A cardiac arrhythmia monitoring platform based on feature selection and classification methods

Anderson P. N. Silva[1], **Gibeon S. Aquino-Júnior**[1],
João C. Xavier-Júnior[2], **Cephas A. S. Barreto**[2]

[1]Informatics and Applied Mathematics Department
Federal University of Rio Grande do Norte
Natal, RN, Brazil

[2]Digital Metrolopis Institute
Federal University of Rio Grande do Norte
Natal, RN, Brazil

{anderson,gibeon}@dimap.ufrn.br,jcxavier@imd.ufrn.br,cephasax@gmail.com

Abstract. *Heart arrhythmia, also known as irregular heartbeat, affects millions of people around the world. One of the ways to detect this cardiac dysrhythmia is by performing an electrocardiogram (ECG) exam which records the electrical activity of the heart. However, this type of exam is always interpreted by a doctor. In order to provide an alternative in heart arrhythmia diagnosis, this paper aims at developing a platform based on Internet of Things infrastructure capable of automatically monitoring and identifying cardiac arrhythmia based on feature selection and classification methods.*

1. Introduction

Considering estimations made by the United Nations (UN) that the Earth will have just over 9 billion inhabitants by 2050, concerns with diseases have become an increasingly alarming topic worldwide. According to the Brazilian Institute of Supplementary Health Studies, in a study carried out with the assistance of the Medical School of the Federal University of Minas Gerais, approximately 829 Brazilians die daily in public and private hospitals due to heart malfunctioning.

To reduce the number of deaths caused by diseases, health professionals have invested resources in tools which can perform diagnoses, whether preventive or emergency ones. This anticipation in the discovery of some diseases such as heart dysfunctions can determine medical treatments, and even prevent patients' sudden deaths. On top of that, special type of algorithms have been used consistently in order to provide immediate benefits to disciplines with reproducible or standardized processes.

Machine learning techniques have been used since the beginning in the healthcare field to identify disease patterns [Ilayaraja and Meyyappan 2013]. Information Technology companies have already begun to develop Machine Learning applications that can remotely track the employees' health or monitor the health of older people. Moreover, many studies have also been focused on monitoring different diseases, such as: high blood pressure, diabetes, and others.

Differently from other works, this work has the main purpose of developing a cardiac arrhythmia monitoring platform based on Internet of Things (IoT) infrastruc-

ture, which enables monitoring, identifying and notifying health professionals, patients and family members in real-time based on feature selection techniques and classification methods. Moreover, two well-known public arrhythmia datasets were used for training and testing the platform, and also three different feature selection methods were used to better understand the particularities (more relevant attributes) of such heart malfunctioning. On top of that, was performed empirical analyses was with both base classifiers (Decision tree, Naive Bayes, Multilayer Perceptron, k-Nearest Neighbor and Support Vector Machine), and ensembles of classifiers (AdaBoost and Random Forest). In addition, the results of the experiments indicate that the use of ensembles of classifiers, and feature selection method improves system response time, and helps to better understand the most relevant attributes in the context of Heart arrhythmia.

2. Background

According to [Mitchell 1997], machine learning is a subarea of Artificial Intelligence responsible for the development of models (hypotheses) generated from data, and that automatically improve with the experience. In this way, machine learning aims at constructing models that can be learned according to samples and past experiences.

As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications [Guyon and Elisseeff 2003, Liu and Motoda 2012]. The objectives of feature selection are manifold, the most important ones being: (a) to avoid overfitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering, (b) to provide faster and more cost-effective models and (c) to gain a deeper insight into the underlying processes that generated the data.

Classification is well known machine learning task. Data classification is the process of creating a prediction model from a learning algorithm. The aim of this model is to predict the value of the class attribute of testing instances. In this paper, we use the following classification methods: k-NN, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Decision tree, Naive Bayes, Random Forest and AdaBoost.

