Feature Extraction Methods for Predicting the Prevalence of Heart Disease

Ngong Ivoline-Clarisse Kieleh
Konya Technical University
Konya, Turkey
ivolinengong@gmail.com

***Abstract*—Heart disease is the number one cause of death in the world and Cardiac Arrhythmias is one of the leading causes of cardiac death in the world today. In this study, features are extracted from an arrhythmias dataset and different kinds of arrhythmias are classified. Random Forests, Boosted Trees, Principal Component Analysis (PCA) and Convolutional Neural Networks (CNN) are the feature extraction techniques used. After the features are extracted, a Support Vector Machine (SVM) classifier is used. From the performance measures calculated it can be seen that CNN is the best feature extraction method with an accuracy of 95.53%.**

***Keywords*—*Feature Extraction, Random Forest, Boosted Trees, Principal Component Analysis (PCA) and Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Heart Disease, Arrhythmias, ECG, Classification*.**

I. INTRODUCTION

An irregular heartbeat is called an arrhythmia – when the heart beats too slowly, too fast or irregularly. Some arrhythmias are harmless but if they are particularly abnormal or result from a weak or damaged heart, arrhythmias can cause serious and even potentially fatal symptoms. Coronary artery diseases have become a major cause of death globally. Annually, an estimated 610.000 people die in the United States die due to different types of heart diseases. 370,000 people die from coronary heart disease which is the most common heart disease [1]. In the United Kingdom, 154.639 people died from cardiovascular diseases in 2014 (25% of deaths in the UK in that year) [2]. It was also reported that in 2016, 3.4million Turkish adults were living with cardiovascular diseases and the prevalence was projected to increase to 5.4million by 2035 [3]. Identifying arrhythmias from ECG recordings is therefore important for clinical diagnosis and treatment not just today but for future generations.

The electrical activity of the heart is represented by an ECG signal. ECG signals are measured using an electrocardiogram which uses electrodes placed on the skin to measure a trace of a person's heartbeat. In the course of a heartbeat, the heart muscles contract and relax causing electrical depolarization and repolarization of these muscles. These events are recorded on the electrocardiogram as deflections on an ECG trace [4]. The different sections of the ECG signals are described by its labels; P, QRS, and T as shown in Fig. 1. The P wave represents atrial contraction that pumps blood to the ventricles, the QRS complex corresponds to ventricular contraction that pumps blood to the lungs and the rest of the body while the T wave represents ventricular repolarization [5].

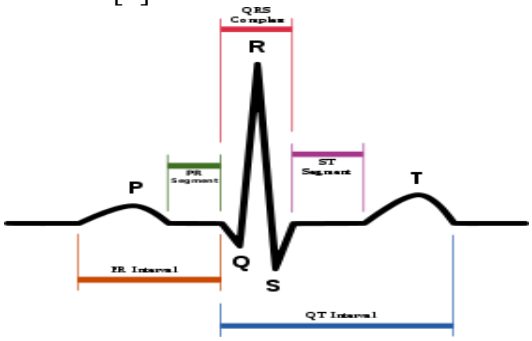


Figure 1 ECG Signal

II. PROBLEM DEFİNİTİON

The aim of this study is to determine the presence or absence of arrhythmia in ECG signals as humans can find identifying and classifying it difficult. Moreover, there is a likelihood of human error during the analysis due to distraction or fatigue. Additionally, the study tries to find the effect of different feature extraction methods on the SVM classifier.

Recently substantial research has been carried out for analyzing ECG signals. According to Karpagachelvi, Arthanari and Sivakumar the majority of these studies are based on Fuzzy Logic Methods, Genetic Algorithm, Artificial Neural Networks, Support Vector Machines and other techniques in signal analysis. In their paper, they do a comparative study the different methods proposed by 10 research works in extracting features thereby giving an overview [6]. Our study adds a broader variety of machine learning methods to the existing ones. A step is taken to further perform feature selection with tree-based methods, a neural network and a linear transformation method. Thereby, cutting across different extremes.

III. PROPOSED METHOD

After the data is acquired, the signals are preprocessed and feature extraction is done. Each of the feature extraction methods are applied and finally the SVM classifier is applied on the dataset with the extracted features to get the results. Hence, the proposed system is composed of 4 major steps for classification.

