Methods for Predicting the Prevalence of Heart Disease

Kieleh Ngong Ivoline-Clarisse 178229001010
Konya Technical University, Computer Engineering

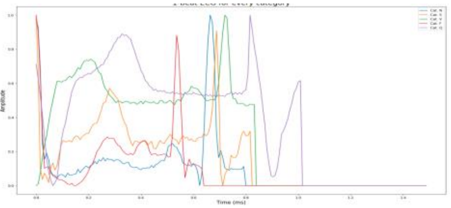
Motivation

Cardiac Arrhythmias is one of the leading causes of cardiac death in the world today.

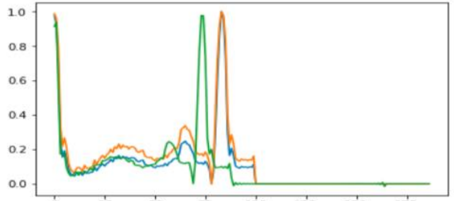
The aim of this study is to automatically classify cardiac arrhythmias and to study the performance of Support Vector Machines with different feature extraction or feature selection techniques.

Data

The model will be trained and tested with data from the MIT-BIH Arrhythmia Dataset. The dataset contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects.The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias infarction. The dateset is made up of **5** categories of Arrythmia, **109446** samples and **187** attributes.

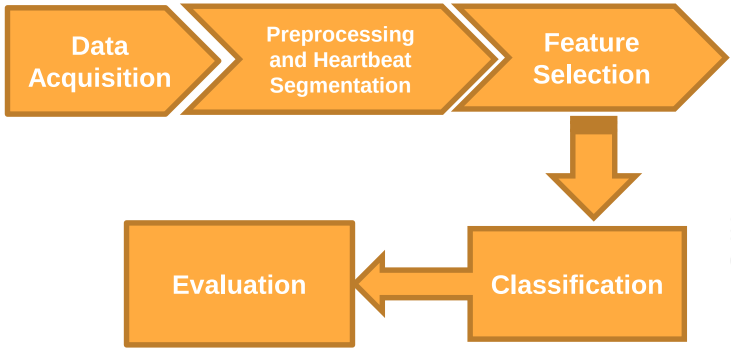


One beat from each category before oversampling



After oversampling

Approach



Methods

The selected data is preprocessed by feature scaling and oversampling to create balanced training data.

4 methods are used to extract/select features and each of the features extracted was classified using SVM.

Principal Component Analysis (PCA)

PCA is a mathematical process that uses linear transformations to map data from high-dimensional fields to low-dimensional fields.

Random Forest

Operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).

Boosted Trees

Learns slowly by sequentially adding shrunken values of the predictions from bootstrapped classification trees fit to the residuals of the previous predictions.

Convolutional Neural Networks (CNN)

A specific type of artificial neural network that uses perceptron, a machine learning unit algorithm, for supervised learning, to analyze data

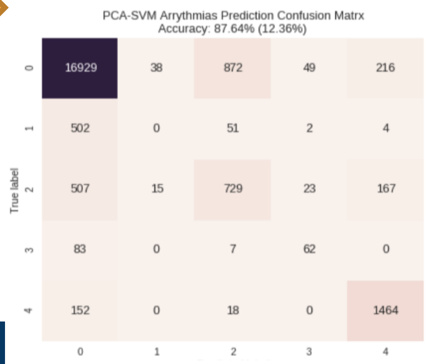
Support Vector Machine (SVM)

Constructs a hyperplane decision boundary in the feature spaces that maximizes the function margin.

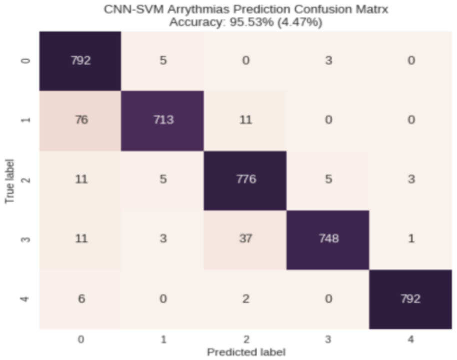
Results

Confusion Matrices For each Feature Extraction Method

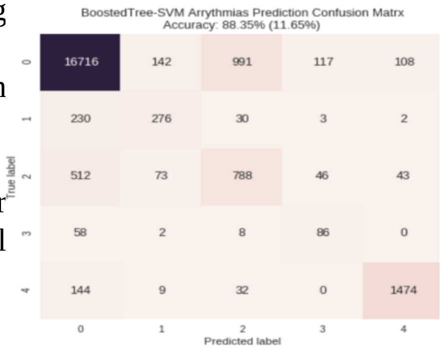
PCA



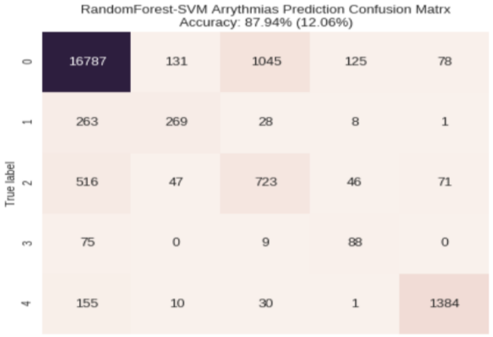
CNN



Boosted Trees



Random Forest



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

Conclusions and Future Work

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. Though this results shows promise, there is much room for improvement. In future steps, techniques to improve these models performances will be applied