An ensemble of classifiers can be defined as a collection of classifiers responsible for producing the final output of the system when aggregated to a combination method [Kuncheva 2004]. The ensembles lead to greater generalization capacity than when working separately. Several studies have shown that combining the results of different classifiers outperform base classifiers. In the of ensemble of classifiers, three main aspects must be considered [Kuncheva 2004], which are:

- The structure of the ensemble system: it defines how the components are organized, and how they will interact [Kuncheva 2004].
- The components of the ensemble system: it is necessary to define which components will be used to compose the system [Kuncheva 2004]:
- The combination method: this method is used to combine results of the output of the classifiers [Kuncheva 2004]. There is a vast number of combination-based methods reported in the literature.

3. Related Works

As the main purpose of this work is to develop a cardiac arrhythmia monitoring platform based on feature selection techniques and classification methods, we will present the works more related to this main goal.

In [Soman and Bobbie 2005], the authors used three machine learning methods that were applied in the 5 tasks of arrhythmia classification, and the most accurate learning methods were evaluated. In [Tsipouras et al. 2005], the authors propose a knowledge-based method for classification of arrhythmic beats and detection of arrhythmic episodes using only the RR interval signal extracted from ECG recordings.

In the work [Oresko et al. 2010] is explored the use of a hierarchical model for the classification of cardiac arrhythmia. In addition, the authors investigate the performance of machine learning techniques for the four-beat classification of cardiac arrhythmia. In this article [Jenny et al. 2014] the authors propose a computer aided ECG diagnostic system for premature ventricular contraction (PVC). The proposed system uses Independent Component Analysis (ICA) to extract resources for cluster k-means and the Fuzzy C-Means classifier (FCM).

In the work [Ahmed and Arafat 2014] is presented the development of a platform based on smartphones for detection of cardiovascular diseases using wearable device, which are able to perform in real time the acquisition of ECG data, extraction of characteristics, and classification of arrhythmias. The author used to classify the machine learning algorithm MLP.

In the paper [Desai et al. 2015], a machine learning approach for computer-aided detection of the five classes of ECG arrhythmia beats is described using Discrete Wavelet Transform (DWT). The paper [Pławiak 2018] presents a methodology that allows the efficient classification of cardiac disorders (17 classes) based on ECG signal analysis and a neural evolutionary system.) To improve the characteristics of the ECG signal, the spectral power density was estimated (using the Welch method and a discrete Fourier transform).

4. The Proposed Monitoring Platform

As mentioned, the main aim of this paper is to develop a platform based on Internet of Things infrastructure capable of automatically monitoring and identifying cardiac arrhythmia. We discuss the relevant details in the next subsection.

4.1. Platform Requirement

In order to build a platform that allows data collection, storage, patient monitoring, prediction and event notification, it was necessary to define a set of functional and non-functional requirements so that the proposed platform can fulfill its purpose.

4.1.1. Non-Functional Requirements

During the planning and analysis of the proposed platform, it was necessary to identify the non-functional requirements necessary for the best functioning of the platform. From this analysis, we discuss some details below:

- In something of great importance as health, especially if it is of organs of greater importance as the heart, the diagnostic time is of fundamental importance, therefore, it can modify completely the mode of intervention. In this sense, the faster the diagnosis, the lower the risk of complications. So one of the requirements of the platform is to perform the analysis of the information in **real-time**. For example, if a significant variation in cardiac electrical behavior occurs in an Intensive Care Unit patient, the proposed solution should activate an alarm within a few seconds.
- **Data privacy** is an essential factor in a system, and especially when the system is in the health area where there is a lot of intimate and sensitive information because of the discriminatory potential they hold. Therefore, a system of permissions will be created on the platform in which you need the permission of the data type to have access to it and to avoid using it in undue ways.

