

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



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



Department of computer Engineering and Applications, GLA University
17 km. Stone NH2, Mathura-Delhi Road, P.O. – Chaumuhan, Mathura, 281406

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

Arrhythmia is a cardio-vascular disease, which when not treated well before time could lead to serious health issues in patients. An early diagnosis of this life-threatening disease would help save the lives of millions of people around us. In this study, an idea is proposed to classify patients into one of the sixteen classes, in which one class represents the case of normal people or those who do not have the disease and the other fifteen classes represent ECG records of various subtypes of arrhythmias. This study is carried out on the dataset taken from the University of California at Irvine Machine Learning (UCI ML) Data Repository. This data set contains huge amount of feature dimensions, which included the records of ECG signals. These variables 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 or absence of arrhythmias as well as to classify them into one of the available classes. 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 used with an accuracy rate of 80.21%.

TABLE OF CONTENTS

Declaration	i
Certificate	ii
Acknowledgment	iii
Abstract	iv
Chapter 1 – Introduction	1
1.1 Motivation and Overview	1
1.2 Objective	1
1.3 Contribution	2
1.4 Organization of Project Report	2
Chapter 2 – Literature Review	3
Chapter 3 – Proposed Work	6
3.1 Data Pre-Processing	6
3.2 Exploratory Data Analysis	8
3.3 Data Modelling	11
Chapter 4 – Implementation and Result Analysis	15
Chapter 5 – Conclusion	18
References	19

INTRODUCTION

1.1 MOTIVATION & OVERVIEW:

A healthy heart beats at a rate of 60-100 beats per minute (bpm). If the heart beats at a rate less than 60 bpm or at a rate more than 100 bpm, it is not considered to be a healthy one. In biological terms, this happens when one is suffering from Cardiac Arrhythmia, which is a cardio-vascular disease. A proper medical assistance to arrhythmic hearts can save lives of millions as this could lead to sudden death of an individual.

Cardiac Arrhythmia is an irregularity in heart rhythms which can be life-threatening if not detected early. It disturbs the regular rhythm of the electrical activity of our heart and causes it to beat either too slow or too fast, and causes nonsequential movement of electrocardiogram signals.

The symptoms of cardiac arrhythmia generally include insufficient pumping of blood, chest pain, shortness of breath, fatigue and unconsciousness in an individual. It is detected using the electrocardiogram signals (ECG). It has two broad categories, namely, Bradycardia and Tachycardia.

Bradycardia makes the heart beat too slow, usually below the rate of 60 bpm, while Tachycardia makes the heart beat faster which could go above 100 bpm.

Due to its non-invasive nature, an electrocardiogram (ECG), which measures the electric activity of the heart, is used widely for detecting heart diseases. It is possible to detect its abnormalities by analyzing the electrical signal of each heartbeat. The electrical signals include the action and impulse waveforms which are produced by different cardiac tissues present in the heart. These ECG signals normally consists of P waves, T waves, and QRS complex. The major parameters which are required to examine patients with cardio-vascular diseases are time duration, shape, and the relationship between P wave, QRS complex, T wave, and R-R interval. And any significant change in these waves indicate an ailment of the heart that may have occurred due to certain specific reasons.

1.2 OBJECTIVE:

In our study, we are going to detect the presence and absence of a cardio-vascular disease - Cardiac Arrhythmia, and also classify it in one of the 16 groups. To classify it into the available classes, till now there exists a program. However, there are some differences between the cardio log's and the programs classification. Taking the cardio log's as a gold standard, in this study, we aim to minimize this difference between the classifications 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 about 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

To develop an automated model for classification of cardio-vascular diseases like Arrhythmia, various methods have been proposed. 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 by Babak 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.

A paper by Vasu et al combined SVM and Random Forest classifiers to classify arrhythmias which resulted in a generalization error of 77.4% [5]. Also, the paper presented the implementation of techniques used by contemporary papers on the dataset.

