Classification of Arrhythmia Using ECG Data

Ashish Sharma[1], Shivani Chauhan[2], Sarvesh Kumar Sharma[3], Sachi Tripathi[4], Satyam Kumar Jha[5]

[1, 2] Department of Computer Engineering & Applications, G.L.A. University, Mathura (U.P.), India

Abstract:

Arrhythmia is a cardiovascular disease, which when not treated well before time could lead to consequential health ailments in patients. So, 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 categorize patients into one of the given sixteen classes, where the first class represents the case of normal people or those who do not have the disease and the rest fifteen classes represent ECG records of various other types of arrhythmias. This study is implemented on the dataset from the UCI ML Data Repository. This data set consists of a huge amount of feature dimensions, which included the records of ECG signals. These variables are reduced using dimensionality reduction techniques. 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 determine 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 surpassed the other classifiers used with an accuracy rate of 80.21%.

1. INTRODUCTION

A healthy human heart beats at a rate of 60-100 bpm. If the heartbeats at a rate less than 60 bpm or 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 cardiovascular disease. Proper medical assistance to arrhythmic hearts can save the lives of millions as this could lead to the sudden death of an individual.

Cardiac Arrhythmia is an irregularity in heart rhythms that 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 very slow or very fast, and causes nonsequential movement of electrocardiogram signals.

The symptoms of cardiac arrhythmia generally include inadequate pumping of blood, chest pain, asthmatic symptoms, fatigue, and loss of conscious in an individual. It is detected using electrocardiogram signals (ECG). It has two main categories, namely, Bradycardia and Tachycardia. Bradycardia makes the heartbeat too slow, which means below the rate of 60 bpm, while in Tachycardia, the heart beats faster which means above 100 bpm.

An electrocardiogram (ECG), which measures the electric activity of the heart of an individual, is used widely for recognizing heart diseases. It is possible to detect its abnormalities by examining the electrical signal of each heartbeat.

The electrical signals include the action and impulse waveforms which are induced by distinct cardiac tissues present in our heart. These ECG signals generally consist of P waves, T waves, and QRS complex. The major criterion which are needed to examine patients with cardiovascular diseases are time duration of waves, its shape, and the relationship between P wave, QRS complex, T wave, and R-R interval. And any significant change in the waves indicates an issue of the heart that may have occurred due to certain specific reasons.

In our study, we are going to determine the presence and absence of a cardio-vascular disease - Cardiac Arrhythmia, and also categorize it in one of the given 16 groups. To classify it into the available classes, till now there exists a program. However, there are some variations between the cardio log’s and the program's classification. In this study, we aim to reduce the difference between the classifications utilizing machine learning techniques by considering cardio log’s as a standard.

The paper 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.

2. RELATED WORK

To develop an automated model for the classification of cardiovascular diseases like Arrhythmia, various methods have been proposed. Significant details in the ECG data are seen in the intervals and amplitudes of the characteristic waves. Any irregularity in the shape and duration of the wave feature is considered to be 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] which consists of 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 algorithm put forward by Babak is bottomed 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 boost 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 the 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, 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 a single ensemble system for categorization. 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 the 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 technique for clustering, and algorithms like SVM, KNN, Random Forest, and Logistic Regression are compared. SVM turned out to be the best model.