4.1.2. Functional Requirement

The functional requirements are discuss below:

- **Collect Data:** the platform must be capable of continuous data collection from the monitored patient. All collections will be recorded as they occur, over a certain period of time.
- **Predict information:** the platform can predict information according to the data being collected and stored by the system. The platform will have the ability to learn criteria that can support future decisions, such as discovering new knowledge, finding unknown patterns in the data.
- **Send Alerts:** the platform must have an alert system that will be responsible for notifying end users about the health of the hospitalized patient. From the moment the platform judges that the information detected is critical the system will issue the alert for the service.

4.2. System Abstraction

Figure 1 system abstraction of the intelligent Cardiac Arrhythmia Monitoring Platform. The system will be designed to ensure support for the various data sources found that use the *Health Level 7* (HL7) protocol in the most varied contexts. It is noted that the general architecture can be divided into three layers: monitoring, *middleware* and services.

The first layer of monitoring is composed of two elements, being: the **sensor** that are the devices used for the monitoring of events in the body, the ECG is the main sensor in the use of this work. The **gateway** will be able to receive the data directly from the sensors and pass it on to the layer above, which is the **middleware**.

The second layer of the architecture is **middleware**, which is responsible for receiving the data from the monitoring layer, processing it, and making it available to the services layer. This layer is composed of one following element: the **intelligence module**. The intelligence module receives the data from the gateway and performs the processing of these using machine learning algorithms for predicting information in real-time flow. After detecting some kind of abnormality in the patient's health, the module will provide notifications information to the service layer.

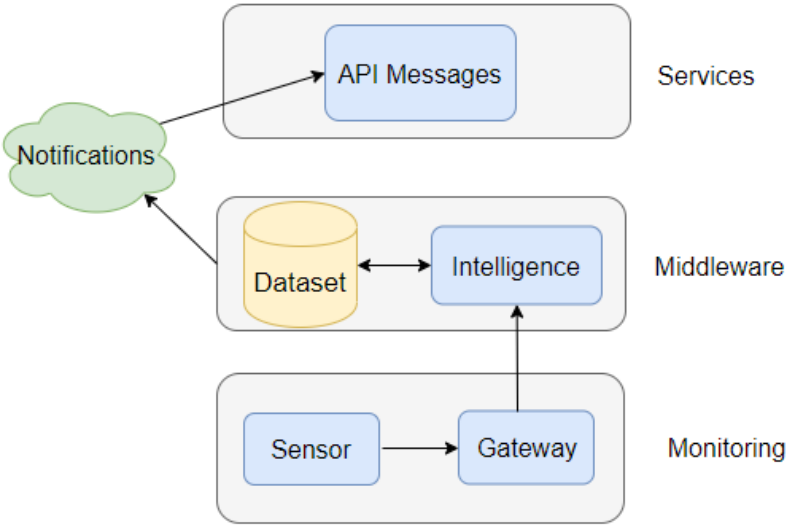

Figure 1. System Abstraction

The third layer of services is responsible for the **services** and programming of interfaces Applications (API). This layer is composed of the **API messages** that will have Representational State Transfer (REST) technology that is responsible for providing information for edge devices such as smart clocks, smartphones and computers.

4.3. The functioning of the Platform

The proposed platform consists of 6 steps such as describe the Figure2

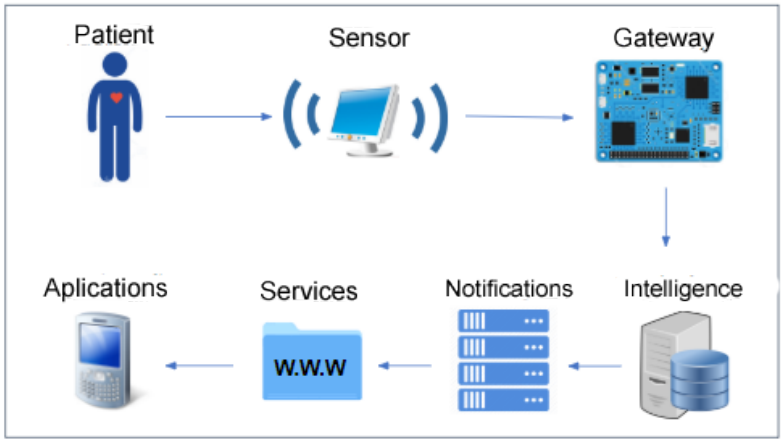

Figure 2. Functioning of the Platform

5. Methodology

In this section, important details will be described such as: datasets and experimental configuration.

