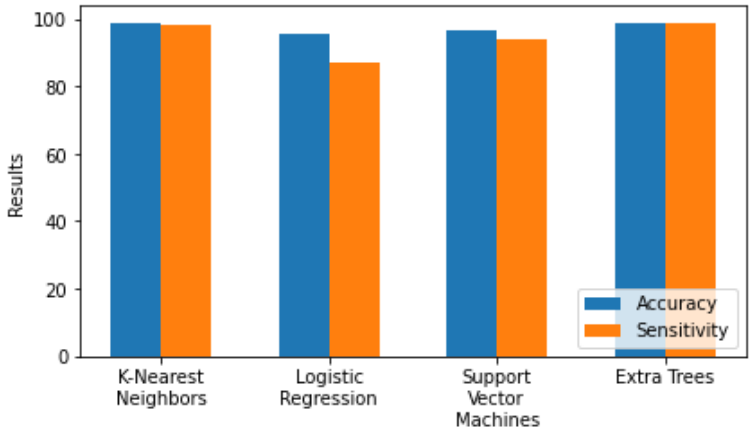

Figure 5. Beat Counter results for diagnose the abnormal cardiac rhythm.
Source: Author.

The Extra Trees method presented the best results for diagnose the abnormal cardiac rhythm, with 98.9% accuracy and 98.8% sensitivity. K-Nearest Neighbors also showed good results, but even so, slightly inferior. These results are very good and proves that a simple equipment, like a smart watch or an oximeter, can contain warning mechanisms recommending that the patient should seek medical assistance. For a detailed diagnosis, however, it is necessary to use more complete techniques, which will be addressed in the next sections of this work

7.2. Fast Fourier Transform

The results obtained using the Fourier technique proved to be unsatisfactory for this application. Previously cited works performed better with the use of this technique, but used fewer classes in their classifications. The best algorithm was KNN with 93,4% accuracy and 87,4% sensitivity. The results are in Figure 6.

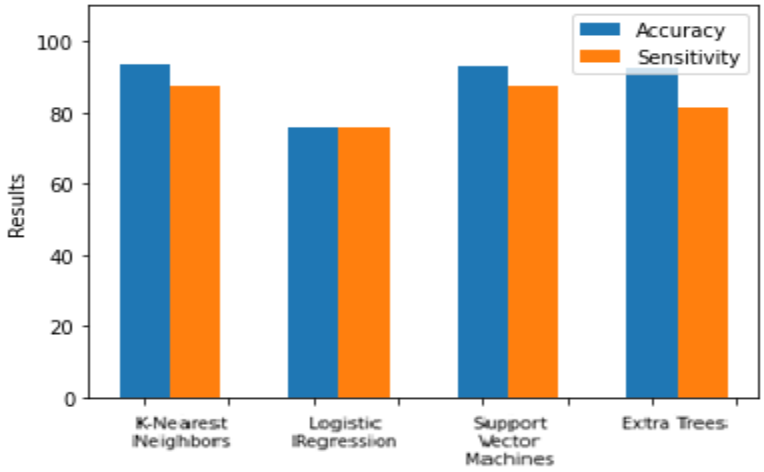

Figure 6. Results using Fourier Transform.
Source: Author.

7.3. Empirical Mode Decomposition

The studied signals were calculated in 14 IMF. All were ordered where the first ones correspond to the high frequency components and the last ones to the low ones. Electrocardiographic signals are always low frequency, so the last 6 IMF do not provide useful information. Similarly, the first 3 IMF provide only noise and baseline loss information. The models were tested in order to prove which IMF would provide the best results, which can be seen in Figure 6, for the Logistic Regression model. Eliminating the first and last IMF is an important form of filtering and significantly impacts the model.

The studied signals were all decomposed in 14 IMF. These IMF were ordered so that the first corresponded to the lowest frequency components and the last to the high frequency ones. Electrocardiographic signals are predominantly low-frequency, so the last 6 IMF did not provide useful information. Likewise, the first 3 IMF only provided information on noise and baseline loss. The models were tested to prove which set of IMF would provide the best results. The result of this analysis, in which Logistic Regression was used, can be seen in Figure 7. As it can be seen, eliminating the first and last IMF is an important form of filtering and significantly impacts the accuracy of the method.