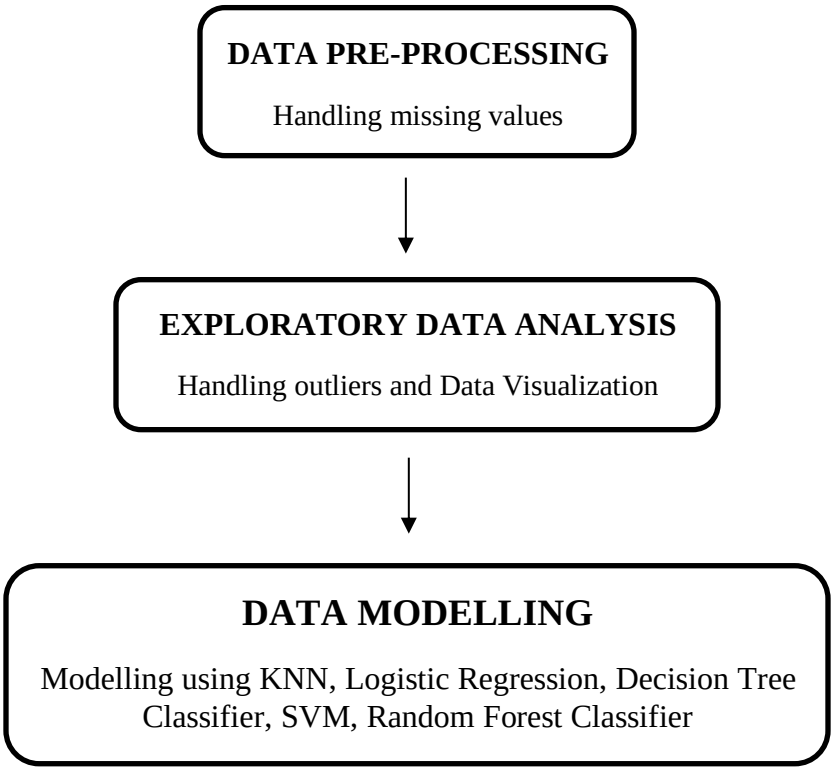
Tae et al [13] used a deep 2D convolutional neural network (CNN) approach and compared AlexNet and VGGNet models for accuracy. The classifier accomplished 99.05% accuracy along with 97.85% sensitivity.

Another approach by Ru-San et al [14] put forth a new strategy 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 heartbeats. This approach showed significant improvement in the time cost of LSTM networks and the accuracy was over 99.0.

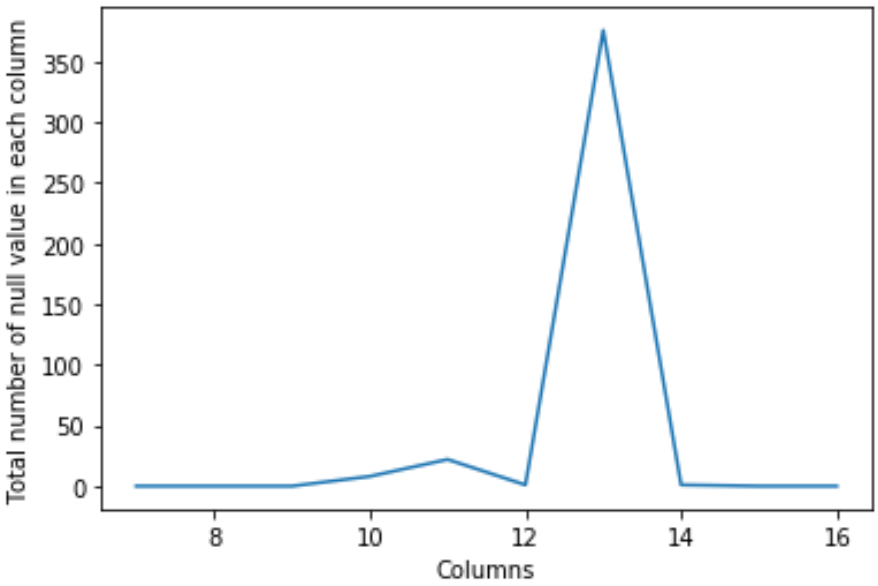
3. PROPOSED METHOD

For our research, we have taken the dataset from the UCI machine learning repository. The proposed work includes cleaning the dataset, visualizing the features, and modeling the data. The first phase includes modeling the data using all the 278 features. During the second phase, only the important features are modeled 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.