A research paper presented by Guilia and Manas et al explored tree classifiers, SVM, Naïve Bayes and Random Forest algorithms [6]. They performed visualization of 2D

plots for SVM with different types of kernels and found that SVM model with 2 classes and 11 features outperformed other models with an average accuracy of 86%.

A paper by Jadhav et al proposed an Artificial Neural Network model for arrhythmic classification [7]. It used Multi-layer perceptron feedforward neural network with static backpropagation to classify data into normal and abnormal classes. The overall accuracy of the model was 86.67%.

A paper published by Shraddha et al in ICCIDS 2018 [8] proposed a Recurrent Neural Network based classification model to separate normal and abnormal beats. In the approach, various different RNN models were quantitatively compared and the accuracies are calculated with 88.1% being the maximum.

A similar approach as Babak et al [3] is presented by Mi Hye et al [9] to classify arrhythmia by reduction in features by Linear Discriminant Analysis (LDA) and SVM classifier. With even smaller learning data available, SVM classifier with reduced dimensions performed better than Multilayer Perceptron classifier.

Osowski et al [10] presented a way to combine neural network classifiers into one ensemble system for the classification. Modifies bayes method, majority voting and Kullback-Leibler divergence methods are used to achieve better accuracy results.

A paper by Saleha et al used Nearest Neighbors, Decision Tree and Naïve Bayes Classifiers. They improved the accuracy of KNN model from 53% to 66.96%. The accuracy of Naïve Bayes was least among all [11].

For the classification of arrhythmia, Prajwal et al [12] used Principal Component Analysis (PCA) for dimensionality reduction, Bag of Visual words approach for clustering, and algorithms like SVM, KNN, Random Forest, and Logistic Regression are compared. SVM turned out to be the best model.

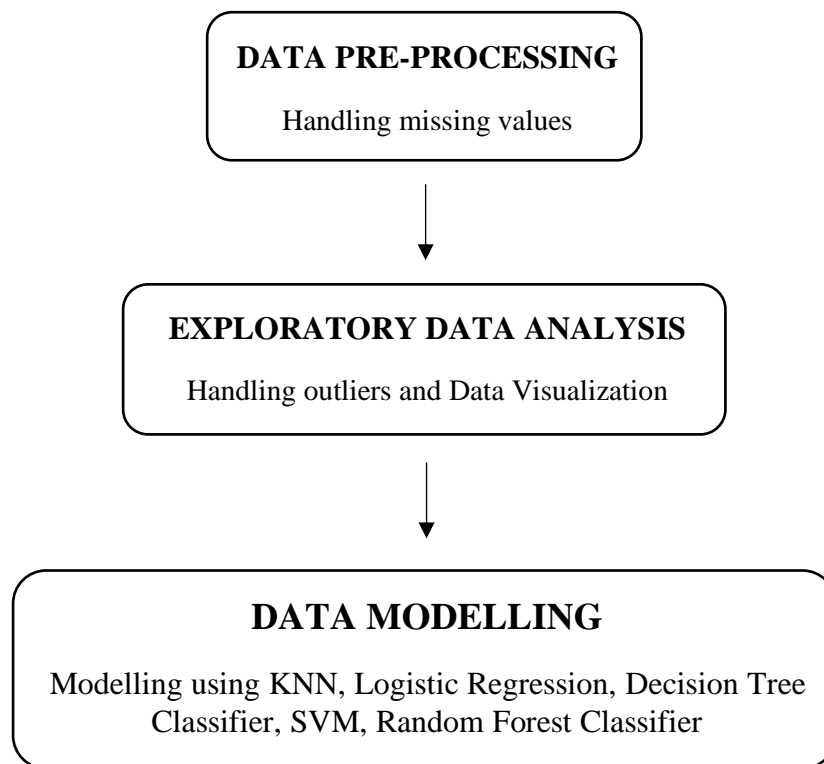
Tae et al [13] used a deep two-dimensional convolutional neural network (CNN) approach and compared AlexNet and VGGNet models for accuracy. The classifier achieved 99.05% accuracy with 97.85% sensitivity.