A. Data Acquisition

The MIT-BIH Arrythmia Dataset is used in this study. The data was gotten from Kaggle. The dataset contains ECG recordings obtained from 47 subjects gotten between 1975 and 1979 in the BIG Arrhythmia Laboratory. The recordings were digitized at 30samples per second per channel with 11-bit resolution over a 10mV range. Also, these recordings were annotated by 2 or more cardiologists. The dataset's content has been summarized:

*Number of Samples: 1*09446
Number of Categories: 5
Number of Features: 187
Sampling Frequency: 125Hz
Data Source: MIT-BIH Arrhythmia Dataset
Classes: ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4]

B. ECG signal Preprocessing and Heartbeat Segmentation

Preprocessing is a very important step and should be done carefully as it influences the final results. The dataset used was preprocessed by Kachuee, Fazeli and Sarrafzadeh [1]. The methodology is described in Section III.A of the paper. The suggested method is simple and effective in extracting the R-R intervals from signals with different morphologies. Moreover, all the extracted beats have identical lengths which are used as inputs to the subsequent processing parts. Amplification and stretching of the signals (was only done in CNN) and Feature scaling (Normalization) were other preprocessing techniques that carried. Additionally, missing values did not pose a problem as no missing values were present in the dataset.

C. Feature Extraction/Selection

The feature extraction stage is a very crucial stage in the classification method. The least discriminative features can be found by various greedy feature selection approaches. However, in practice, many features depend on each other or on an underlying unknown variable. Therefore a single feature can be used to represent a combination of multiple types of information and removing this feature can cause the loss of important information.
Some methods used for feature extraction include Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Independent Component Analysis. PCA and other feature selection methods are used in this study.

1. Principal *Component Analysis (PCA)*

PCA is a mathematical procedure that uses linear transformations to map data from high dimensional spaces to low dimensional space.[6] Technically, PCA finds the eigen vectors of a covariance matrix that have the highest eigen vlaues and then use these values to project the data to a new subspace which can either have equal or less dimensions. Practically, it converts a matrix of n features into a new dataset of n or less than n features. It therefore reduces the number of features by building new and smaller number variables which capture a significant portion of the information found in the original features. PCA is carried out following the steps below:

1. The mean vector of data is calculate
2. The covariance matrix of the data is calculated
3. The eigenvalue and eigenvector of the covariance matrix is evaluated
4. The principal components are formed using the eigenvectors of the covariance matrix as coefficients of the weights.

PCA may not perform well for all datasets but its low sensitivity to noise, increased efficiency and low memory capacity makes it advantageous to use.

2. *Boosted Trees*

Boosted Tree is an ensemble of classification or regression trees. It sequentially applies weak learners to the incrementally changing data then

creates a series of decision trees that produce an ensemble of weak prediction models. Simply put, the boosted tree starts with a small tree and builds other tree models which it adds to the small tree. An observation is assigned a higher weight in the next iteration if it was misclassified. The weighted sum of the decisions made by the trees produces the final classifier. Boosting is a flexible nonlinear regression algorithm and it increases the accuracy of trees. However, the speed and clarity to humans. Gradient boosting tries to reduce these issues by generalizing the tree boosting. The library used in the study is XGBoost.

3. *Random Forest*

Random Forest is an ensemble learning technique based on bagging that works by constructing a multitude of decision trees at training time. For classification, it outputs the mode of the classes while for regression it calculates the mean prediction for the individual trees. [8]One advantage of random tress is that it corrects overfitting of the training set which is very common in decision trees. Random Forest in the project was implemented through the scikit learn library in python.

4. *Convolutional Neural Networks (CNN)*

CNN are a class of deep neural networks, a variation of multilayer perceptrons that have been designed to need very little preprocessing. They are regularly applied in analyzing visual images. CNNs were inspired by the biological processes in the visual cortex of an animal. Neurons in the cortex respond to stimuli in the receptive field which is a made up many neurons that partially overlap each other [9]. A major advantage of CNNs is that unlike other image classification algorithms, a CNN based classification system automatically learns feature representation. Hence, no need for prior knowledge or for the filtering to be explicitly programmed. One of the reasons why CNN can be good feature selector.

