CLASSIFICATION OF ARRHYTHMIA USING ECG DATA

A project report submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By

Sachi Tripathi (181500598)

Sarvesh Kumar Sharma (181500625)

Satyam Kumar Jha (181500627)

Shivani Chauhan (181500678)

Under the guidance of

Dr. Ashish Sharma

Department of Computer Engineering & Applications

Institute of Engineering & Technology

GLA UNIVERSITY

Mathura – 281406, India

Dec, 2020

Department of computer Engineering and Applications, GLA University
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuhan, Mathura, 281406

DECLARATION

We hereby declare that the work which is being presented in the B.Tech. Project "**Classification of Arrhythmia Using ECG Data**" in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Science & Engineering of GLA University, Mathura, is an authentic record of our own work carried under the supervision of **Dr. Ashish Sharma**, Assistant Professor, Institute of Engineering & Technology.

The contents of this project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

Sign ______________________

Name of Candidate: Sachi Tripathi

University Roll No.: 181500598

Sign ______________________

Name of Candidate: Sarvesh Kumar Sharma

University Roll No.: 181500625

Sign ______________________

Name of Candidate: Satyam Kumar Jha

University Roll No.: 181500627

Sign ______________________

Name of Candidate: Shivani Chauhan

University Roll No.: 181500678

CERTIFICATE

This is to certify that the above statements made by the candidate are correct and true to the best of my knowledge and belief.

____

Supervisor

Dr. Ashish Sharma

Assistant Professor

Institute of Engineering & Technology

GLA University

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the Machine Learning Project undertaken during B. Tech. (Third Year). This project in itself is an acknowledgment of the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.

Our heartiest thanks to **Dr. (Prof). Anand Singh Jalal**, Head of Dept., Department of CEA for providing us with an encouraging platform to develop this project, which thus helped us in shaping our abilities towards a constructive goal.

We owe a special debt of gratitude to **Dr. Ashish Sharma**, Assistant Professor, Institute of Engineering & Technology, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness, and perseverance have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.

We also would not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Sachi Tripathi (181500598)

Sarvesh Kumar Sharma (181500625)

Satyam Kumar Jha (181500627)

Shivani Chauhan (181500678)

ABSTRACT

Cardiac Arrhythmia is a life-threatening disease, causing serious health issues in patients, when left untreated. An early diagnosis of arrhythmias would be helpful in saving millions of lives. This study is conducted to classify patients into one of the sixteen subclasses, among which one class represents absence of disease and the other fifteen classes represent electrocardiogram records of various subtypes of arrhythmias. The research is carried out on the dataset taken from the University of California at Irvine Machine Learning Data Repository. This data set contains large amount of feature dimensions which are reduced using dimensionality reduction techniques. In order to classify, various algorithms such as K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Random Forest, Linear SVM as well as Kernelized SVM are employed over original data to detect the presence and absence of arrhythmias. The accuracies are then improved by using Principal Component Analysis (PCA) over the original dataset. The models are then evaluated and compared using their accuracy and recall values. The results showed that on applying PCA over the data, Kernelized SVM outperforms the other classifiers with an accuracy rate of 80.21%.

TABLE OF CONTENTS

INTRODUCTION

1.1 MOTIVATION & OVERVIEW:

An early detection and accurate medical assistance to heart disease patients can save human lives as heart diseases can be life-threatening causing sudden death.

Cardiac Arrhythmia is a form of irregularity in heart rhythms and, in some cases, results in heart disease, which poses serious threats to human lives. It is a type of disease that disturbs the smooth rhythm of heart’s electrical system and causes the heart to beat either too slow or too fast, to race, and to skip beats and causes nonsequential movement of heart signals.

Generally, arrhythmia is identified and analyzed from an ECG recording along with the symptoms such as insufficient pumping of blood from heart, shortness of breath, fatigue, chest pain, and unconsciousness. Arrhythmias are generally divided into two broad categories, that is, Bradycardia and Tachycardia. Bradycardia causes the heart to beat too slow, which is usually below the rate of 60 beats per minute (bpm), while Tachycardia makes the heart beat faster which could go up to 100 bpm.

An electrocardiogram (ECG) measures the electric activity of the heart and has been widely used for detecting heart diseases due to its simplicity and non-invasive nature. By analyzing the electrical signal of each heartbeat, i.e., the combination of action impulse waveforms produced by different specialized cardiac tissues found in the heart, it is possible to detect some of its abnormalities. ECG signals are normally made of P waves, T waves, and QRS complex. The significant parameters required for the examination of heart-patients are time duration, shape, and the relationship between P wave, QRS complex, T wave, and R-R interval. Any abrupt change in these parameters indicates an ailment of the heart that may occur due to a wide range of reasons.