Another approach by Ru-San et al [14] proposed a new approach based on long-duration ECG fragments and employed 1D Convolutional Neural Network (CNN) for the classification. The overall accuracy obtained turned out to be 91.33%.

Ru-San et al [15] further increased the model accuracy by applying LSTM and CAE-LSTM methods. Convolutional autoencoders reduced the signal sizes of arrhythmic heart beats. This approach showed significant improvement in time cost of LSTM networks and the accuracy was over 99.0%.

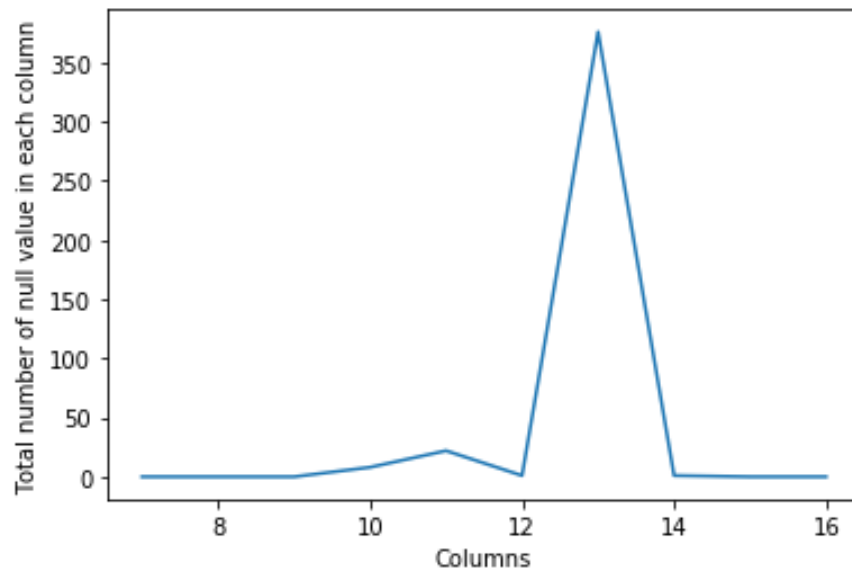
PROPOSED WORK

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.



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 more than 350 missing values.



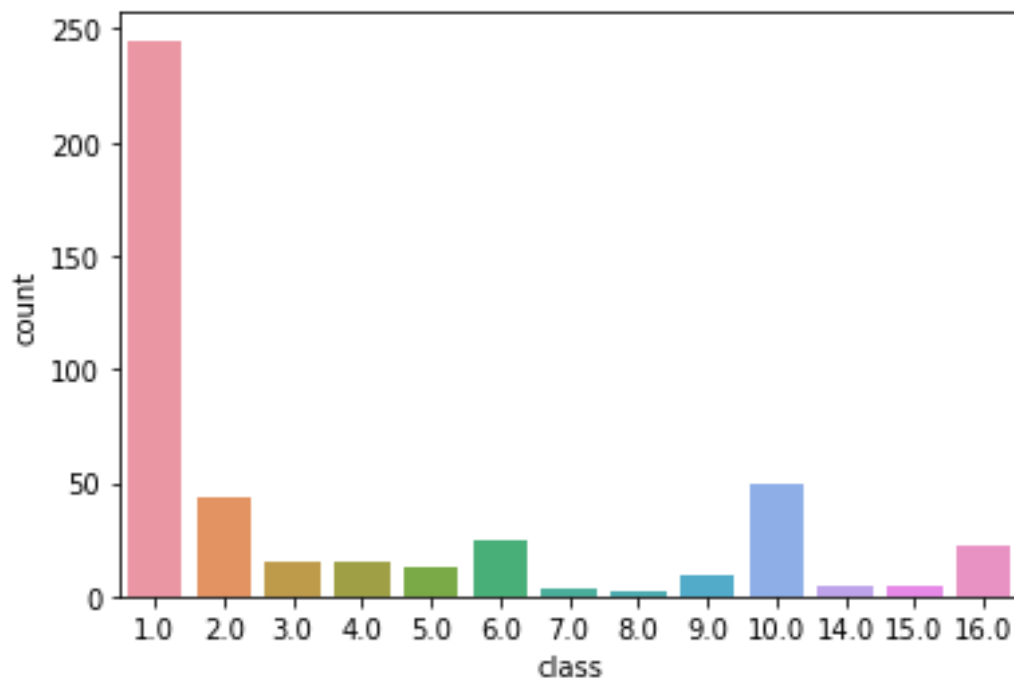
So, we dropped that column and imputed the other columns containing missing data using their mean values. To make our data more understandable, we replaced the column names with appropriate names as per the details given in the UCI machine learning repository.

