A Novel Unified Diagnostics Model for Predicting Risk Levels in Coronary Heart Disease based on Machine Learning Methods

Thendral Puyalnithi, Ashvin Panicker
VIT University, Vellore, TN, India
thendral.p@vit.ac.in, ashvin.panicker@gmail.com

Abstract: **In this paper, we analyse Cardiac Datasets using Supervised Machine Learning methods like decision trees, Naive Bayes algorithm, SVM, k-Nearest Neighbours algorithm and Ensemble Classifiers. We have a proposed a System which uses the following three data sets collected from the UCI Machine Learning Repository, SPECT heart dataset, Heart disease dataset, Echocardiogram dataset and also have created a dataset from the ECG dataset with risk level as class to detect the severity of Heart disease. This system will be a supportive tool for Cardiologists. In addition, we test each classification algorithm and score them based on their precision. We have compared the ensemble methods of Bagging and Boosting with different number of tree estimators, test data sizes and train data sizes and used the accuracy results to arrive at the best classifiers at each stage in the System.**

Keywords - Data Analysis, Data Mining, Machine Learning, Classification, Training, Testing, Prediction, Scoring, Evaluation, Ensemble, Bagging, Boosting

I. INTRODUCTION

Heart disease shares a major part of the global burden of lifestyle diseases. It is now the world's leading causes of death, claiming 17.3 million lives each year. India has seen a rapid transition in its heart disease burden over the past couple of decades. Earlier, non-modifiable factors like age, gender, family history were mainly responsible for heart disease. But over the past few decades, heart disease seems to have surpassed all the boundaries and now controllable risk factors like diet, physical inactivity and stress largely determine the risk of heart disease.[1]

In India, out of the estimated population of more than 1.27 billion dispersed across various geographical regions, about 45 million people suffer from coronary artery disease. According to current estimates, India will soon have the highest number of cases of cardiovascular disease in the world. It is estimated to account for 35.9% deaths by the year 2030.[1]

Cardiac hospitals in India perform over 2,00,000 open heart surgeries per year, one of the highest, worldwide. There has been a steady annual rise to the tune of 25-30 per cent per year in the number of coronary interventions over the past several years.[2]

A heart attack happens when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. If blood flow isn't restored quickly, the section of heart muscle begins to die.[3] Heart attacks most often occur as a result of coronary heart disease (CHD), also called coronary artery disease. CHD is a condition in which a waxy substance called plaque builds up inside the coronary arteries. These arteries supply oxygen-rich blood to your heart. When plaque builds up in the arteries, the condition is called atherosclerosis. The buildup of plaque occurs over many years. Eventually, an area of plaque can rupture (break open) inside of an artery. This causes a blood clot to form on the plaque's surface. If the clot becomes large enough, it can mostly or completely block blood flow through a coronary artery. If the blockage is not treated quickly, the portion of heart muscle fed by the artery begins to die. Healthy heart tissue is replaced with scar tissue. This heart damage may not be obvious, or it may cause severe or long-lasting problems.[3]

The major causes of heart attack are age, angina, cholesterol, diabetes, dietary habits, infection, genes, hypertension, obesity, overweight, physical inactivity, smoking, being HIV positive, stress and calcium supplements

The common methods for diagnosis are ECG (Electrocardiogram), Blood Tests, Echocardiography, Coronary Angiography, Cardiac Catheterization, Single Photon emission computed tomography (SPECT) for heart.[3]

Machine learning algorithms can play a vital role in heart data analysis in the current day. It will help assist physicians in their diagnostics.

Machine learning methods include Supervised (Classification) and Unsupervised (Clustering) algorithms. In our work since we are using labelled data sets, we have used classification techniques to predict. The aim of our work is to predict the severity levels in coronary Heart Disease. We have analysed the performance of Decision Tree, kNN, Naive Bayes, SVM and Ensemble techniques.

Classification and Regression:

Classification is about training a System with labelled data and the labelled data being a class, we will be finding the appropriate class for an unknown set of data, whereas in Regression we are having a labelled attribute which is a continuous value, so we will be training a system to predict a Continuous value. Both the methods need a Trained system and for training a system we need a Labelled or Classified Data set.