1.2 OBJECTIVE:

In our study, we are going to distinguish between the presence and absence of cardiac arrhythmia and classify it in one of the 16 groups. For the time being, there exists a computer program that makes such a classification. However, there are differences between the cardio log's and the programs classification. Taking the cardio log's as a gold standard we aim to minimize this difference by means of machine learning tools.

1.3 CONTRIBUTION:

The project proposes a diagnostic system built using Machine Learning. The data contains high dimensionality which is reduced using Principal Component Analysis (PCA). For training our model, Kernelized Support Vector Machine (SVM) is used which enhances the results produced by the original data set.

1.4 ORGANIZATION OF PROJECT REPORT:

The report is organized in five chapters. A discussion on previously implemented arrhythmic classification models and techniques is provided in Chapter 2. Chapter 3 presents our proposed work to classify arrhythmia into one of the 16 classes, all the algorithms used and pseudo-codes. The implementations and results of all the models are contained in Chapter 4 along with the snapshot of simulated work, and finally, the report is concluded in Chapter 5.

LITERATURE REVIEW

Various methods have been proposed to develop an automated model for classification of arrhythmia. Useful information in the ECG is found in the intervals and amplitudes of the characteristic waves. Any abnormality in the wave shape and duration of the wave feature is considered as arrhythmia. Using Logistic Model Tree (LMT), the classifier classifies the 11 different arrhythmias [1].

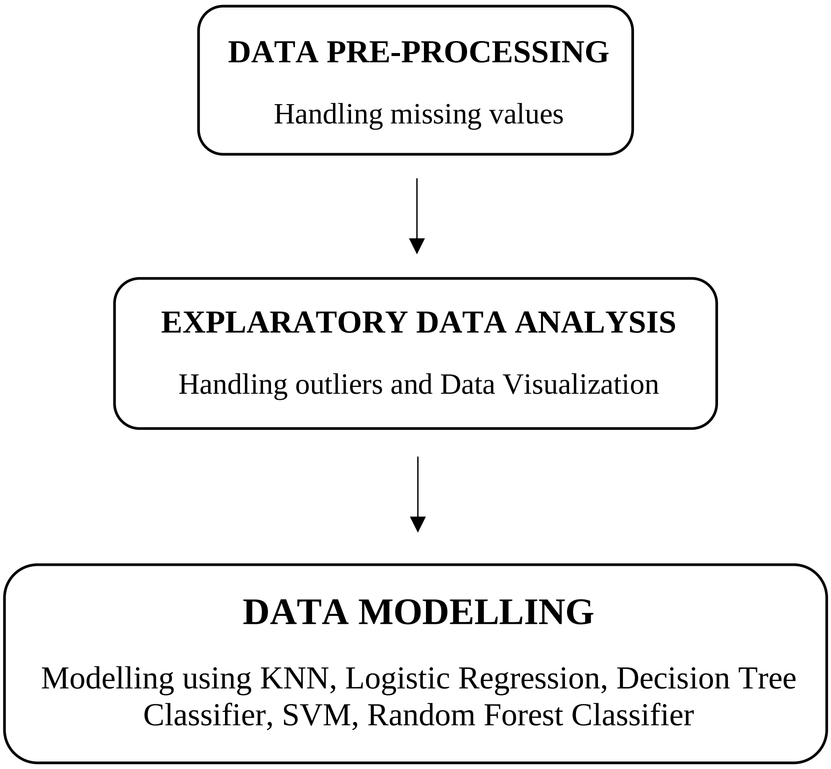
Multi-Class classification of cardiac arrhythmia by Anwar et al proposed SVM based approaches [2] including One-Against-One (OAO), One-Against-All (OAA), and error-correction code (ECC) using improved feature selection.

Another paper by Babak et al presents an SVM based classification using reduced features of heart rate variability (HRV) signal. The proposed algorithm is based on the generalized discriminant analysis (GDA) feature reduction scheme [3].

Nasiri et al presented a new approach for classification by combining both SVM and genetic algorithm approaches [4]. The genetic algorithm is used to improve the generalization performance of the SVM classifier to better classify ECG signals.