Out of the 452 observations, 245 observations were of normal people, that means those who do not have an arrhythmic heart. We have 12 different types of arrhythmia. Among these 12 types, the most representative ones are the 'coronary artery disease' and 'Right bundle branch block'.

The features included age, sex, weight, height of individuals and other related information. We explicitly observed that the number of features were relatively higher when compared to the number of observations in the dataset. Finally, we separated the target attribute from the features of our data.

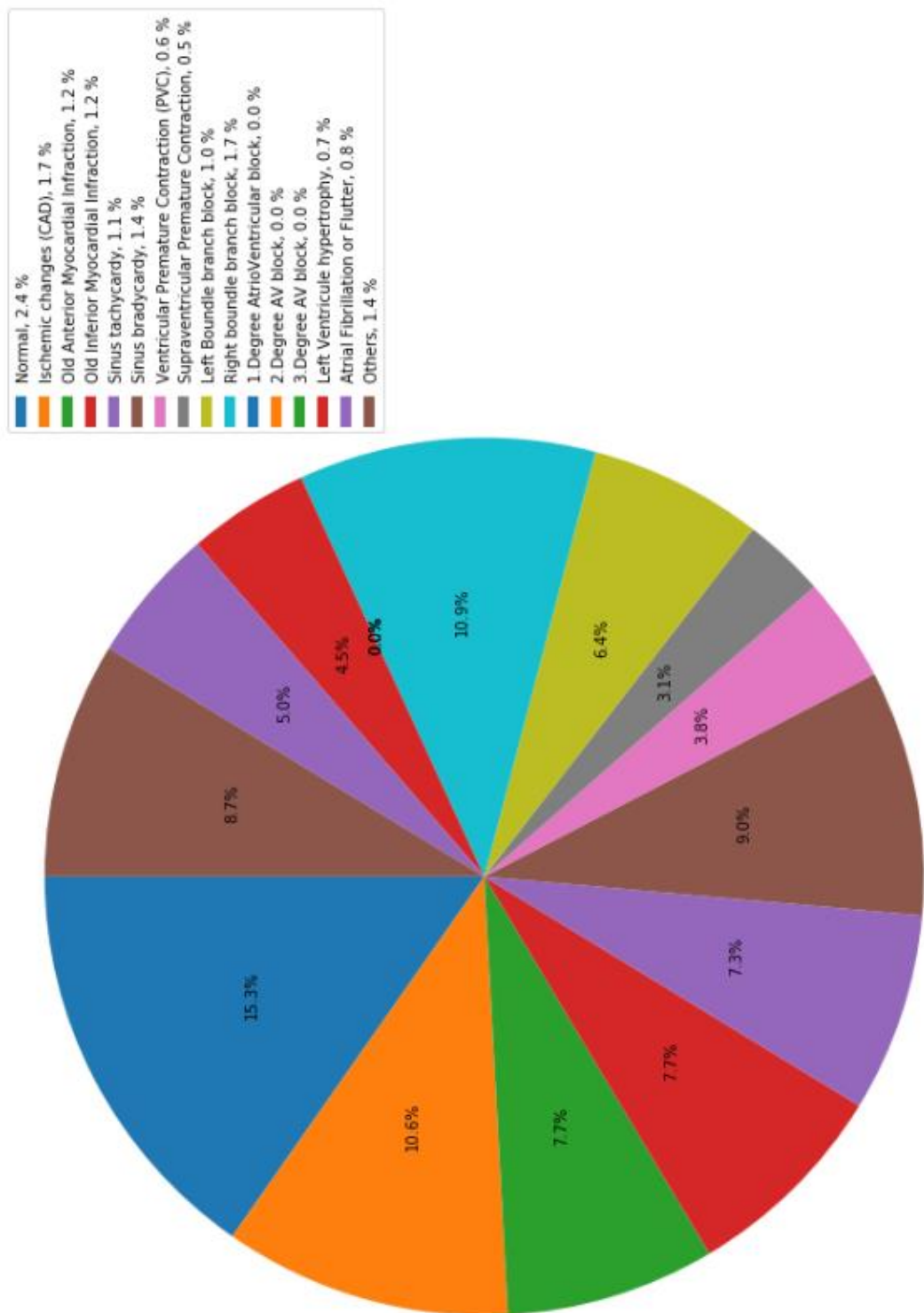
3.2 EXPLORATORY DATA ANALYSIS:

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

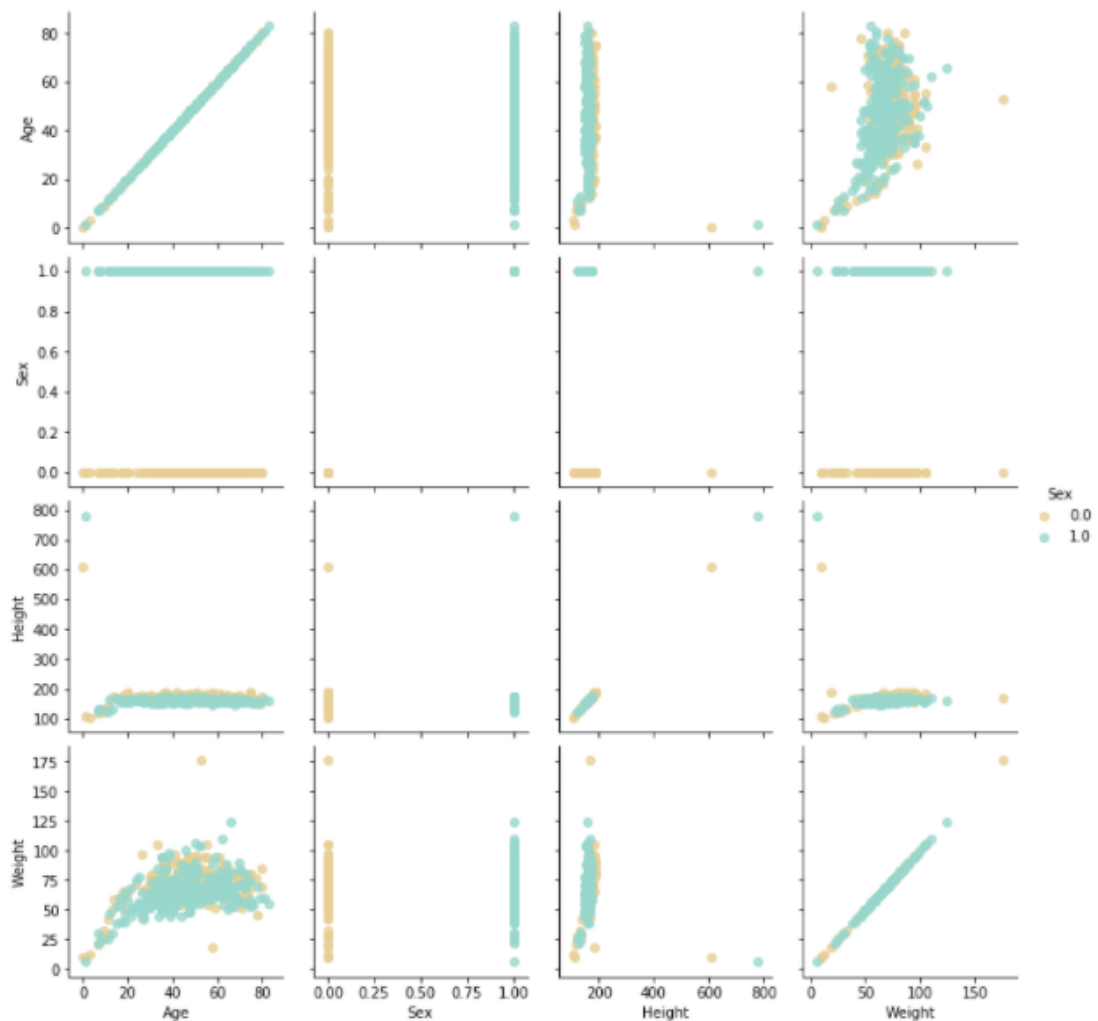


Out of 452 samples, 245 were of class 1 which is for 'normal people'. Also, Atrio-Ventricular block Arrhythmia is not available in the dataset. The samples of classes 7 and 8 are also very few, making our dataset highly imbalance.

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



In order to find the outliers, we visualized the pairwise distribution of a few features and found that the features like height and weight contained outliers and replaced those with their expected values.



The outliers of other features were visualized using boxplots and bar plots. The outliers showed some kind of similarity. We did not remove these outliers assuming they might belong to the classes with few instances.

We then perform feature scaling using standard scaler from sklearn library 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:

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

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

$$\Pr(Y = c | \vec{X} = x) = \frac{e^{\beta^{(c)}_0 + x \cdot \beta^{(c)}}}{\sum_c e^{\beta^{(c)}_0 + x \cdot \beta^{(c)}}}$$

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, p_i 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 form

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

We focus on the soft-margin classifier since choosing a sufficiently small value for λ 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.

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

To increase the accuracies of our model, we applied Principal Component Analysis (PCA) to our dataset.

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.

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

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.

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

Since, it is a classification problem, we will compare the precision and recall values to evaluate our models. We will maximize the Sensitivity i.e. the percentage of unhealthy people who are correctly identified as unhealthy ones.