5.1. Datasets

Two well-known datasets were used which are widely used in the literature being MIT-BIH[1] and UCI[2] Dataset.

- MIT-BIH Arrhythmia dataset: the data consist of 48 records, 30 min in length, extracted from 24 hours of ECG acquisition. The signs were acquired from 47 patients between 1975 and 1979 at the Boston Beth Israel Hospital Arrhythmia Laboratory, aged 23-89 years, of whom 22 were women and 25 men. The heart rate was marked and classified manually by specialists in 15 classes on the type of arrhythmia.
- UCI Arrhythmia dataset: this dataset contains 279 attributes, 206 of which are linear valued and the rest are nominal. The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones.

5.2. Implementational Issues

In order to evaluate the performance of the classification methods, we created five versions of the datasets, randomizing the original instances. In this way, each classification method were applied to two original datasets and their five versions, performing 12 datasets in total. For simplicity reason, allow us to call this as the first scenario.

From the first scenario (12 datasets), we applied different feature selection techniques (attribute evaluator and search method), such as: CfsSubsetEval with GreedyStepwise, WrapperSubsetEval (J48) with GreedyStepwise, and WrapperSubsetEval (Naive Bayes) with GreedyStepwise. In this way, we obtained three more scenarios.

In this paper, in order to obtain a better estimation of the accuracy rates, we applied 10-fold cross validation, and 4 different percentage splits (90%, 75%, 66% and 50%) for all classification methods.

The Friedman statistical test was applied to evaluate the performance of all classifiers in four different scenarios. It is important to emphasize that the Friedman test is applied directly on the accuracy values of all classifiers. If any significant difference is detected, a posthoc test will be applied.

6. Results

In order to validate our proposed platform, we conducted an empirical analysis which will be described in details in the next subsections.

6.1. Performance Analysis

Table 1 presents the average accuracy and standard deviation for all seven classifiers. Note that AdaBoost/J48 classifier obtained the best accuracy results in both datasets.

As it can be noticed in Table 2, AdaBoost/J48 classifier was significantly better than the other classifiers.

[1] https://www.physionet.org/physiobank/database/mitdb/

[2] https://archive.ics.uci.edu/ml/datasets/arrhythmia

Table 1. Scenario 1 - Original Datasets

Datasets / Classifier	MIT arrhythmia	UCI arrhythmia	Average / Standard Deviation
Ada/J48	**93,08 1,49**	**91,53 1,15**	**89,55 1,56**
J48	70,65 4,86	64,47 4,11	52,88 3,52
kNN	81,86 3,17	78,00 2,63	71,22 2,54
MLP	81,41 2,47	68,01 4,20	74,71 3,33
NB	75,79 2,44	72,70 4,47	74,25 3,46
RF	91,73 1,79	71,12 4,70	81,43 3,25
SVM	75,24 2,14	63,04 4,73	69,14 3,43

Table 2. Scenario 1 - Original Datasets

Nemenyi p-values, with no further adjustment

	Ada/J48	J48	kNN	MLP	NB	RF
J48	0,00000					
kNN	0,00000	0,00008				
MLP	0,00000	0,87580	0,01376			
NB	0,00000	0,00000	0,38330	0,00000		
RF	0,00480	0,62260	0,00000	0,05167	0,00000	
SVM	0,00000	0,83746	0,01825	1,00000	0,00000	0,04037

Table 3 presents the accuracy results for all classifiers in scenario 2. Note that the Random Forest classifier obtained the best overall results. However, it outperformed the AdaBoost/J48 only in the UCI arrhythmia dataset.