Decision Tree: Decision tree learning uses a decision tree as a predictive model, which maps observations about an item to conclusions about the item's target value. A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.[4]
There are many specific decision-tree algorithms. Notable ones include:

- ID3 (Iterative Dichotomiser 3)
  ID3 uses information gain as its attribute selection measure.
- C4.5 (successor of ID3)
  C4.5, a successor of ID3, uses an extension to information gain known as gain ratio.
- CART (Classification And Regression Tree)
  The Gini index is used in CART. [5]

SVM: Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.[4]

kNN: Nearest neighbour classifiers are based on learning by analogy, that is, by comparing a given test with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in an *n*-dimensional space. In this way, all the training tuples are stored in an *n*-dimensional pattern space. When given an unknown tuple, a kNN classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are regarded as the *k* "nearest neighbours" of the unknown tuple.[4]

Naive Bayes: Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class-conditional independence. It is made to simplify the computations involved and, in this sense, is considered "Naive", literally meaning lacking wisdom or inexperienced.[4]

Ensemble Techniques:

An ensemble for classification is a composite model, made up of a combination of classifiers. The individual classifiers each submit their result and a class label prediction is returned by the ensemble based on the collection of results from each individual classifier. Ensembles tend to be more accurate than their component classifiers and are hence often used to improve classification accuracy.[4] The following techniques are used by us in this system:

Bagging- The bagging algorithm creates an ensemble of classification models for a learning scheme where each model gives an equally weighted prediction.[4]
Essentially bagging algorithms build multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction.[5]

Boosting- A boosting algorithm creates an ensemble of classifiers where each one gives a weighted vote. Similar to bagging it resamples the data and assigns higher weights to the incorrectly predicted outcomes and lower weights to the correctly predicted results while maintaining the overall equilibrium of weight throughout the data set

II. PROPOSED SYSTEM

We have proposed a unified diagnostics model for predicting the risk levels in coronary heart disease.
We have used four cardiac datasets for training the System. Three of these datasets are obtained from the UCI Machine learning repository[6], while we have created a separate dataset using the ECG dataset as a background for it.We refer to this dataset as the Generated ECG dataset The details about the three datasets are given below.

A. Dataset Description

1. General Heart Disease Database -

Created by - Hungarian Institute of Cardiology,Budapest

Date of creation - July, 1988

Number of Instances (rows) - 294

Number of Attributes (columns) - 14

Attributes -

- *age* - Age in Years
- *sex* - Gender (1 = male, 0 = female)
- *cp* - Chest Pain Type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- *trestbps* - Resting Blood Pressure (in mm Hg on admission to the hospital)
- *chol* - serum cholesterol in mg/dl
- *fbs* - fasting blood sugar > 120 mg/dl (1 = true; 0 = - false)
- *restecg* - resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality
    - Value 2: showing probable or definite left ventricular hypertrophy
- *thalach* - maximum heart rate achieved
- *exang* - exercise induced angina (1 = yes; 0 = no)
- *oldpeak* - ST depression induced by exercise relative to rest
- *slope* - The slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
- *ca* - number of major vessels (0-3) coloured by fluoroscopy
- *thal* - 3 = normal; 6 = fixed defect; 7 = reversible defect
- *num (class attribute)* - diagnosis of heart disease (angiographic disease status)
    - Value 0: < 50% diameter narrowing (Normal Heart)
    - Value 1: > 50% diameter narrowing (Abnormal Heart)

Missing values were imputed (filled in) by Orange 3 software

2. **SPECT Heart Database -**

Created by - Medical College of Ohio, OH, U.S.A.