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

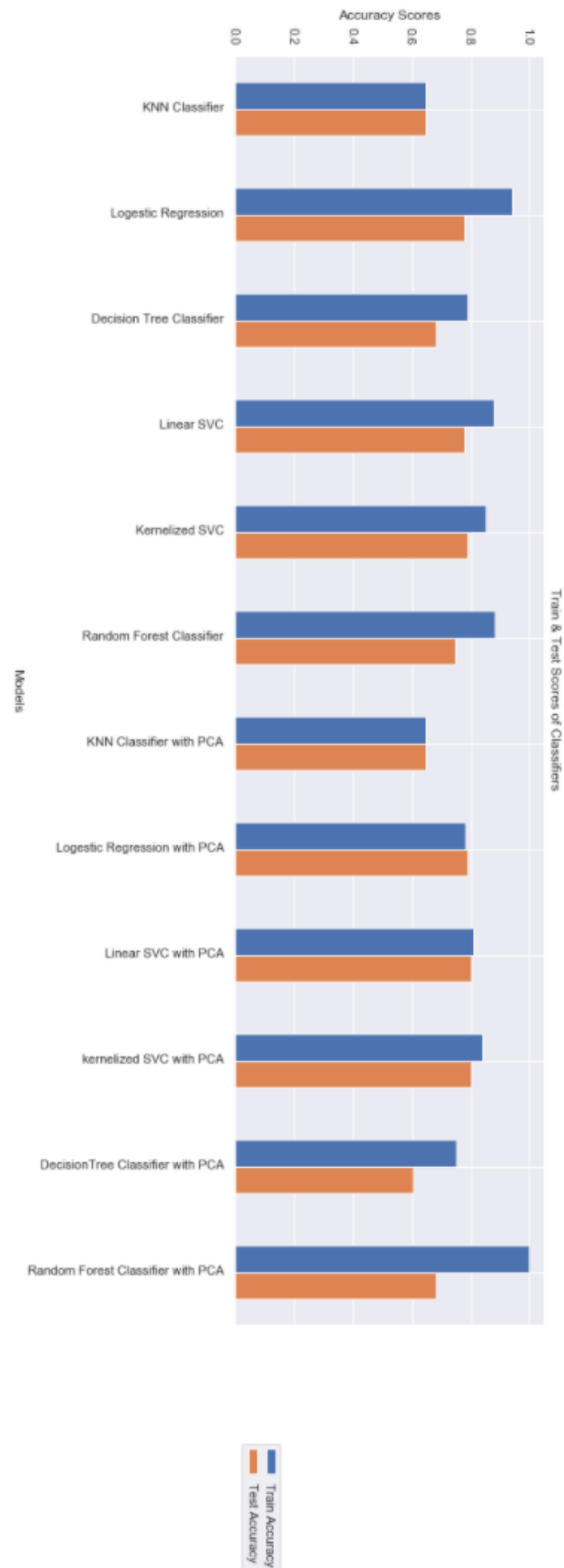
After performing Principal Component Analysis, when we trained our models, we found no improvement in KNN results, also, Random Forest too did not yield better results.

However, we obtained improvised results from Kernelized SVM model with an accuracy of **80.21%**.

Below, we can see the training and testing accuracies of all the models.

	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

Now, 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
- [5] Vasu Gupta, Sharan Srinivasan, Sneha S Kudli (2012) Prediction and Classification of Cardiac Arrhythmia
- [6] Guilia Guidi and Manas Karandikar (2014) Classification of Arrhythmia using ECG Data
- [7] S. M. Jadhav, S. L. Nalbalwar, Ashok Ghatol (2010) Artificial Neural Network based Cardiac Arrhythmia Classification using ECG data
- [8] Shraddha Singh, Saroj Kumar Pandey, Urja Pawar, Rekh Ram Janghel (2018) Classification of ECG Arrhythmia using Recurrent Neural Networks
- [9] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, Sun Kook Yoo (2005) Support Vector Machine based Arrhythmia Classification using Reduced Features
- [10] S. Osowski, T. Markiewicz, L. Tran Hoai (2008) Recognition and Classification system of arrhythmia using ensemble of neural networks
- [11] Saleha Samad, Shoab A. Khan, Anam Haq, Amna Riaz (2014) Classification of Arrhythmia
- [12] Prajwal Shimpi, Sanskruti Shah, Maitri Shroff, Anand Godbole (2017) A Machine Learning approach for the Classification of Arrhythmia
- [13] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyoun Kang, Dohyoun Kim, Daeyoung Kim, Young-Hak Kim (2018) ECG arrhythmia classification using a 2D Convolutional Network
- [14] Ozal Yaldirim, Pawel Plawiak, Ru-San Tan, U. Rajendra Acharya (2018) Arrhythmia Detection using Deep Convolutional Neural Network with long duration ECG signals
- [15] Ozal Yaldirim, Ulas Baran Baloglu, Ru-San Tan, Edward J. Ciaccio, U. Rajendra Acharya (2019) A new approach for arrhythmia classification using Deep coded features and LSTM networks