D. *Learning/classification*

Support Vector Machines (SVM) has been used for classifying the extracted features in the system. SVM is a supervised learning algorithm that performs classification for 2 linearly separable classes by finding a hyperplane which separates the input space with a maximum margin. Consider the points in our dataset to be: (x_1,y_1) …$(x_n…y_n)$ and we have to classify between two classes 0 and 1. We get the best hyperplane by doing:[10]

$$w.x_i + b \geq 1, if\ y_i = 1 \qquad (1)$$

$$w.x_i + b \leq 0, if\ y_i = 0 \qquad (2)$$

Where w is the weight vector and b the bias. According to the equation 1, any result above 1 is classified as belong to class 1 while in equation 2 any value below 0 is classified as belonging to class 0. However, this works only for linearly separable data. Our data is expected to be non-linearly separable, therefore a kernel function is required to transform the problem into a linearly separable space.

$$K(x_i, y_i) = \varphi(x)\varphi(x_j) \qquad (3)$$

For a non-linearly separable data with 2 classes, the solution is as follows:[10]

$$f(x) = sign(\sum_i \alpha_i\, y_j \varphi(x)\varphi(x_i) + b) \qquad (4)$$

IV. EXPERİMENTS/RESULTS

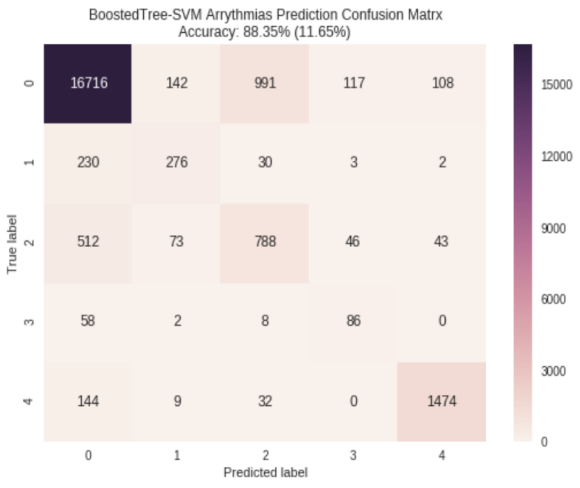
The models were then evaluated and we got the confusion matrix. The confusion matrix is a technique for summarizing the performance classification of algorithm. When dealing with confusion matrix it is common to hear terms like true positive (TP), true negative (TN), false positive (FN) and false negative (FN). From the confusion matrix we got the accuracy, sensitivity and specificity. These can be expressed in terms of TP, TN, FN and FP.

Sensitivity = TP / (TP + FN) (5)

Specificity = TN / (TN + FP (6)

Accuracy = (TN + TP) /(TN+TP+FN+FP) (7)