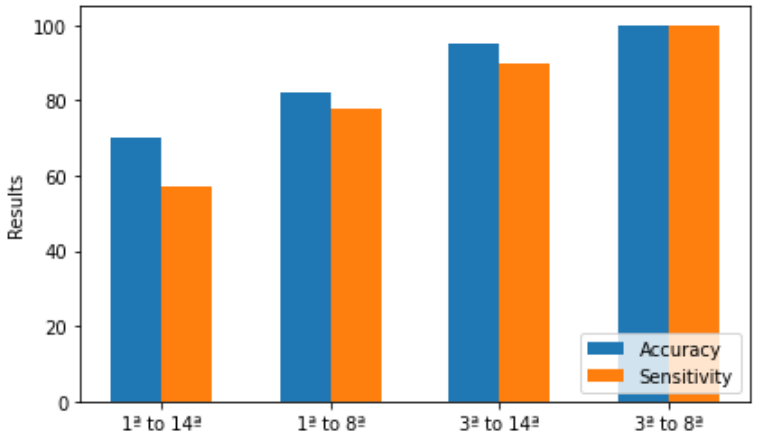

Figure 7. Results using different IMF.
Source: Author.

Another important factor when studying EMD is the order of the IMF autoregressive model. As already mentioned, each IMF is a well-behaved time series that can be described by an autoregressive model. The order of this model determines the accuracy of this description. Very low orders are not able to faithfully represent the IMF, while very high orders incorporate a lot of errors and noise in the representation. To determine the best order of the AR model, different values were tested. The results of this analysis are shown in Figure 8. It is observed that the order 8 model presented the best results.

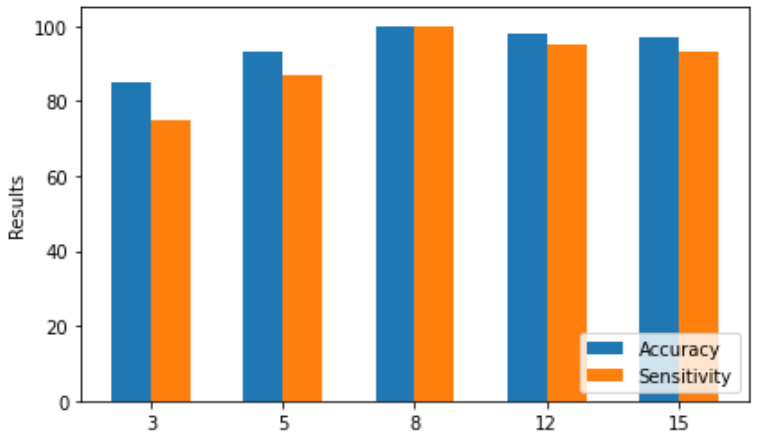

Figure 8. Results for different IMF AR model order.
Source: Author.

The method that presented the best results using EMD was the Logistic Regression. Based on the results presented in Figures 7 and 8, in the processing only the IMF from 3rd to

8th and AR model of 8th order were considered. This method obtained the best possible result, with 100% accuracy and sensitivity, hitting all tests.

7.4. Techniques Comparison

Models based on Beat Counter, FFT and EMD presented very different results for the complete classification of the studied anomalies. The Beat Counter did not provide good results, being unfeasible for this type of analysis. FFT achieved good accuracy, especially when used with the K-Nearest Neighbors and Extra Trees methods. EMD, on the other hand, provided very good results, with emphasis on the Logistic Regression, in which it obtained 100% accuracy. Only with K-Nearest Neighbors the FFT was a better technique than EMD, proving the importance of comparing methods using different inputs. Extra Trees proved to be the most robust technique in terms of varying signal processing techniques, with very similar results for the three modeling approaches. These conclusions can be easily obtained by analyzing Figure 9, which shows the precision presented by the various machine learning methods for the three types of input data.

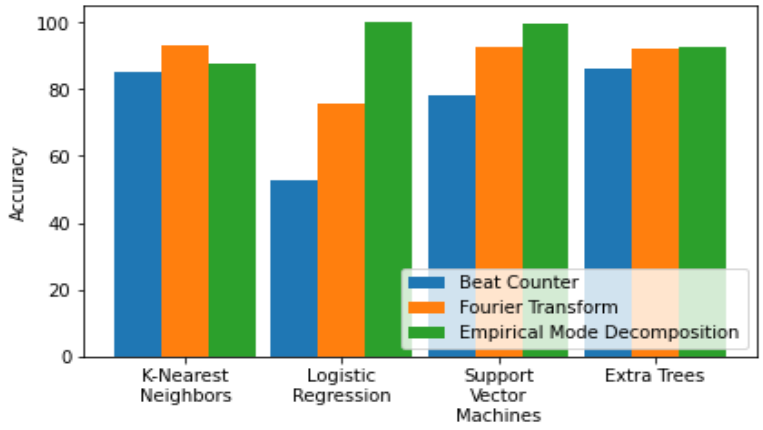

Figure 9. Accuracy of the models comparing each signal processing and Machine Learning technique

Source: Author.