Table 3. Scenario 2 - Feature Selection (CfsSubsetEval with GreedyStepwise)

Datasets / Classifier	MIT arrhythmia	UCI arrhythmia	Average / Standard Deviation
Ada/J48	**89,05 1,96**	72,02 4,85	80,54 3,40
J48	87,75 1,26	69,29 4,29	78,52 2,77
kNN	88,74 2,11	63,25 3,70	76,00 2,90
MLP	71,32 2,96	68,09 3,39	69,71 3,18
NB	61,19 2,44	69,50 4,47	65,35 3,45
RF	88,39 1,89	**75,13 3,62**	**81,76 2,75**
SVM	61,40 3,52	69,21 4,65	65,31 4,08

It is important to emphasize that Random Forest classifier was significantly better than the other classifiers. However, it was not the case when compared to AdaBoost/J48 (see Table 4).

Table 5 presents the accuracy results for all classifiers in scenario 3. Again the Random Forest classifier obtained the best overall results. However, it outperformed AdaBoost/J48 only in the UCI arrhythmia dataset. As it can be seen in Table 6, Random Forest classifier was significantly better than kNN, MLP and SVM classifiers.

Table 7 presents the accuracy results for all classifiers in scenario 4. Again the

Table 4. Scenario 2 - Feature Selection (CfsSubsetEval with GreedyStepwise)

Nemenyi p-values, with no further adjustment						
	Ada/J48	**J48**	**kNN**	**MLP**	**NB**	**RF**
J48	0,00127					
kNN	0,00000	0,60833				
MLP	0,00000	0,03556	0,82700			
NB	0,00000	0,00057	0,16946	0,92227		
RF	0,99983	0,00027	0,00000	0,00000	0,00000	
SVM	0,00000	0,00075	0,19471	0,94054	1,00000	0,00000

Table 5. Scenario 3 - Feature Selection (WrapperSubsetEval/J48 with GreedyStepwise)

Datasets / Classifier	**MIT arrhythmia**		**UCI arrhythmia**		**Average / Standard Deviation**	
Ada/J48	**93,03**	**1,56**	68,34	4,25	80,68	2,91
J48	91,60	1,12	72,35	3,60	81,98	2,36
kNN	89,01	1,81	61,12	3,88	75,07	2,84
MLP	80,83	2,68	70,10	3,53	75,46	3,11
NB	75,03	1,90	66,40	5,29	70,72	3,60
RF	91,82	3,74	**72,80**	**3,53**	**82,31**	**3,63**
SVM	75,59	3,92	60,46	5,37	68,02	4,64

Table 6. Scenario 3 - Feature Selection (WrapperSubsetEval/J48 with GreedyStepwise)

Nemenyi p-values, with no further adjustment						
	Ada/J48	**J48**	**kNN**	**MLP**	**NB**	**RF**
J48	0,99903					
kNN	0,00000	0,00000				
MLP	0,01376	0,00231	0,07351			
NB	0,00000	0,00000	0,93481	0,00195		
RF	0,67863	0,92227	0,00000	0,00001	0,00000	
SVM	0,00000	0,00000	0,01956	0,00000	0,30841	0,00000

Random Forest classifier obtained the best overall results. However, it was outperformed by AdaBoost/J48 and Naive Bayes in MIT arrhythmia and UCI arrhythmia datasets.

Again, as it can be seen in Table 8, Random Forest classifier was significantly better than kNN, MLP and SVM. classifiers. Moreover, Random Forest was not significantly better than AdaBoost/J48 in any scenario.