Figure 2 Boosted Tree Confusion Matrix



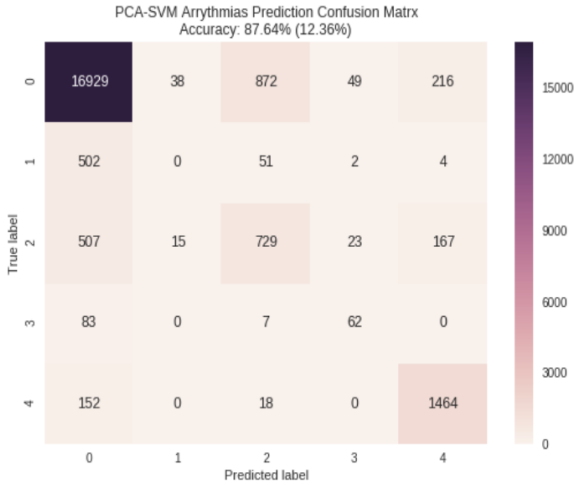


Figure 3 PCA Confusion Matrix

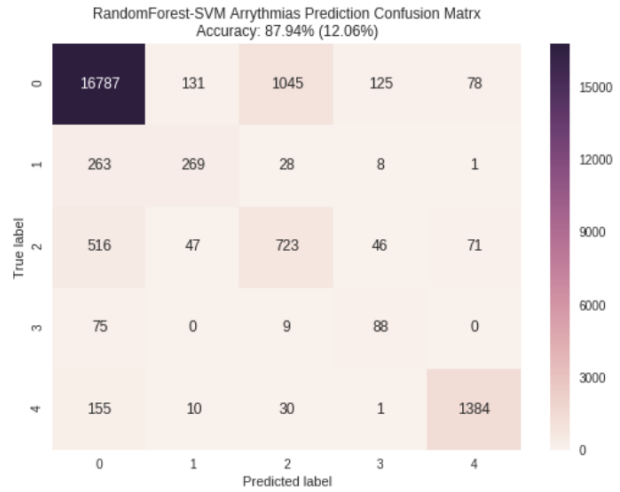


Figure 4 Random Forest Confusion Matrix

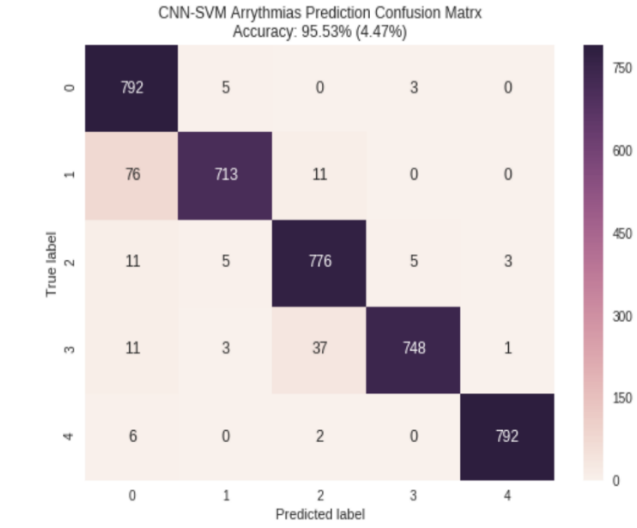


Figure 5 CNN Confusion Matrix

The results are summarized in Table 1:

Table 1 Performance Measures

SVM	PCA (%)	Boosted Trees (%)	Random Forest (%)	CNN (%)
Accuracy	87,64	88.30	87.94	95.53
Sensitivity	83.77	82.30	82.46	55.18
Specificity	82.14	83.03	82.71	79.74

V. CONCLUSİONS

In this study, 4 different machine learning algorithms were used for feature extraction/selection from Arrythmia ECG signals and classification was done with SVM. Among the feature extraction methods, the CNN model performed better than any other model. Followed by the boosted trees, then Random Forest and lastly PCA.

We show that the CNN-SVM model represents an initial version of a potentially useful tool. The performance measures of the classifier for each extraction or selection method were observed in terms of accuracy, sensitivity and specificity. With CNN an accuracy of 95. % was obtained. Though this results shows promise, there is much room for improvement. In future steps, techniques to improve these models performances will be applied

VI. REFERENCES

[1] CDC, NCHS, “Heart Disease Facts”,2017. [Online].Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 5-Jan-2019].

[2] Office of National Statistics, “Deaths Registered in England and Wales,” 2014.

[3] Y. Balbay, I. Gagnon, S. Malhan, M. Ergun, G. Sutherland, A. Dobrescu , G. Villa, G. Ertuğru , M. Habib, “Modeling the burden of cardiovascular disease in Turkey”, Turkish Society of Cardiology, 2018.

[4] S. McCandlish, T. Barrella,“Identifying Arrhythmia from Electrocardiogram Data”, unpublished,2014

[5] H. Montazeri, “ Hybrid Neuro-Fractal Analysis Of Ecg Signals To Predict Ischemia”, Arak University, 2008.

[6] S. Karpagachelvi, M. Arthanari, M. Sivakumar, “ECG Feature Extraction Techniques - A Survey Approach”, International Journal of Computer Science and Information Security, 2010.

[7] M. Suganthy, P. Ramamoorthy, “Principal Component Analysis Based Feature Extraction, Morphological Edge Detection and Localization for Fast Iris Recognition”, Journal of Computer Science, 2012

[8] Wikipedia contributors, "Random forest," *Wikipedia, The Free Encyclopedia,* [Online].Available: https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=877940975 (Accessed: 15-Jan-2019).

[9] Wikipedia contributors, "Convolutional neural network," *Wikipedia, The Free Encyclopedia,* [Online].Available: https://en.wikipedia.org/w/index.php?title=Convolutional_neural_network&oldid=876940492(Accessed: 15-Jan-2019).

[10] Bayram,Kizrak, Bolat, “Classification of EEG Signals by using Support Vector Machines”, Conference Paper · June 2013