PROPOSED WORK

For our project, we have taken the dataset from UCI machine learning repository. The proposed work includes cleaning of the dataset, visualizing the features and modelling the data. The first phase includes modelling the data using all the 278 features. During the second phase, only the important features are modelled using principal component analysis (PCA). The steps involved are as shown in the figure.

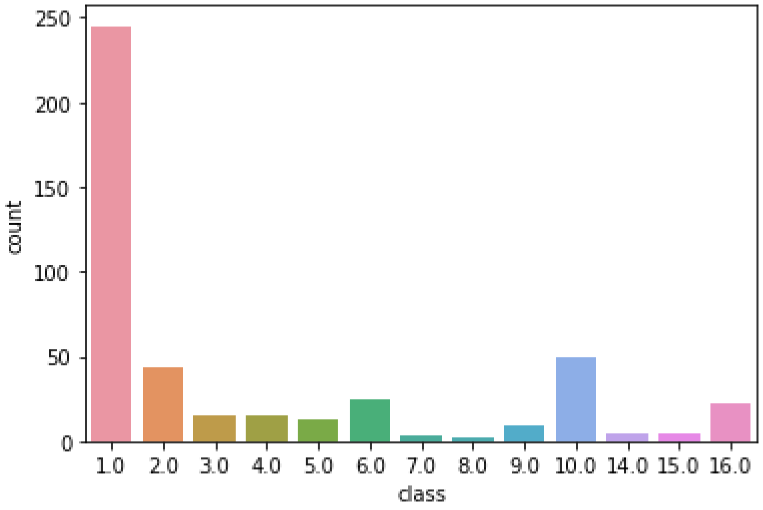

DATA PRE-PROCESSING
Handling missing values
EXPLARATORY DATA ANALYSIS
Handling outliers and Data Visualization
DATA MODELLING
Modelling using KNN, Logistic Regression, Decision Tree Classifier, SVM, Random Forest Classifier

3.1 DATA PRE-PROCESSING:

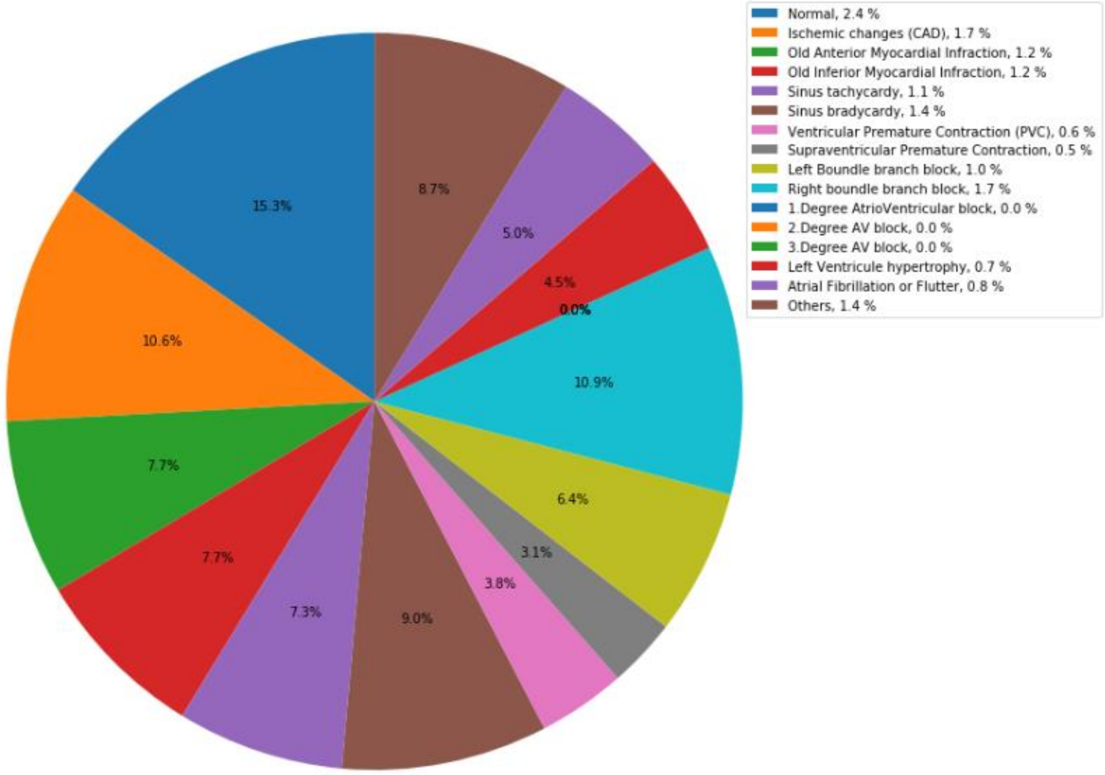
The first step of our project is data cleaning. We observed that out of 279 attributes, 5 of them contained missing values. Upon digging the data further, we found that an attribute contained almost 350 missing values. So, we dropped that column and imputed the other columns containing missing data using their mean values. To make our data more understandable, we replaced the column names with appropriates names as per the details given in the UCI machine learning repository. Finally, we separated the target attribute from the features of our data.