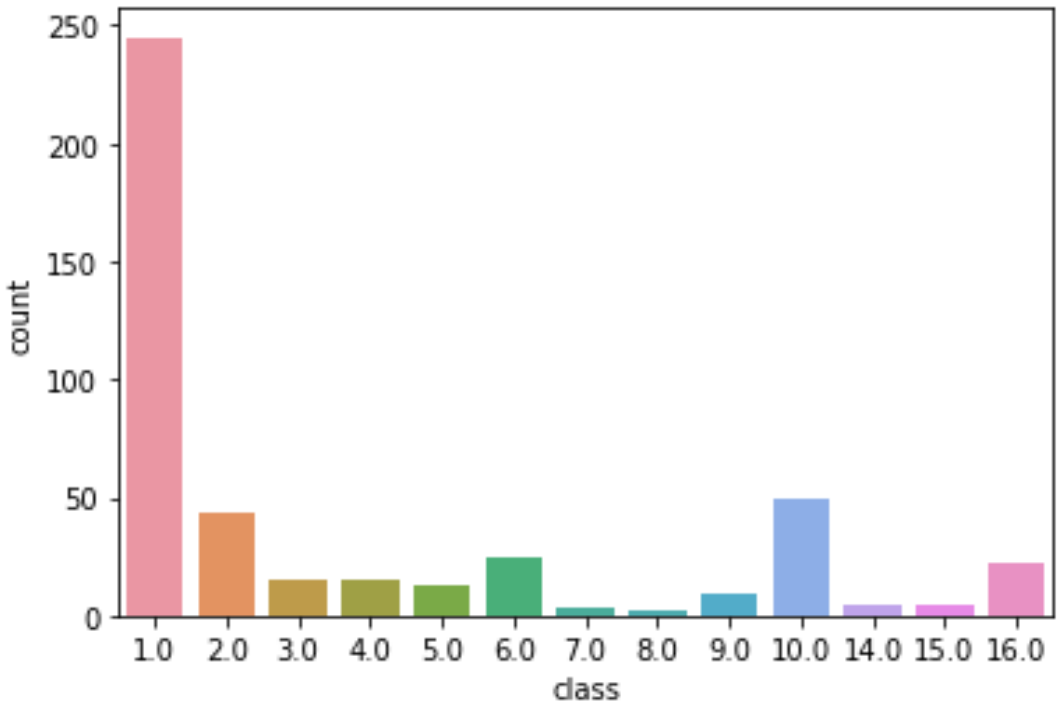
So, we dropped that column and imputed the other columns containing missing data using their mean values. To make our data more understandable, we replaced the column names with appropriates names as per the details given in the UCI machine learning repository.

Out of the 452 observations, 245 observations were of normal people, which 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, the height of individuals, and other related information. We explicitly observed that the number of features was relatively higher when compared to the number of observations in the dataset. Finally, we separated the target attribute from the features of our data.

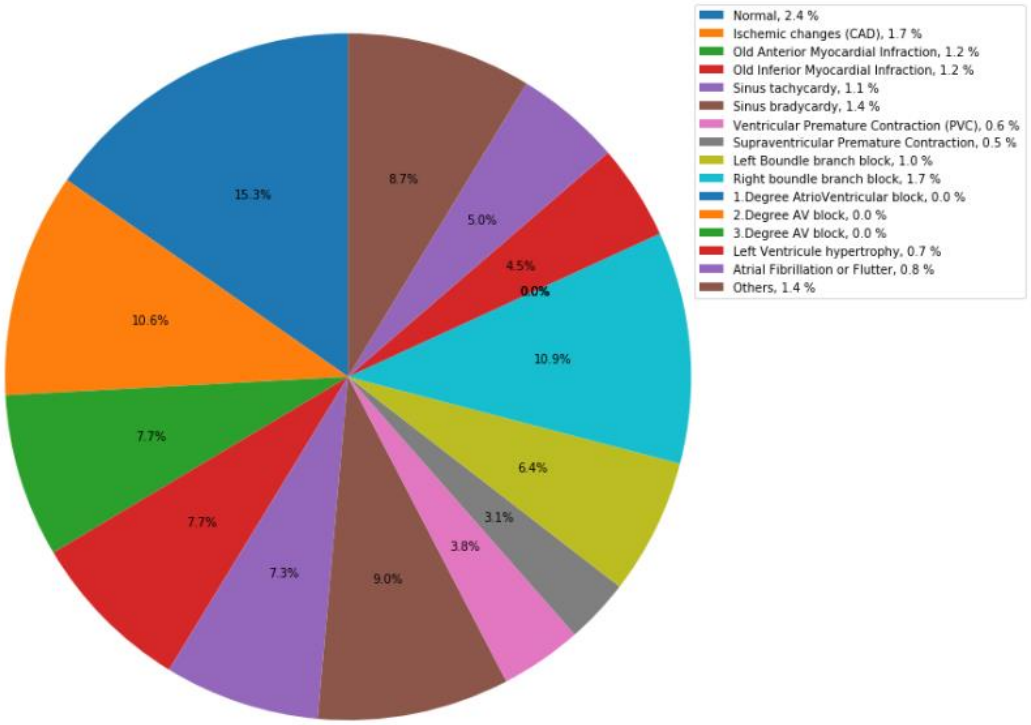
3.2 EXPLORATORY DATA ANALYSIS:

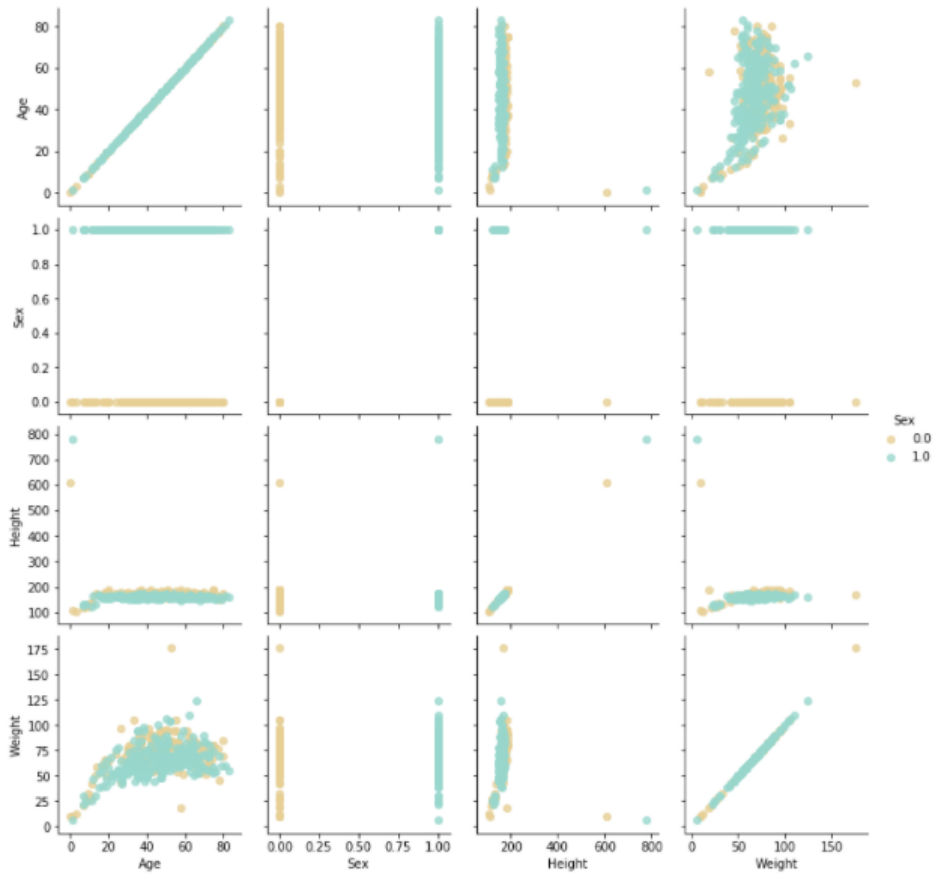
The count of each class in our dataset is plotted using the count plot as shown.



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 imbalanced.