7. Conclusion

In this paper, a platform proposal for the cardiac arrhythmia monitoring platform was presented, and also an empirical analysis was conducted in order to evaluated the accuracy performance of seven classification methods (Adaboost, k-NN, Naive Bayes, Decision

Table 7. Scenario 4 - Feature Selection (WrapperSubsetEval/Naive Bayes with GreedyStepwise)

Datasets / Classifier	MIT arrhythmia	UCI arrhythmia	Average / Standard Deviation
Ada/J48	**93,14 1,67**	66,94 3,00	80,04 2,33
J48	91,23 1,32	67,29 3,82	79,26 2,57
kNN	89,06 1,75	61,49 4,19	75,27 2,97
MLP	81,41 2,47	68,01 4,20	74,71 3,33
NB	75,79 2,44	**72,70 4,47**	74,25 3,46
RF	91,73 1,79	71,12 4,70	**81,43 3,25**
SVM	75,24 2,14	63,04 4,73	69,14 3,43

Table 8. Scenario 4 - Feature Selection (WrapperSubsetEval/Naive Bayes with GreedyStepwise)

Nemenyi p-values, with no further adjustment						
	Ada/J48	**J48**	**kNN**	**MLP**	**NB**	**RF**
J48	0,83746					
kNN	0,00000	0,00003				
MLP	0,00098	0,09191	0,34485			
NB	0,06942	0,74492	0,01478	0,88447		
RF	0,78187	0,08230	0,00000	0,00000	0,00039	
SVM	0,00000	0,00000	0,11988	0,00005	0,00000	0,00000

Tree, Random Forest, SVM and MLP). Moreover, two well-known cardiac arrhythmia datasets were used in this analysis. Feature selection techniques were also applied to theses two datasets.

In general, some significant improvements were observed in some cases after applying feature selection techniques, besides the reduction of the dataset and consequently an improvement in system response time. Among the seven different classification methods, ensembles of classifiers, such as Random Forest and AdaBoost (J48) have outperformed all base classification methods.

The results indicate that the use of feature selection can help us to better understand the most relevant attributes in the datasets, and also to maintain a high degree of accuracy, which shows that machine learning can assist with a good degree of accuracy in the diagnosis of cardiac arrhythmia.

7.1. Future Works

- It is planned to carry out new case studies, and with this, it is intended to evaluate all the operating flows and requirements present in the platform.
- Creat partnerships with local doctors to determine the creation of a dataset regional of cardiac arrhythmia, thus, contributing to science and well-being of peoples.
- Using committees with different numbers of members and investigate other classifiers bases.

References

Ahmed, R. and Arafat, S. (2014). Cardiac arrhythmia classification using hierarchical classification model. In *Computer Science and Information Technology (CSIT), 2014 6th International Conference on*, pages 203–207. IEEE.

Desai, U., Martis, R. J., Nayak, C. G., Sarika, K., and Seshikala, G. (2015). Machine intelligent diagnosis of ecg for arrhythmia classification using dwt, ica and svm techniques. In *India Conference (INDICON), 2015 Annual IEEE*, pages 1–4. IEEE.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Ilayaraja, M. and Meyyappan, T. (2013). Mining medical data to identify frequent diseases using apriori algorithm. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on*, pages 194–199. IEEE.

Jenny, N. Z. N., Faust, O., and Yu, W. (2014). Automated classification of normal and premature ventricular contractions in electrocardiogram signals. *Journal of Medical Imaging and Health Informatics*, 4(6):886–892.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

Mitchell, T. M. (1997). *Machine Learning. First edition*.

Oresko, J. J., Jin, Z., Cheng, J., Huang, S., Sun, Y., Duschl, H., and Cheng, A. C. (2010). A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):734–740.

Pławiak, P. (2018). Novel methodology of cardiac health recognition based on ecg signals and evolutionary-neural system. *Expert Systems with Applications*, 92:334–349.

Soman, T. and Bobbie, P. O. (2005). Classification of arrhythmia using machine learning techniques. *WSEAS Transactions on computers*, 4(6):548–552.

Tsipouras, M. G., Fotiadis, D. I., and Sideris, D. (2005). An arrhythmia classification system based on the rr-interval signal. *Artificial intelligence in medicine*, 33(3):237–250.