3.2 EXPLARATORY DATA ANALYSIS:

The count of each class in our dataset is plotted using countplot as shown.



Now, let’s visualize the percentage distribution of the counts using a pie chart for better clarity.



Out of 452 samples, 245 were of class 1 which is for ‘normal people’. Also, Atrio-Ventricular block Arrhythmia is not available in the dataset. The samples of classes 7 and 8 are also very few, making our dataset highly imbalance.

In order to find the outliers, we visualized the pairwise distribution of a few features and handled those outliers using boxplots.

We then perform feature scaling and split our dataset using 80% as training dataset and 20% as testing data.

3.3 DATA MODELLING:

The data is modelled using various algorithms. Let us see each of them, one by one.

K-NEAREST NEIGHBOURS:

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

The K parameter decides how many of the nearest neighbors have to be considered to determine the property of our unknown point.

KNN uses a similarity metric to determine the nearest neighbors. This similarity metric is more often the Euclidean distance between our unknown point and the other points in the dataset. The general formula for Euclidean distance is:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

where q_1 to q_n represent the attribute values for one observation and p_1 to p_n represent the attribute values for the other observation.

LOGISTIC REGRESSION:

If Y takes on more than two values like in our case, say k of them, we can still use logistic regression. Instead of having one set of parameters β_0, β, each class c in 0: (k

−1) will have its own offset $\beta^{(c)}_0$ and vector $\beta^{(c)}$, and the predicted conditional probabilities will be

$$\Pr\left(Y = c|\vec{X} = x\right) = \frac{e^{\beta_0^{(c)} + x\cdot\beta^{(c)}}}{\sum_c e^{\beta_0^{(c)} + x\cdot\beta^{(c)}}}$$

DECISION TREE:

A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continuous value).

Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The splitting process continues until no further gain can be made or a preset rule is met, e.g. the maximum depth of the tree is reached.

RANDOM FOREST:

The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees. Gini-index is often used to how branching is done by nodes in a decision tree.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, *pi* represents the relative frequency of the class we are observing in the dataset and *c* represents the number of classes.

SUPPORT VECTOR MACHINE:

The main objective of SVM is to find the optimal hyperplane which linearly separates the data points in two components by maximizing the margin. The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1. Computing the (soft-margin) SVM classifier amounts to minimizing an expression of the for

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2.$$

We focus on the soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly-classifiable input data.

The kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. There are also no constraints on the form of this mapping, which could even lead to infinite-dimensional spaces. The Sigmoid Kernel (Hyperbolic Tangent) comes from the Neural Networks field, where the bipolar sigmoid function is often used as an activation function for artificial neurons.

$$k(x, y) = \tanh(\alpha x^T y + c)$$

PRINCIPAL COMPONENT ANALYSIS:

In order to reduce the dimensionality of dataset consisting of large inter-related variables, while retaining as much as possible of the variation present in the dataset, we use PCA. The features are transformed into new set of variables called principal components, which are uncorrelated. Covariance is a metric measured between two variables. It gives a measure of how changes in one-dimension affect changes in the other.

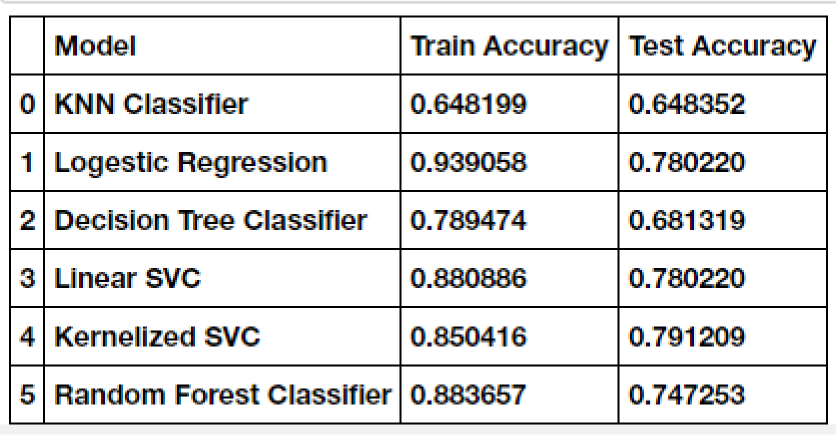
$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