Date of creation - January, 2001

Number of Instances (rows) - 80

Number of Attributes (columns) - 23

- *OVERALL_DIAGNOSIS* - 0,1 (**class attribute**,binary)
- *F1* - 0,1 (the partial diagnosis 1, binary)
- *F2* - 0,1 (the partial diagnosis 2, binary)
- *F3* - 0,1 (the partial diagnosis 3, binary)
- *F4* - 0,1 (the partial diagnosis 4, binary)
- *F5* - 0,1 (the partial diagnosis 5, binary)
- *F6* - 0,1 (the partial diagnosis 6, binary)
- *F7* - 0,1 (the partial diagnosis 7, binary)
- *F8* - 0,1 (the partial diagnosis 8, binary)
- *F9* - 0,1 (the partial diagnosis 9, binary)
- *F10* - 0,1 (the partial diagnosis 10, binary)
- *F11* - 0,1 (the partial diagnosis 11, binary)
- *F12* - 0,1 (the partial diagnosis 12, binary)
- *F13* - 0,1 (the partial diagnosis 13, binary)
- *F14* - 0,1 (the partial diagnosis 14, binary)
- *F15* - 0,1 (the partial diagnosis 15, binary)
- *F16* - 0,1 (the partial diagnosis 16, binary)
- *F17* - 0,1 (the partial diagnosis 17, binary)
- *F18* - 0,1 (the partial diagnosis 18, binary)
- *F19* - 0,1 (the partial diagnosis 19, binary)
- *F20* - 0,1 (the partial diagnosis 20, binary)
- *F21* - 0,1 (the partial diagnosis 21, binary)
- *F22* - 0,1 (the partial diagnosis 22, binary)

3. **Generated Echocardiogram Database -**

Number of Instances (rows) - 204

Number of Attributes (columns) - 7

Attributes -

- *fact_shortening* - a measure of contractility around the heart lower numbers are increasingly abnormal
- *EPSS (in mm)* - E-Point Septal Separation, another measure of contractility. Larger numbers are increasingly abnormal.
- *lvdd* - Left Ventricular end-Diastolic Dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
- *wall_motion_score* - A measure of how the segments of the left ventricle are moving.
- *wall_motion_index* - wall_motion_score divided by number of segments seen
- *mult* - a derivate variable which can be ignored
- *class* - risk level of heart attack from 0 - 6 (0 = Lowest risk, 6 = Highest risk)

We have used certain attributes from the Echocardiogram dataset from the UCI machine learning repository to create a dataset with a generated class as risk level of heart attack. We have also generated a separate set of data which pertains to people who have not had a heart attack but may be prone to it. We have done this by assigning random values in the appropriate ranges. When the class variable 1, we have generated tuples for that class where all but 1 attribute is in the range of a person not prone to heart attack. When the class variable is 2, we have generated tuples for that class where all but 2 attributes are in the range of a person not prone to heart attack. We perform similar actions for classes 3 and 4. Likewise for class 5, we have generated tuples where all the attributes are out of the range of a person not prone to a heart attack. For class of 6 we have used the tuples from the ECG dataset since they were all heart attack cases.
We have referred to the normal values for heart data.[9] This makes the dataset optimal for our system.

4. **Echocardiogram Dataset**

Created by - The Reed Institute, Miami, FL, U.S.A

Date of Creation - February, 1989

Number of Instances - 131

Number of Attributes - 11

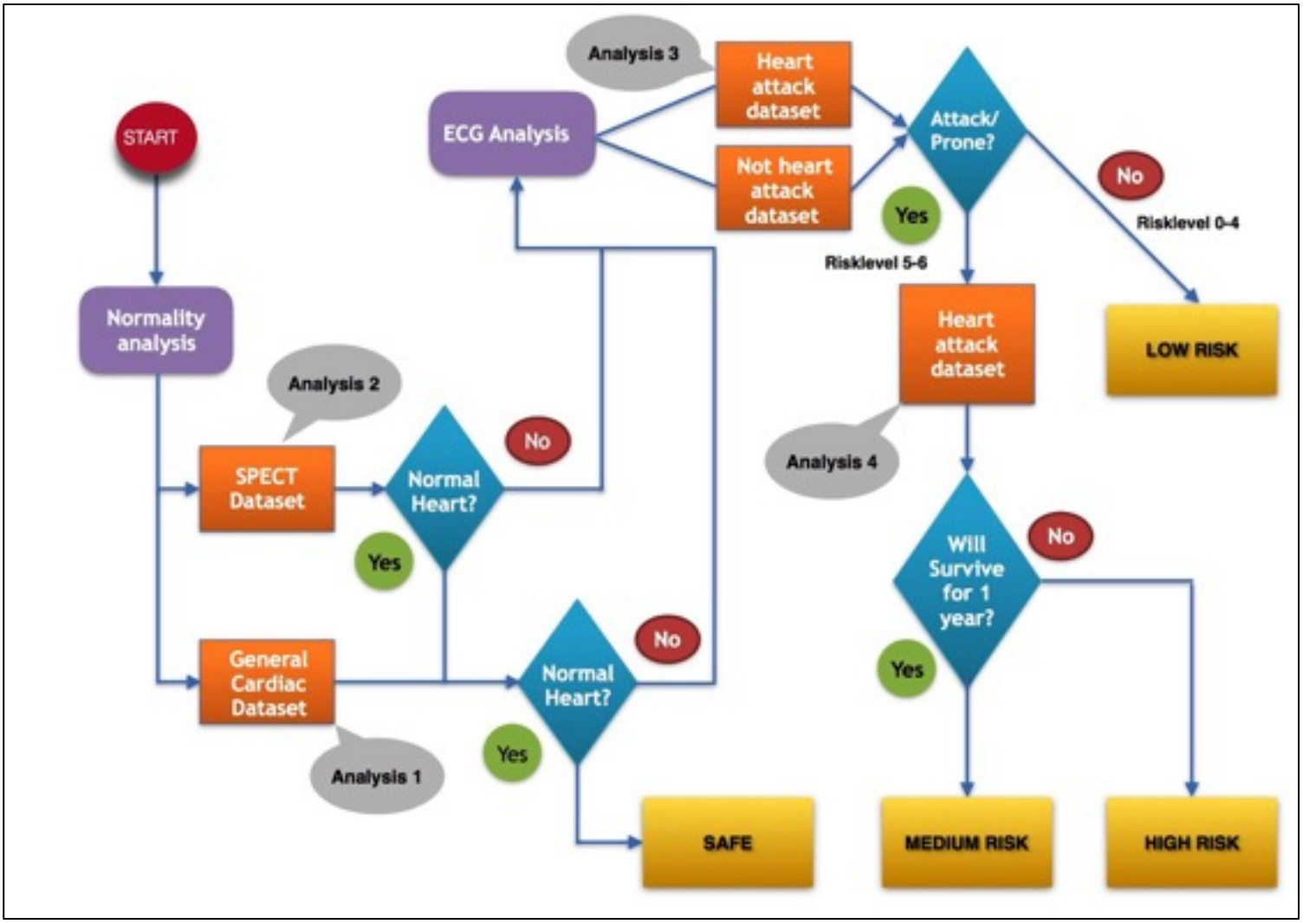
Attributes -

- *survival* - The number of months patient survived (has survived,if patient is still alive). Because all the patients had their heart attacks at different times, it is possible that some patients have survived less than one year but they are still alive. Check the second variable to confirm this. Such patients cannot be used for the prediction task mentioned above.
- *still-alive* - a binary variable. 0 means dead at end of survival period, 1 means still alive.
- *age-at-heart-attack* - age in years when heart attack occurred.
- *pericardial-effusion* - Binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid.
- *fractional-shortening* - A measure of contractility around the heart lower numbers are increasingly abnormal
- *EPSS (in mm)* - E-Point Septal Separation, another measure of contractility. Larger numbers are increasingly abnormal.
- *lvdd* - Left Ventricular end-Diastolic Dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
- *wall_motion_score* - A measure of how the segments of the left ventricle are moving.
- *wall_motion_index* - wall_motion_score divided by number of segments seen
- *mult* - A derivate variable which can be ignored.
- *alive-at-1* - Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

B. Training the system:

The Proposed System has been trained with the above mentioned datasets using both the Orange3 machine learning software[7] and the Scikit package[8] based in python. We have used C4.5, Naive Bayes, k Nearest Neighbour, SVM and Ensemble Classification techniques to train the System. Various Evaluation parameters are used to find the best classification algorithm for every data set. The flowchart of Unified Diagnostics approach is given below.

C. Flowchart of the Proposed Model



Key

- Classifier
- Training Datasets
- Analysis
- Classes

E. Explanation of flowchart

The system starts by checking the normality of the heart by using the SPECT dataset and the General Cardiac dataset as training sets. If both the analysis tests show a positive result then the system predicts that the heart is normal and displays result as SAFE.

In case either one of the analysis gives a negative result then the system predicts that the heart is abnormal and proceeds for ECG analysis of the heart. This section uses the Generated ECG data and analyses the risk level of heart attack, 0-4 being low risk and 5-6 being prone to heart attack. If the risk level is between 0 and 4 then the system displays result as LOW RISK.

Otherwise the system predicts that the person is attack prone and further tries to predict whether he will survive for 1 year. This analysis uses the pure ECG dataset as a training set. In this analysis if a positive result is obtained, the system displays result as MEDIUM RISK.

If a negative result is obtained it displays the result as HIGH RISK and declares that the person will not be alive after 1 year

III. RESULTS AND ANALYSIS

A. Method of Evaluation

In Orange3 we are able to compare and analyse the efficiencies of the different algorithms separately that are used to test and score the data set

Here are the scoring methods that we have considered to determine the efficiency of each individual method of Classification:

AUC - Area Under the Curve using the trapezoidal rule The trapezoidal rule works by approximating the region under the graph of the function as a trapezoid and calculating its area.[7]

Precision- The precision is the ratio **tp / (tp + fp)** where **tp** is the number of true positives and **fp** the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.[7]

General Cardiac Data

	AUC	Precision
ID3	0.935	0.969
kNN	0.732	0.744
SVM	0.783	0.678
NB	0.824	0.786

SPECT Data

	AUC	Precision
ID3	0.724	0.969
kNN	0.711	0.963
SVM	0.695	0.966
NB	0.768	0.983

Generated ECG Data

	AUC	Precision
ID3	0.912	0.893
kNN	0.821	0.815
SVM	1.000	1.000
NB	0.921	0.922

ECG Data

	AUC	Precision
ID3	0.769	0.368
kNN	0.686	0.321
SVM	0.861	0.857
NB	0.686	0.357

Bagging Classifier Evaluation Scores

Data Set 1 (General Cardiac Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.802	0.809	0.802	0.802
	0.6	0.813	0.830	0.836	0.813
	0.7	0.825	0.815	0.820	0.820
	0.8	0.817	0.817	0.813	0.817

Data Set 2 (SPECT Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.724	0.675	0.675	0.675
	0.6	0.770	0.770	0.770	0.770
	0.7	0.732	0.696	0.714	0.696
	0.8	0.671	0.703	0.671	0.703

Data Set 3 (Generated ECG Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.754	0.745	0.764	0.754
	0.6	0.715	0.723	0.731	0.739
	0.7	0.692	0.685	0.706	0.692
	0.8	0.652	0.664	0.658	0.664

AdaBoost Classifier Evaluation Scores

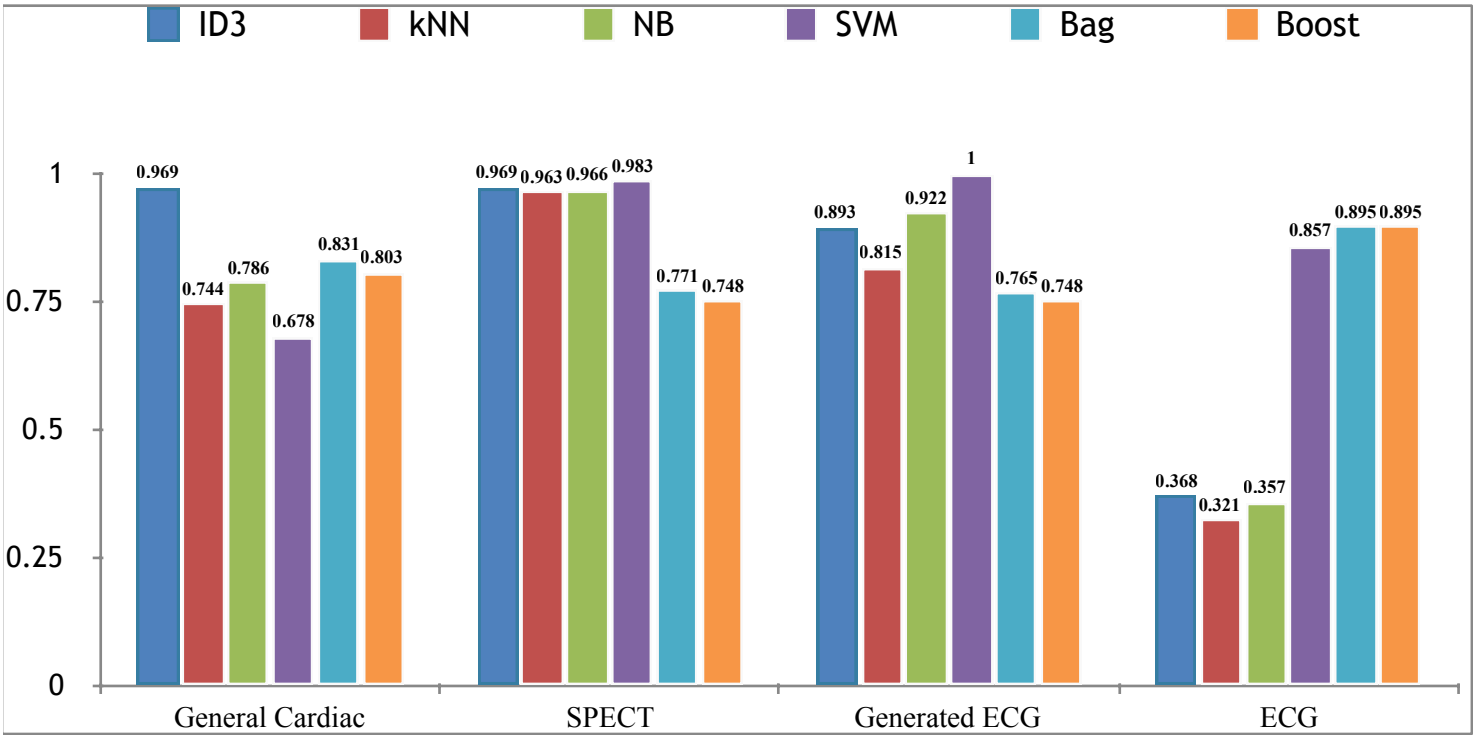
Data Set 1 (General Cardiac Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.802	0.789	0.761	0.782
	0.6	0.768	0.723	0.745	0.745
	0.7	0.791	0.796	0.796	0.776
	0.8	0.758	0.775	0.775	0.754

Data Set 2 (SPECT Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.625	0.625	0.625	0.625
	0.6	0.729	0.729	0.729	0.729
	0.7	0.625	0.625	0.625	0.625
	0.8	0.671	0.671	0.671	0.671

Data Set 3 (Generated ECG Dataset)		No. of models			
		50	**100**	**150**	**200**
Test size	**0.5**	0.696	0.696	0.686	0.696
	0.6	0.642	0.642	0.642	0.642
	0.7	0.748	0.741	0.748	0.741
	0.8	0.658	0.652	0.658	0.658

Data Set 4 (ECG Dataset)		No. of models			
		50	100	150	200
Test size	0.5	0.878	0.878	0.878	0.878
	0.6	0.873	0.873	0.873	0.873
	0.7	0.869	0.869	0.880	0.880
	0.8	0.866	0.895	0.895	0.895

Data Set 4 (ECG Dataset)		No. of models			
		50	100	150	200
Test size	0.5	0.878	0.848	0.863	0.863
	0.6	0.873	0.873	0.873	0.860
	0.7	0.880	0.880	0.880	0.880
	0.8	0.895	0.885	0.885	0.885



IV. CONCLUSION

Finally we have compared the accuracy scores for all the datasets using all the classification techniques discussed earlier and can clearly see the best classifiers for each dataset in the graph above.

V. REFERENCES

[1] http://www.thehealthsite.com/

[2] http://food.ndtv.com/health/

[3] http://www.nhlbi.nih.gov/

[4] Data Mining - Concepts and Techniques (3rd Edition) Jiawei Han, Micheline Kamber

[5] wikipedia.org/wiki/Decision_tree_learning

[6] https://archive.ics.uci.edu/ml/

[7] https://orange.biolab.si

[8] http://scikit-learn.org

[9] http://www.echopedia.org/wiki/Normal_Values_of_TTE