International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Cardio-Vascular Disease Prediction based on Ensemble technique enhanced using Extra Tree Classifier for Feature Selection

Baranidharan B, Abhisikta Pal, Preethi Muruganandam

Abstract: ***Cardio vascular disease is the major death factor in the last decade. Most of the patients diagnosed with the CVD at the later stage die even though advanced treatments are given. Earlier identification of heart disease may reduce the death rate. The cost of medical diagnosis makes it perverse for the large number of people to the early fix of the disease. Machine learning and Data mining techniques are successful in medical diagnosis through non-invasive methods. In developing such models, Feature selection is very important since it affects the accuracy of the diagnosis. In this research, the feature selection is done through Extra Tree classifier method for identifying the most important feature combination for predicting the heart disease. Cleveland and Statlog datasets are used for developing and testing the model. Base classifiers such as Support Vector Machine (SVM), K Nearest Neighbour (KNN), Decision Tree, Logistic Regression, Naïve Bayes and Vote are tested over all original 13 features from datasets, 9 feature combination and 6 feature combination. It is observed that Vote classifier using 9 and 6 feature combination gives the best accuracy and F1 score.***

Keywords: Cardio-vascular disease, Feature selection, Support vector machine, Ensemble classifiers, Naïve bayes

I. INTRODUCTION

Coronary Artery Disease (CAD) is identified as the number one killer disease in the world by World Health Organization (WHO). According to a 2015 article, it is estimated that around 110 million are affected by CAD. It leads to 17.9 million deaths which is 31% of death in 2016 [1]. In the highly developed countries like United States (US) and United Kingdom (UK) the case is still bad. CAD is the state of narrowing down of the blood vessels which are carrying blood to heart muscles.The primary reason for this narrowing down ofblood vessels are due to plaque built in it, life style or some hereditary factors. Early diagnosis of CAD risk will improve the treatment process and increases the survival rate of the patients.

Mostly, the heart functioning is diagnosed using echogram or electrocardiogram (ECG) tests. The medical specialist will identify the irregularities in the normal functioning of heart through ECG [2] signals. But in certain cases, the ECG also does not capture the severity of the CAD.

Manuscript published on 30 September 2019
* Correspondence Author
Baranidharan B*, Associate Professor, Department of CSE, SRM IST, Chennai. Email: baranidb@srmist.edu.in
Abhisikta Pal, UG student, Department of CSE, SRM IST, Chennai. Email: abhisiktapal.99@gmail.com
Preethi Muruganandam, UG Student, Department of CSE, SRM IST, Chennai. Email: preethimuruganandam@gmail.com

Angiogram is the most widely used standard for diagnosing the severity of CAD but it is very costly and invasive method. The cost of Angiogram makes it very difficult for rural people to afford it. So a less complex, less costly, affordable and accurate model needs to be built with the help of recent technological advancements.

Machine learning [ML] based predictive systems are being developed by Tech companies and academic institutions along with their partner hospitals. Lot of the classification models are being built for CAD diagnosis using ML techniques. But most of these are built around data sets from UCI repository. Heart disease UCI data sets [3] contains 14 variables where 13 are independent variables and 1 dependent variable. Age, Sex, Chest pain, Resting blood pressure (Trestbps), Cholesterol, Fasting blood sugar (Fbs), Restecg, Thalach, Exang, Oldpeak, slope, Ca and Thal are the independent variable or input variables. Class is the output variable which takes the values from 0 to 4 where 0 represents no heart disease and the values 1 to 4 represent the severity of heart disease. In recent times, the industrial giants like Microsoft [4], Google [5] have also involved the heart disease prediction model in cooperation with hospitals. The individual classifier algorithms give better results for particular data sets and fail to achieve the same for other data sets. So, it is observed that the combination of these classifiers with optimized feature selection algorithms will improve the prediction.

Let us have a look at individual classifiers:

- **Logistic Regression:** Logistic regression is the simplest of all the classifiers. It uses logits also called score based on probabilistic method for identifying the class of new input.
- **Naïve Bayes:** Naïve Bayes classifiers is based on Bayesian theorem. It is built on the assumption that each attribute or feature is completely independent of each other and has its impact on the output. Though it is a simple classifier sometimes its performance is much better than advanced classifiers.
- **Decision Tree:** Decision Tree builds prediction model based on a tree structure. Like in normal tree structure, it has root node, intermediate and leaf nodes. The root node represents the base feature for data set division, other important features are located in the next subsequent levels of the tree structure.
- **K Nearest Neighbour:** K Nearest Neighbour (KNN) is a non-parametric classifier algorithm. It classifies new data based on its distance with the K nearest already classified data. When the data set is huge, knn gives better performance than most other classifiers.

- **Support Vector Machine:** SVM is the large margin classifier which classifies the positive and negative data points witha large boundary between them. SVM is a strong classifierwhich does not suffer from overfitting problem unlike other similar classifiers.
- **Ensemble model:** Ensemble model [6] is the combination or aggregation of more than one classifiers. There are three techniques of building an ensemble model: (i) **Bagging**, where the same type of classifiers are used and the final decision is based on the vote from individual classifiers, (ii) **Boosting** is also similar like bagging but classifiers are in sequence and the performance of previous model affects the next model and (iii) **Voting Classifier**, which combines different classifiers and the final decision is based on the vote.

The presence of improper features in the dataset reduces the efficiency of the data mining techniques. Before identifying the best techniques there must be proper identification of best feature combinations. When the best feature combination is applied over the techniques it is expected to havea good improvement in accuracy and other performance metrics.

Feature extraction is the technique of producing a new set of m features from original n features. It combines the original features so that the effect of redundancy and inconsistency can be mitigated.Feature selection will identify the most important or significant features from the given dataset and leaves the less significant features so that accuracy and other performance metrics can be improved.

In this work, our contribution comes in two aspects, (i) Using proper feature selection to obtain more relevantfeatures for classification and (ii) Developing a classifier which is a combination of strong classifier so that consistent improvement can be achieved.In medical expert systems, the proper feature reduction mechanism will reduce the cost of diagnosis by prescribing the patients to undergo the most relevant medical examination and leaving unnecessary medical examination.

II. LITERATURE REVIEW

Alizadehsani et al [7] have built the various classification models for CAD based on Sequential Minimal Optimization (SMO), Bagging, Artificial Neural Networks and Naives Bayes. It is claimed that SMO and Bagging are giving higher accuracy at the level of 89% and ANN at 85% and Naives Bayes at very low level.

Srinivasan et al [8] had developed a heart care system based on 15 attributes for identifying the morbidity of people working in coal mines in Singaneri, Andhra Pradesh. Compared with other classifiers, Decision Tree gives better results.

Heart rate variability (HRV) is used as the main criteria for determining CAD by Melillo et al [9] It is observed that HRV predictive model identifies the risk of cardio vascular disease in a better way than even Echographic parameters.

An Intelligent Heart Disease Prediction System was developed by Palaniappan et al [10] based on Naïve Bayes, Neural Network and Decision Tree. It is concluded that Naïve Bayes has highest accuracy followed by Neural Network and Decision Tree. The experiment was conducted over 909 instances from UCI heart disease repository with equal split of training and testing set.

Pouriyeh et al [11] have conducted a comparison study on the different classifiers over Cleveland data set. Naive Bayes, Support Vector Machine, Radial Basis Function, Multi-Layer Perceptron, K Nearest Neighbour, Single Conjunctive Rule Learner and Decision Tree was taken for the comparison study. Apart from that the ensemble techniques like bagging, boosting and stacking was applied over individual classifiers to enhance the results. Finally, they concluded that SVM enhanced with boosting technique gives the highest accuracy than all others.

Latha et al [12] analysed the effect of bagging, boosting, stacking and voting combined with feature selection. Bagging technique shows good improvement over the weal classifiers than strong classifiers. Using the ensemble classification they achieved a maximum increase of 7% in accuracy.

Vivekandan et al [13] used Differential evolution [DE] algorithm for feature selection. Further the classification have been done by using Fuzzy analytic hierarchy process (Fuzzy AHP) and feed forward neural network. An accuracy of 83% is achieved by this hybrid model.

The individual classifier algorithms gives better results for the particular data sets and fails to achieve the same for other data sets. So, it is observed that the combination of these individual classifier as a single ensemble model will improve the prediction.

III. MEDICAL DATASET& PROPOSED MODEL

Most of the AI researchers uses the UCI heart disease data set which is made of from four different sources:

- Cleveland data set [14]
- Hungarian Institute of cardiology [15]
- University Hospital, Zurich, Switzerland [16]
- University hospital, Basel, Switzerland [17]

Cleveland data set is widely by ML research community. Totally, it has 303 records of 76 different variables. But out of that only 14 variables are identified to be closely related with heart disease, they are Age, Sex, Chest pain, Resting blood pressure (Trestbps), Cholesterol, Fasting blood sugar (Fbs), Restecg, Thalach, Exang, Oldpeak, slope, Ca, Thal and Class. Class is the output variable which takes the values from 0 to 4 where 0 represents no heart disease and the values 1 to 4 represents the severity of heart disease. Table I shows the description of each variables.

Table – I: Data set variables description

S.No	Variable	Description
1	Age	Age of the person in years
2	Sex	Gender 1 – Male 0 - Female
3	Chest pain	1 – typical angina 2 – atypical angina 3 – non anginal pain 4 - asympomatic
4	Resting Blood Pressure	Blood pressure in mm Hg during hospital admission
5	Cholestral	Serum cholesterol in mg/dl

6	Fasting Blood Sugar (fbs)	If (fbs>120mg/dl) 1 = true 0 = false
7	Restecg	Electrocardiography 0 – Normal 1 – may be some problem 2 – definite problem
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina 1 – Yes 0 - No
10	Oldpeak	Induced ST depression due to exercise
11	Slope	Slope of the ST segment during peak exercise 1 – Upsloping 2 – flat 3 - downsloping
12	Ca	Number of blood vessels coloured by fluoroscopy Values ranges from 0 to 3
13	Thallium Scan	It is a method of analysing blood flow to heart muscles 3 = normal 6 = fixed defect 7 = reversable defect
14	Class	It is the output or dependable variable 0 = No heart disease 1, 2, 3 and 4represents the severity of the heart disease

Another important dataset used by research community is Statlog [18] dataset.It also has the same input attributes like Cleveland dataset.

In this paper, the existing data mining models and the proposed model based on Vote classifier is tested over both Cleveland dataset and Statlog dataset. In the experimentation procedure, the heart disease prediction is converted into a binary model (i.e.) classifying into positive and negative classes. So, in the Cleveland dataset, the output 'Class' variable values are assigned 0 or 1. The previous values like 2,3 and 4 are reassigned the value 1 which represents the presence of heart disease. Since the total number of record is only 303, 10-fold cross validation is done in the experiments to ensure better results. In the case of Statlog dataset the output variable 'Class' has two values 'present' and 'absent' which is again converted into numerical values 1 and 0 respectively. In original Cleveland dataset 6 entries out of 303 are having missing values and in some of the comparisons only 297 entries are considered but in this work the missing values are replaced with the suitable values and all 303 records are used for model building. Gaussian Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression and K Nearest Neighbour are used for the experimentation purpose since these classifiers are proved to the best and traditional ones. Extra Tree Classifier from *SKLEARN* package is used for feature selection.

A. Extra Tree Classifier:

Extra tree classifier [19] generates randomized multiple decision trees with different sub-samples without bootstrapping. It avoids the problem of over-fitting and results in better accuracy.Three important parameters for Extra tree classifies are: (i) ***M***, represents the total number of trees to be generated, (ii) ***K*** represents the number of attributes chosen for tree construction and (iii) n_{min}denotes minimum required samples.When compared with random forest mechanism, extra tree classifier differs by choosing random K attribute and random split values for generating the tree. The attribute which shows lesser bias-variance is identified as the best split attribute.In this paper, Extra tree classifier is used for feature selection. Fig.1. depicts the input feature and the final selected features from the Extra Treeclassifier.

Fig. 1 – Extra Tree for Feature Selection

B. Vote Classifier:

In the proposed work, Vote classifier is built by combining SVM, Logistic Regression and Naïve bayes. The best 6 features obtained using Extra tree classifier are given as input to the Vote classifier. Fig. 2 shows the Vote classifier designed using SVM, Naïve bayes and Logistic regression. Since the above classifier are designated as strong classifier, Vote classifier is built based on these classifier.

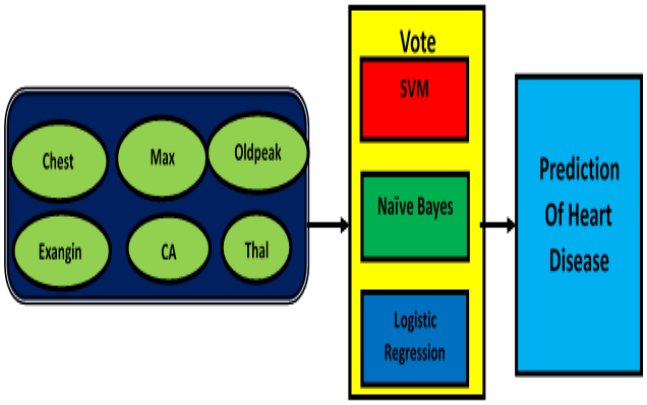

Fig. 2. Vote classifier based on SVM, Naïve bayes and Logistic Regression

IV. RESULT AND ANALYSIS

The classifiers used for comparison in this research are deployed and tested in the system having configuration of Intel i5 processor 7th generation, 8GB RAM, Windows 10 operating system and in Jupyter notebook environment. The inbuilt classifier models from Sci Kit (sklearn) library is used.

A. Comparison Metrics:

For comparing the different classifiers performance, the metrics such as accuracy, precision, recall and F1 scores are used. Equation 1, 2, 3 and 4 shows the formula for computation of the above mentioned performance measures. *True positive* and *True Negative* are the instances rightly predicted as positive and negative samples respectively. *False positive* instances are predicted as positive whereas really it is a negative sample and the *False Negatives* are predicted as negative samples but really these are positive samples.

$$Accuracy = \frac{TP+TP}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} (2)$$

$$Recall = \frac{TP}{TP+FN} (3)$$

$$F1 = 2X\frac{Precision * Recall}{Precision + Recall} (4)$$

In [20], authors have claimed that the following 9 attributes such as Sex, Chest pain, Fasting blood sugar, Electrocardiographic, Exercise induced angina, Oldpeak, Slope of the peak, Ca and thal are the best feature combination. In this research, Chest pain, maximum heart rate, exercise induced angina, oldpeak, ca and thal are identified as the best 6 feature combination for predicting the heart disease. The best 6 features are identified through Extra Tree classifier technique. Apart from identifying the best 6 features, a vote classifier is built using the combination of SVM, Naïve bayes and Logistic regression techniques.

B. CLEVELAND DATASET:

Table II, III, IV and V shows the accuracy, precision, recall and F1 score respectively for 13, 9 and 6 feature combinations. Fig 3, 4, 5 and 6 is the graphical representation of the Tables I, II, III and IV respectively. As an individual classifier Logistic regression is showing the highest accuracy of 0.8446 using 6 feature combination. As already mentioned the weak classifiers such as KNN and decision tree is showing the lower accuracy than other classifiers. But it is observed that KNN shows an improvement in accuracy of around 15% when the feature set is reduced from 13 to 9.But on considering Accuracy and F1 scores, the proposed model Vote classifier with 6 feature combination shows improved results than all other models. Since F1 score is based on Precision and Recall, it is the better metric for comparison along with accuracy.

Table – II: Accuracy of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8381	0.8283	0.8379
K Nearest Neighbor	0.6505	0.8180	0.7618
Decision Tree	0.7318	0.7649	0.7424
Logistic Regression	0.8249	0.8181	0.8446
Naïve Bayes	0.8415	0.8150	0.8381
Vote	0.8479	0.8382	0.8479

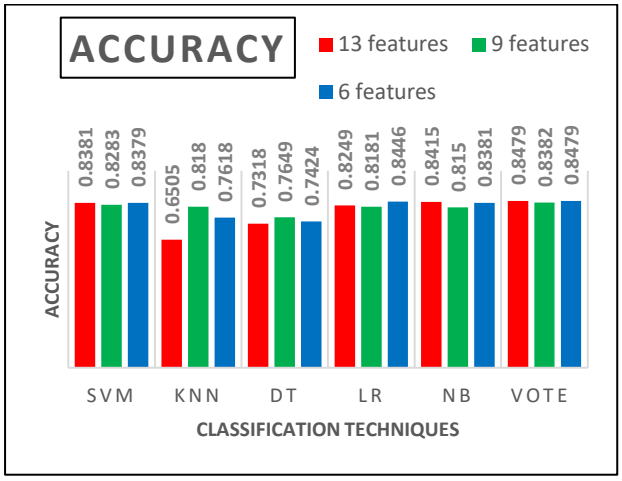

Fig.3.Accuracy vs ClassificationTechniques

Table –III: Precision of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8491	0.8276	0.8599
K Nearest Neighbor	0.6485	0.8303	0.7656
Decision Tree	0.7217	0.7604	0.7219
Logistic Regression	0.8421	0.8387	0.8677
Naïve Bayes	0.8433	0.8060	0.8579
Vote	0.8698	0.8446	0.8753

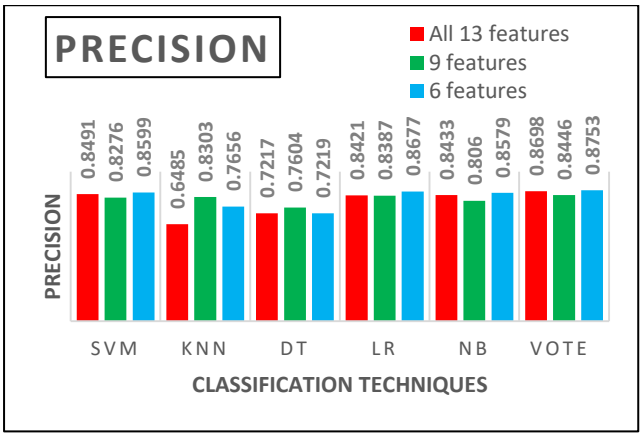

Fig. 4. Precision vs Classification Techniques

Table – IV: Recall of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.7878	0.7974	0.7844
K Nearest Neighbor	0.5721	0.7821	0.7096
Decision Tree	0.7025	0.7323	0.7152
Logistic Regression	0.7777	0.7751	0.7934
Naïve Bayes	0.8054	0.7916	0.7857
Vote	0.7974	0.8065	0.7934

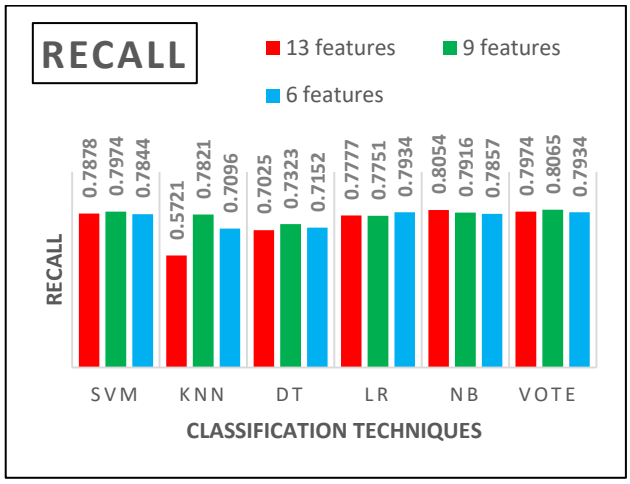

Fig. 5. Recall vs Classification Techniques

Table – V: F1 score of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8125	0.8085	0.8161
K Nearest Neighbor	0.5975	0.8000	0.7329
Decision Tree	0.7012	0.7371	0.7137
Logistic Regression	0.8022	0.7985	0.8249
Naïve Bayes	0.8191	0.7938	0.8167
Vote	0.8268	0.8209	0.8280

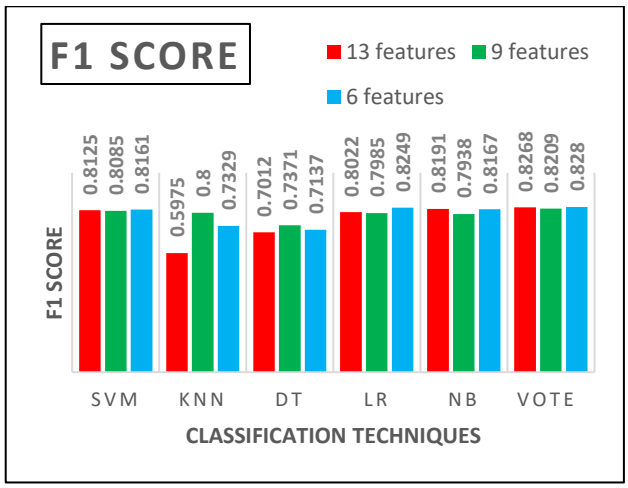

Fig. 6. F1 Score vs Classification Techniques

C. STATLOG DATASET:

The second set of experiments were conducted over statlog dataset. As already mentioned statlog has 270 patients details as against 303 in Cleveland. But Statlog is considered as pure dataset since all the values are proper and no missing values in statlog dataset. On Statlog dataset, Vote classifier with 6 feature is giving better or equal performance with Vote classifier using 9 feature combination. Table VI, VII, VIII and IX shows the accuracy, precision, recall and F1 scores of Statlog dataset using different data mining techniques with 13, 9 and 6 feature combination. So as Fig. 7, 8, 9 and 10 are its graphical representation.

In the case of SVM, Logistic Regression, Naïve bayes and Vote classifiers the accuracy increases clearly with feature selection. It is due the removal of unwanted features from the dataset which affects the accuracy. Logistic regression as the individual classifier gives the highest accuracy with 0.8518 using 6 features and Vote classifier gives the same accuracy as 0.8518 in both 9 and 6 feature combination. Also based on F1 score, Vote classifier is identifying as showing consistent performance for 9 and 6 feature combination.

Table – VI: Accuracy of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8296	0.8407	0.8481
K Nearest Neighbor	0.6703	0.8333	0.7555
Decision Tree	0.7518	0.7407	0.7555
Logistic Regression	0.8333	0.8407	0.8518
Naïve Bayes	0.8407	0.8407	0.8444
Vote	0.8370	0.8518	0.8518

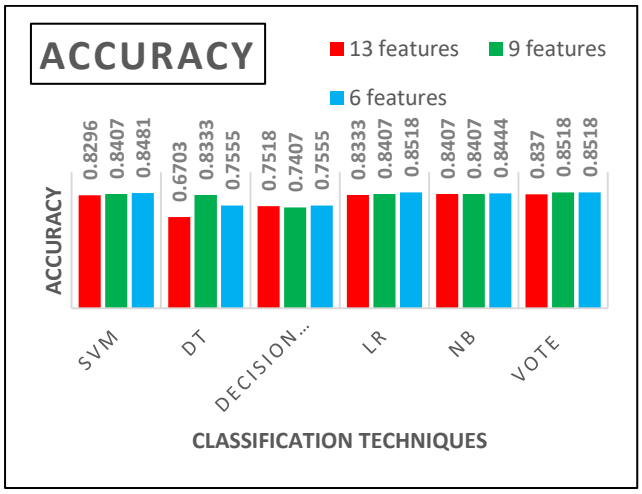

Fig.7. Accuracy vs Classification Techniques

Table – VII: Precision of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8269	0.8331	0.8523
K Nearest Neighbor	0.6470	0.8532	0.7478
Decision Tree	0.7272	0.7383	0.7533
Logistic Regression	0.8234	0.8379	0.8586
Naïve Bayes	0.8470	0.8290	0.8411
Vote	0.8469	0.8446	0.8515

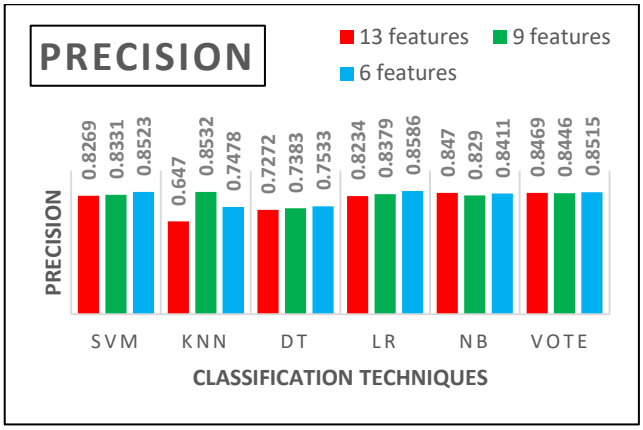

Fig. 8. Precision vs Classification Techniques

Table – VIII: Recall of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.7858	0.8175	0.8021
K Nearest Neighbor	0.5639	0.7692	0.6876
Decision Tree	0.7371	0.7229	0.7042
Logistic Regression	0.8014	0.8094	0.7989
Naïve Bayes	0.7833	0.8169	0.8092
Vote	0.7846	0.8247	0.8073

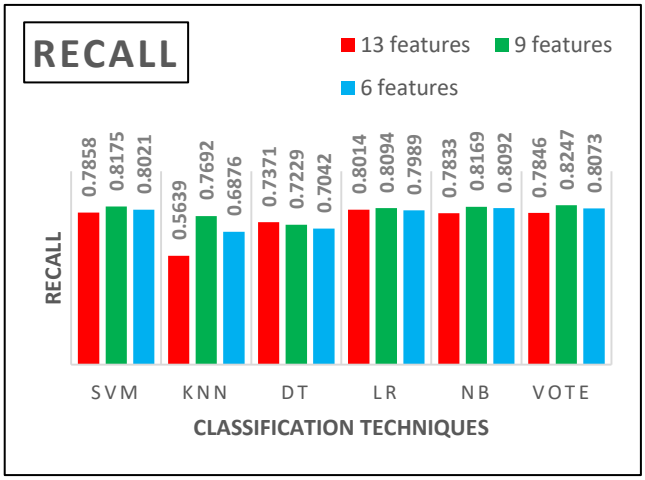

Fig. 9. Recall vs Classification Techniques

Table – IX: F1 score of models in 13, 9 and 6 feature combination

Models	All 13 features	9 features	6 features
Support Vector Machine	0.8006	0.8160	0.8197
K Nearest Neighbor	0.5989	0.7996	0.7060
Decision Tree	0.7197	0.7118	0.7144
Logistic Regression	0.8070	0.8152	0.8202
Naïve Bayes	0.8103	0.8168	0.8165
Vote	0.8058	0.8264	0.8223

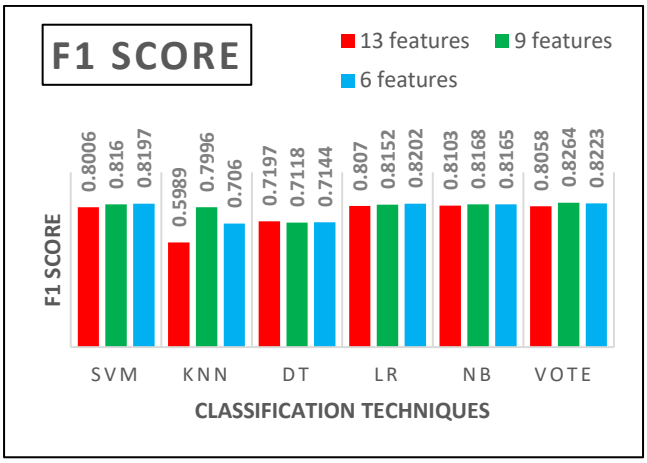

Fig. 10. F1 Score vs Classification Techniques

In both the Cleveland and Statlog datasets, Vote classifier gives better results than all the individual classifiers. In the case of weak classifier KNN, feature shows a steep increase in accuracy upto 15% in both the datasets.

V. CONCLUSION AND FUTURE WORK

Feature selection is an important technique in developing the predictive models for medical diagnosis. In this paper, the accuracy, precision, recall and F1 measures of six data mining models are compared with 13 feature, 9 feature and 6 features. Chest pain, maximum heart rate, exercise induced angina, oldpeak, ca and thal are the best 6 features identified using Extra-tree technique.The experiments were done over Cleveland and Statlog datasets to verify the impact of feature selection. It is clearly observed that based on accuracy and F1 score, Vote classifier shows much better results than other classifiers with 9 and 6 features. The effect of feature selection is felt more in KNN, with 15% increase in accuracy and 20% increase in precision, recall and F1 scores. It shows that the weak classifiers can be benefitted a lot using proper feature selection mechanism. In future, new feature selection techniques can be used for better feature selection with new combination of data mining techniques for Vote classifiers.

REFERENCES

1. Raghupathi, Wullianallur. "Data mining in health care." *Healthcare Informatics: Improving Efficiency and Productivity* 211 (2010): 223.
2. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. https://archive.ics.uci.edu/ml/datasets/heart+Disease
4. https://www.firstpost.com/tech/news-analysis/microsoft-and-apollo-hospitals-launch-ai-model-to-predict-heart-disease-risk-4988071.html

5. https://www.theverge.com/2018/2/19/17027902/google-verily-ai-algorithm-eye-scan-heart-disease-cardiovascular-risk
6. Dietterich, Thomas G. "Ensemble methods in machine learning." In *International workshop on multiple classifier systems*, pp. 1-15. Springer, Berlin, Heidelberg, 2000.
7. Alizadehsani, Roohallah, Jafar Habibi, Mohammad Javad Hosseini, HodaMashayekhi, ReihaneBoghrati, Asma Ghandeharioun, BehdadBahadorian, and Zahra Alizadeh Sani. "A data mining approach for diagnosis of coronary artery disease." *Computer methods and programs in biomedicine* 111, no. 1 (2013): 52-61.
8. Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques." In *2010 5th International Conference on Computer Science & Education*, pp. 1344-1349. IEEE, 2010.
9. Melillo, Paolo, Raffaele Izzo, Ada Orrico, Paolo Scala, Marcella Attanasio, Marco Mirra, Nicola De Luca, and Leandro Pecchia. "Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis." *PloS one* 10, no. 3 (2015): e0118504.
10. Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In *2008 IEEE/ACS international conference on computer systems and applications*, pp. 108-115. IEEE, 2008.
11. Pouriyeh, Seyedamin, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, and Juan Gutierrez. "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease." In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pp. 204-207. IEEE, 2017.
12. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." *Informatics in Medicine Unlocked* 16 (2019): 100203.
13. Vivekanandan, T., and N. Ch Sriman Narayana Iyengar. "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease." *Computers in biology and medicine* 90 (2017): 125-136.
14. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
15. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
16. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
17. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
18. Dua, Dheeru, and E. KarraTaniskidou. "UCI machine learning repository." *University of California, Irvine, School of Information and Computer Sciences* (2017).
19. Uddin, M.T. and Uddiny, M.A., 2015, May. Human activity recognition from wearable sensors using extremely randomized trees. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1-6). IEEE.
20. Amin, Mohammad Shafenoor, Yin Kia Chiam, and Kasturi DewiVarathan. "Identification of significant features and data mining techniques in predicting heart disease." *Telematics and Informatics* 36 (2019): 82-93.

AUTHORS PROFILE

Dr.B.Baranidharan has completed his Master of Technology in Computer Science and Engineering from SRM IST, Chennai and PhD in Wireless Sensor Networks (specialization) from SASTRA Deemed University, Thanjavur. Currently, he is working as Associate Professor in the department of CSE, SRM IST. He is having more than 10 years of academic experience and have published 22 papers in various International Journals and Conferences. Earlier his research involved about designing new clustering architecture for Wireless Sensor Networks and Internet of Things using various computational techniques. His current research includes Artificial Intelligence, Machine learning, Deep learning and Internet of Things.

Abhisikta pal has completed her high school with science from carmel school, Sarengabad, Kolkata. Currently she is pursuing B tech in Computer Science and Engineering from SRM Institute of science and technology,Chennai . She aspires to have a career in the field of data analysis . Her goal in life is innovation for the better utilisation of existing resources. She aims to research further in the field of data science to build models that can unfold more efficient ways to lead the human life . She has completed courses on Deep learning and machine learning from coursera and has done a project on big data analysis on the hadoop platform . She is highly motivated to excel in the field of data science.

Preethi Muruganandam is currently a third year student pursuing BTech in Computer Science and Engineering from SRM IST Kattankulathur Chennai. Hailing from Chennai, she completed her higher secondary education from DPS Gurugram with a very good academic record. Preethi is passionate in the fields of Machine Learning and Artificial Intelligence and has completed courses and research in those areas. She plans to earn a PhD in AI and has high aspirations of becoming an AI researcher.