An eigenvector or characteristic vector of a linear transformation, is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled.

$$A\vec{v} = \lambda\vec{v}$$

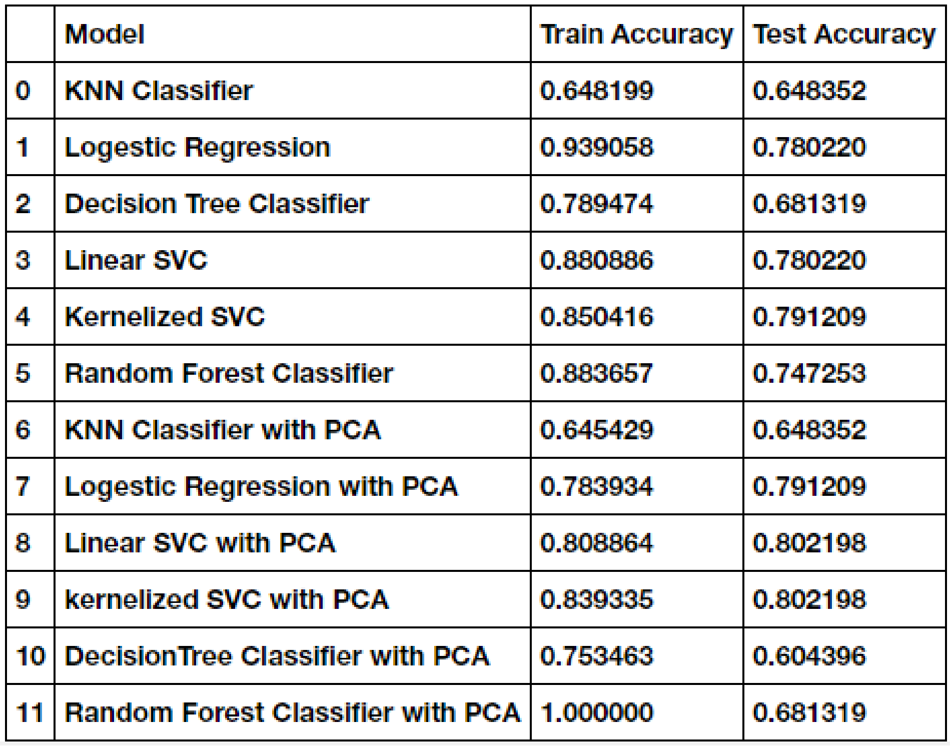
IMPLEMENTATION & RESULT ANALYSIS

When trained over original data, Kernelized SVM proved to be the best among other classifiers in terms of recall value, with an accuracy percent of 79.12%. Also, Logistic Regression showed better training accuracy.