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


Normal, 2.4 %
Ischemic changes (CAD), 1.7 %
Old Anterior Myocardial Infraction, 1.2 %
Old Inferior Myocardial Infraction, 1.2 %
Sinus tachycardy, 1.1 %
Sinus bradycardy, 1.4 %
Ventricular Premature Contraction (PVC), 0.6 %
Supraventricular Premature Contraction, 0.5 %
Left Boundle branch block, 1.0 %
Right boundle branch block, 1.7 %
1.Degree AtrioVentricular block, 0.0 %
2.Degree AV block, 0.0 %
3.Degree AV block, 0.0 %
Left Ventricule hypertrophy, 0.7 %
Atrial Fibrillation or Flutter, 0.8 %
Others, 1.4 %
15.3%
8.7%
5.0%
4.5%
0.0%
10.9%
10.6%
6.4%
7.7%
3.1%
7.7%
3.8%
7.3%
9.0%


Age
Sex
Height
Weight
Sex
0.0
1.0

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.

4. MATHEMATICAL MODEL

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 a method known as 'feature similarity' in order to forecast the values of latest datapoints which implies the new data points are going to be assigned values according to how exactly it matches the points within the training set.

The K value determines the number of the closest neighbors which are needed to be observed to work out the property of the unknown point.

The algorithm employs a metric to work out the closest neighbors. The similarity metric used is mostly Euclidean distance between the unknown point and the other points in the dataset. The formula for calculating 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 a single observation and p_1 to p_n represent the attribute values for the opposite observation.

LOGISTIC REGRESSION:

If Y takes on over two values like in our case, say k of them, we can still use logistic regression.

In logistic regression, every class c in 0: (k-1) has its offset β(c)0 and vector β(c), apart from having a set of parameters β0, β, and also the calculated conditional probabilities are going to 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 algorithm comprises of a tree in which a node represents a feature (attribute), each link or each branch represents a call (rule) and each leaf gives an outcome (categorical or continuous value). The decision tree algorithm learns a method to excellently split the dataset into small subsets to forecast the target value.

The method of splitting the dataset continues till no additional gain is formed or a predetermined rule is attained, e.g., the foremost depth of the tree is gained.

RANDOM FOREST:

Random Forest consists of numerous decision trees with alike nodes but with dissimilar data that results in distinct leaves. It then unites the solutions of multiple decision trees to search out a solution, which represents the common choice of all the decision trees. Gini-index is commonly accustomed to how branching is finished by nodes in an exceeding decision tree.

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

This formula uses the category and chance to work out the Gini of every branch on a node, a determinant that of the branches is a lot of seemingly to occur. Here, pi represents the ratio of the category we tend an observant within the dataset and c represents the number of categories.

SUPPORT VECTOR MACHINE:

The motive of SVM is to hunt out the optimal hyperplane which segregates the data points into two components linearly by maximizing the margin. The point which is above or on the hyperplane is classified as class +1, and also that below the hyperplane is classified as class -1. The (soft margin) SVM Classifier is calculated using the given expression:

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

We specialize in the soft-margin classifier since selecting a less value for lambda produces the hard-margin classifier for a linearly-classifiable data.

The kernel mechanism maps the details into higher dimensional space anticipating that during this mapping the details might have become a lot of simply separated or higher structured. There are no constraints on the shape of this mapping, which may even emerge into infinite-dimensional space. The Sigmoid Kernel (Hyperbolic Tangent) is derived from the Neural Networks concepts, where the bipolar sigmoid operate is usually used as associate degree activation operate for artificial neurons.

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

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

PRINCIPAL COMPONENT ANALYSIS:

To scale back the spatial property of the dataset consisting of huge inter-related variables, whereas holding the maximum amount as attainable of the variation gift within the dataset, we use PCA. The options are reworked into a new set of variables referred to as principal parts, that are unrelated. The variance could be a metric measured between 2 variables. It offers a live off however changes in one-dimension affect changes within the alternative.

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

An eigenvector, which is also known as a characteristic vector of a linear transformation, may be a nonzero vector that changes at the foremost by a scalar unit once that linear transformation is applied thereto. The eigenvalue can be found out as:

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

Since it's a classification drawback, we'll compare the exactness and recall values to judge our models. we'll maximize the Sensitivity i.e., the proportion of unhealthy those who are properly known as unhealthy ones.

5. RESULTS

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

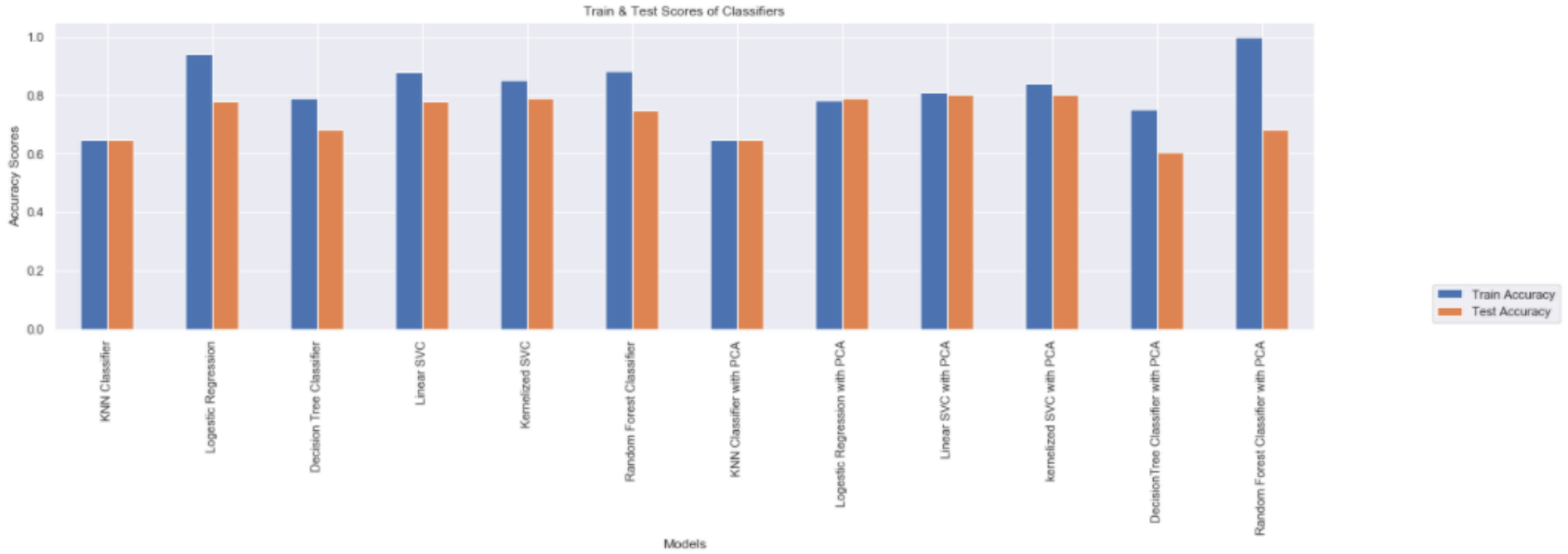
After performing Principal Component Analysis, when we trained our models, we found no improvement in KNN results, also, Random Forest too did not yield better results.

However, we obtained improvised results from the 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.


Train & Test Scores of Classifiers
Accuracy Scores
1.0
0.8
0.6
0.4
0.2
0.0
KNN Classifier
Logestic Regression
Decision Tree Classifier
Linear SVC
Kernelized SVC
Random Forest Classifier
KNN Classifier with PCA
Logestic Regression with PCA
Linear SVC with PCA
kernelized SVC with PCA
DecisionTree Classifier with PCA
Random Forest Classifier with PCA
Models
Train Accuracy
Test Accuracy

6. DISCUSSION & CONCLUSION

This study suggests a strategy for the categorization of arrhythmia using ECG data by implementing various machine learning techniques. After cleaning and pre-processing, the data is modeled using diverse 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 modeled. The results show that Kernelized SVM outperformed other classifiers when data with reduced features are 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 the Kernelized SVM model can be used to diagnose cardiovascular 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 an 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.