Another important metric is processing time. The Beat Counter was extremely fast for training, which proves its simplicity. For the FFT calculation the processing was fast, however many inputs are provided to the models which makes them slow. This is evidenced by the difference in the processing time of models using the same data. The opposite is true

with EMD. Signal processing is much more costly, but training is extremely fast. Calculating the IMF requires a high computational cost, since a lot of noise must be inserted into the system by the CEEMDAN methodology. However, the linear autocorrelation provides the coefficients that are the few inputs of the Machine Learning models.

Another important metric for comparing the methods is processing time. The Beat Counter was extremely fast for training, which proves its simplicity. For the calculation of the FFT the processing was fast, however many inputs are provided to the models which makes them slow. This is evidenced by the difference in the processing time of the models using the same data. The opposite is true with EMD, where signal processing is much more expensive, but training is extremely fast. The calculation of the IMF requires a high computational cost, since a lot of noise must be inserted in the system by the CEEMDAN methodology. However, in this case, only the coefficients of the AR model are given as inputs to the Machine Learning methods. Figure 10 shows the processing time spent by the various machine learning methods for the three types of input data.

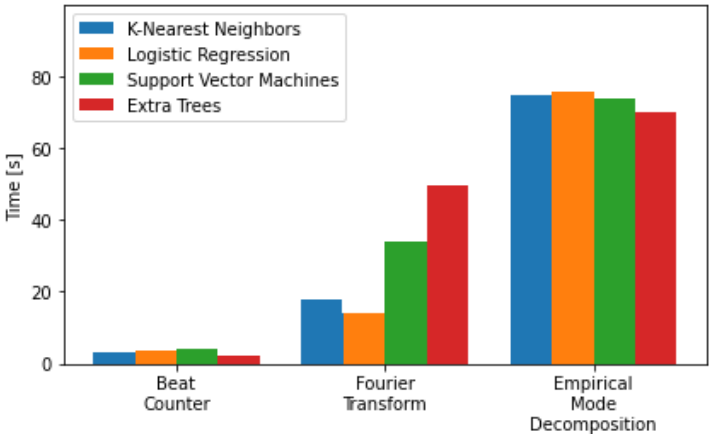

Figure 10. Processing time of each model.
Source: Author.

8. CONCLUSIONS

As in other works, the use of the Fourier Transform allowed a good classification among the arrhythmias present in the ECG. However, the use of the Empirical Decomposition by Modes technique allowed achieving the perfect result of 100% accuracy. Thus, it can be stated that EMD is the best for classifying signals with characteristics similar to those of the

electrocardiogram. This work presents an important contribution to the area, achieving the best possible result for this database.

The automation of medical reports is a very important tool for the early diagnosis of diseases. In cardiology, equipment with such capacity can diagnose different types of cardiac arrhythmias. Whether just for an alert recommending that the patient be referred to a doctor or for the professional to have information for the treatment.

Of the three signal processing methodologies studied, the Beat Counter is the simplest and easiest to measure. The model with this technique was not able to separate all 16 arrhythmias studied. However, it was very effective in diagnosing the simple presence of anomalies. This technique can be implemented in devices such as oximeters or smart watches, allowing the individual to quickly look for a trained professional. This technique was unable to distinguish which arrhythmia was present in the test, which other techniques did efficiently.

As in other works, the use of the Fourier Transform allowed a good classification among the arrhythmias present in the ECG. However, the use of the Empirical Mode Decomposition technique allowed achieving the perfect result of 100% accuracy. Thus, it can be said that the EMD is the best input to classify signals with characteristics similar to those of the electrocardiogram. This work presents an important contribution to the area, achieving the best possible result for the studied database.

REFERENCES

Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H. & Adam, M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences*, v. 405, p. 81–90, https://doi.org/10.1016/j.ins.2017.04.012.

Allen, C., Andrei, C. L., Carrero, J. J. & Goulart, A. C. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, v. 390, n. 10100, p. 1151–1210, https://doi.org/10.1016/S0140-6736(17)32152-9.

Chen, Y., Zheng, W., Li, W. & Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters*, v. 144, p. 1–5, Doi: 10.1016/j.patrec.2021.01.008.

Das, M. & Ari, S. (2013). Analysis of ECG signal denoising method based on s-transform. *IRBM*, v. 34, n. 6, p. 362–370, Doi: 10.1016/j.irbm.2013.07.012.

Ecobar, F. B. (2019). Eletrocardiograma e anormalidades cardíacas. [S.l.: s.n.].

Fan, Y., Bai, J., Lei, X., Zhang, Y., Zhang, B., Li, K.-C. & Tan, G. (2020). Privacy preserving based logistic regression on big data. *Journal of Network and Computer Applications*, v. 171, p. 102769, https://doi.org/10.1016/j.jnca.2020.102769.

Jain, S., Ahirwal, M., Kumar, A., Bajaj, V. & Singh, G. (2017). QRS detection using adaptive filters: A comparative study. ISA Transactions, v. 66, p. 362–375, Doi: 10.1016/j.isatra.2016.09.023.

Jiang, Y., Wang, X.-G., Zou, Z.-J. & Yang, Z.-L. (2021). Identification of coupled response models for ship steering and roll motion using support vector machines. *Applied Ocean Research,* v. 110, p. 102607, https://doi.org/10.1016/j.apor.2021.102607.

Josephus, B. O., Nawir, A. H., Wijaya, E., Moniaga, J. V. & Ohyver, M. (2020). Predict mortality in patients infected with covid-19 virus based on observed characteristics of the patient using logistic regression. *Procedia Computer Science*, v. 179, p. 871–877, 2021, 5th International Conference on Computer Science and Computational Intelligence, Doi: 10.1016/j.procs.2021.01.076.

Li, Q., Rajagopalan, C. & Clifford, G. D. A machine learning approach to multi-level ECG signal quality classification. *Computer Methods and Programs in Biomedicine*, v. 117, n. 3, p. 435–447, 2014.

Ma, H. & Li, J. (2021). A sub-linear time algorithm for approximating k-nearest-neighbor with full quality guarantee. *Theoretical Computer Science*, v. 857, p. 59–70, Doi: 10.1016/j.cmpb.2014.09.002.

Mailagaha Kumbure, M., Luukka, P. & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters*, v. 140, p. 172–178, Doi: 10.1016/j.patrec.2020.10.005.

Makungwe, M., Chabala, L. M., Chishala, B. H. & Lark, R. M. (2021). Performance of linear mixed models and random forests for spatial prediction of soil ph. *Geoderma,* v. 397, p. 115079,

Doi: 10.1016/j.geoderma.2021.115079.

Martis, R. J., Acharya, U., Prasad, H., Chua, C. K. & Lim, C. M. Automated detection of atrial fibrillation using Bayesian paradigm. *Knowledge-Based Systems,* v. 54, p. 269–275.

Pławiak, P. (2018). Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. *Expert Systems with Applications*, v. 92, p. 334–349, Doi: 10.1016/j.eswa.2017.09.022.

Mohana, R. M., Reddy, C. K. K., Anisha, P. & Murthy, B. R. (2021). Random Forest algorithms for the classification of tree-based ensemble. *Materials Today: Proceedings*, Doi: 10.3390/ma14185342.

Nakagawa, S., Hochin, T., Nomiya, H., Nakanishi, H. & Shoji, M. (2021). Prediction of unusual plasma discharge by using support vector machine. *Fusion Engineering and Design,* v. 167, p. 112360.

Panhalkar, A. R. & Doye, D. D. (2021). Optimization of decision trees using modified African buffalo algorithm. *Journal of King Saud University - Computer and Information Sciences*, Doi: 10.1016/j.jksuci.2021.01.011.

Purcell, C. A., Alvis-Guzman, N., Bensenor, I. M., Carvalho, F., Castro, F., Fernandes, J. C., Fernandes, E. & Freitas, (2020). M. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, v. 395, n. 10225, p. 709–733, Doi: 10.1016/S0140-6736(20)30045-3.

Rezende, L. F. M. de, Azeredo, C. M., Canella, D. S., Carmo Luiz, O. do, Levy, R. B. & Eluf-Neto, J. (2016). Coronary heart disease mortality, cardiovascular disease mortality and all-cause mortality attributable to dietary intake over 20 years in Brazil. International *Journal of Cardiology*, v. 217, p. 64–68, Doi: 10.1016/j.ijcard.2016.04.176.

Singhal, A., Singh, P., Fatimah, B. & Pachori, R. B. (2020). An efficient removal of power-line interference and baseline wander from ECG signals by employing Fourier decomposition technique. *Biomedical Signal Processing and Control*, v. 57, p. 101741, Doi: 10.1016/j.bspc.2019.101741.

Silva Souza, J. da. (2019). Análise de atributos de classificação para o diagnóstico de falhas em rolamentos baseados em SVM.

Xiao, R., Cui, X., Qiao, H., Zheng, X., Zhang, Y., Zhang, C. & Liu, X. (2021). Early diagnosis model of Alzheimer's disease based on sparse logistic regression with the generalized elastic net. Biomedical Signal *Processing and Control*, v. 66, p. 102362, Doi: 10.1016/j.bspc.2020.102362.

Yazdani, S. & Vesin, J.-M. (2016). Extraction of QRS fiducial points from the ECG using adaptive mathematical morphology. *Digital Signal Processing*, v. 56, p. 100–109.

Zidelmal, Z., Amirou, A., Ould-Abdeslam, D. & Merckle, J. (2013). ECG beat classification using a cost sensitive classifier. *Computer Methods and Programs in Biomedicine*, v. 111, n. 3, p. 570–577, Doi: 10.1016/j.cmpb.2013.05.011.