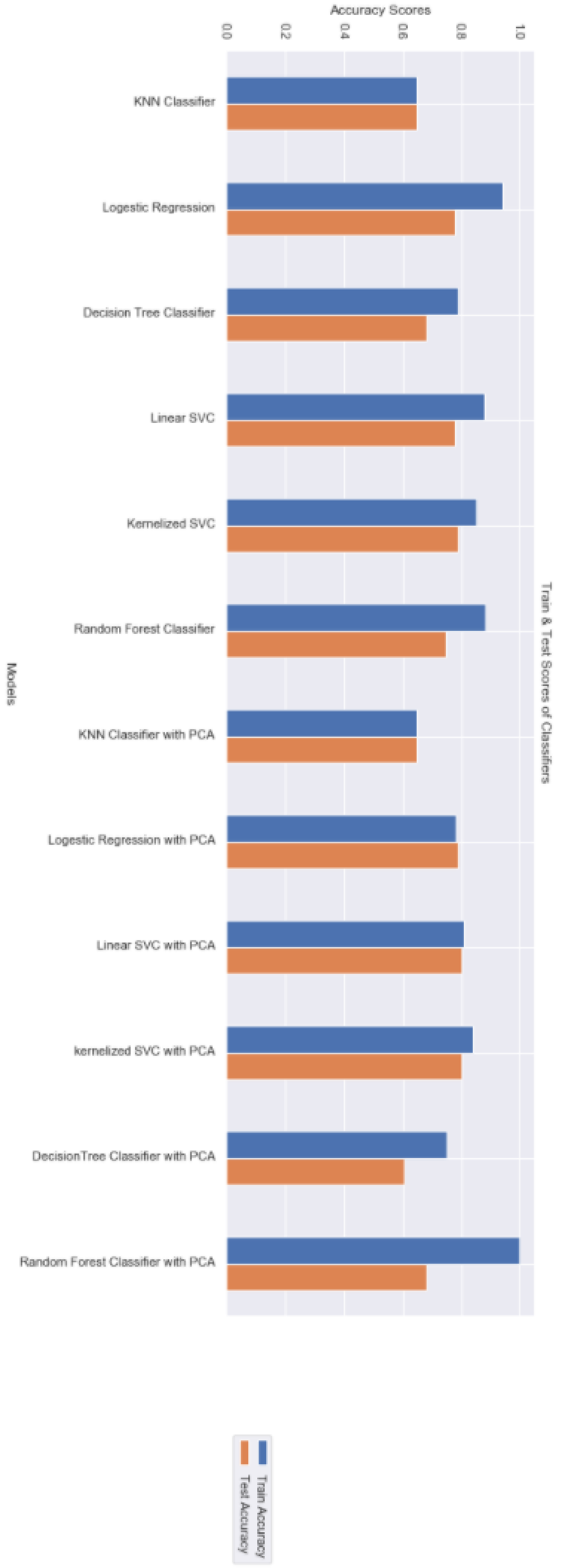
	Model	Train Accuracy	Test Accuracy
0	KNN Classifier	0.648199	0.648352
1	Logestic Regression	0.939058	0.780220
2	Decision Tree Classifier	0.789474	0.681319
3	Linear SVC	0.880886	0.780220
4	Kernelized SVC	0.850416	0.791209
5	Random Forest Classifier	0.883657	0.747253

After performing Principal Component Analysis, when we trained our models, we found no improvement in KNN results, also, Random Forest too did not yield better results. However, we obtained improvised results from Kernelized SVM model with an accuracy of **80.21%**.



	Model	Train Accuracy	Test Accuracy
0	KNN Classifier	0.648199	0.648352
1	Logestic Regression	0.939058	0.780220
2	Decision Tree Classifier	0.789474	0.681319
3	Linear SVC	0.880886	0.780220
4	Kernelized SVC	0.850416	0.791209
5	Random Forest Classifier	0.883657	0.747253
6	KNN Classifier with PCA	0.645429	0.648352
7	Logestic Regression with PCA	0.783934	0.791209
8	Linear SVC with PCA	0.808864	0.802198
9	kernelized SVC with PCA	0.839335	0.802198
10	DecisionTree Classifier with PCA	0.753463	0.604396
11	Random Forest Classifier with PCA	1.000000	0.681319

Let us visualize our final results.



CONCLUSION

This study proposes a method for classification of arrhythmia using ECG data by implementing various machine learning techniques. After cleaning and pre-processing, the data is modelled using multiple machine learning algorithms like K-Nearest Neighbors, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Linear SVM, and Kernelized SVM. To improve the accuracy of the model, Principal Component Analysis (PCA) is performed over the data to reduce its dimensions and then the data is modelled. The results show that Kernelized SVM outperformed other classifiers when data with reduced features is trained, as PCA reduced the complexity of the original data. The model predicts the absence or presence of cardiac arrhythmia and classifies it into one of the 16 classes with an accuracy of 80.21%. Our results suggest that Kernelized SVM model can be used to diagnose cardio-vascular diseases like arrhythmia in hospitals.

REFERENCES

[1] V. Mahesh, A. Kandaswamy, C. Vimal, and B. Sathish (2009) ECG Arrhythmia Classification Based on Logistic Model Tree

[2] Anam Mustaqeem, Syed Muhammad Anwar, and Muahammad Majid (2018) Multi-Class Classification of Cardiac Arrhythmia Using Feature Selection and SVM Invariants

[3] Babak Mohammadzadeh, Seyed Kamaledin Setarehdan, & Maryam Mohebbi (2008) Support Vector Machine – based arrhythmia classification using reduced features of heart rate variability signal

[4] Jalal A. Nasiri, Mahmoud Naghibzadeh, H. Sadoghi Yazdi, and Bahram Naghibzadeh (2009) ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm