Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

*

1st Abhishek Vyas
MT19086
abhishek19086@iiitd.ac.in

2nd Amrit Mohan Kaushik
MT18022
amrit18022@iiitd.ac.in

3rd Chirag Chawla
MT19089
chirag19089@iiitd.ac.in

4th Megha Mathur
MT19104
megha19104@iiitd.ac.in

***Abstract*—Heart disease is a big problem now a days due to the life style adopted by most of the people. They are addicted to the unhealthy eating habits which leads to many heart related diseases such as cardiac arrest, etc. In clinical world predicting such type of diseases is a typical task due to the enormous amount of patients and variety in symptoms like shortage of breathe, weight-loss, weight-gain, etc. but Machine Learning techniques has helped to do so very efficiently. Such Techniques helps to analyse a large amount of data for predicting possibilities and decision making. There are various techniques which are previously used for predicting cardiovascular diseases such as Multi Perception of ANN but doesn't provide the appropriate accuracy. Other such techniques were also imposed but doesn't able to provide appropriate result and are cost ineffective in terms of prediction nature. In this paper, an approach is introduced to collect only suitable features(only clinical features) that can help to increase the accuracy of the prediction model in order to predict the cardiovascular diseases and its stages in medical field accurately. Some of the methods are Decision-tree, KNN, random forest, ANN,SVM, Language models, and hybrid random forest linear model(HRFLM). HRFLM is the proposed model and considered as better model among all the above mentioned model since it provides better accuracy and f1-score when all clinical data attributes are considered. HRFLM is the hybrid model which has the advantage of both random-forest and linear model like lasso. Accuracy and F1-score for each model is calculated with and without considering non-clinical content and compared. The proposed hybrid model could be used in future for detecting early stage symptoms and can aid in controlling the adversity of such heart related diseases.**

***Index Terms*—Machine Learning, Cardiovascular Disease, HRFLM, F1-Score**

I. Introduction

In today's time, since each one of us are prone to unhealthy lifestyle and unsuitable environment, it is required to provide appropriate techniques to test the presence of any disease. Now-a-days heart related problems are very common among the generation such as cardiac arrest, diabetes, high blood pressure levels, high cholesterol and irregular pulse rates.Nature of heart diseases are complicated and difficult to predict and it was observed that each year among the total deaths per country, more than 55 percent deaths are occurred due to any heart related problems, so there is a need to predict such diseases in advance properly so that proper treatment shall be given to patients.For predicting such diseases appropriate tools and techniques are needed which could increase the precision of such predictions. Many data mining related techniques are used for this purpose such as decision-tree, KNN, etc. Before applying any such data-mining algorithm, efficient feature selection mechanisms need to be adopted so that model could be train well and produces more efficient outcomes.

A. Related Work

There are various prediction models which has been used in predicting the heart related diseases, such as ANN, CNN, SVM, etc., in past years by many researchers so that heart related problems shall be diagnosed efficiently and timely. ANN [2] was considered as highly efficient model since it gave the highest accuracy among other data mining related algorithms. Various classification algorithms were also used without segmentation like Convolutional neural network(CNN) [5] [6]. Many researchers are also working towards building models that could predict heart diseases in an efficient manner and that can decrease the cost associated with such predictions. Many data mining algorithms or machine learning algorithms [7] [8] are also deployed for this purpose and they are proved to provide better and accurate results as compare to other techniques.

B. Limitations of existing methods

Various data mining and machine learning algorithms or models were applied on provided data-set to predict cardiovascular diseases on various features provided. These machine learning techniques are retrieving accurate results but lacks proper precision as well as they haven't perform analysis without considering non-clinical data such as Age and sex. They also doesn't scale well enough for clinical data or attributes and therefore there is a need to change previous approaches. For getting higher accuracy for clinical precision, hybrid approach needs to be adopted.

C. Proposed Strategies

Since Cardio-vascular diseases are very sensitive in nature, therefore there exists a need to predict it efficiently with all clinical data available. Existing methodologies are giving high accuracy but lacks precision when it comes to clinical content.

To mitigate this problem, we have implemented models such as SVM,Naive-Bayes, KNN, Random-Forest, language model etc. Among them SVM is more accurate prediction. In order to increase the overall precision, we have introduced a Hybrid approach of random forest and linear model such as Lasso. By introducing such hybrid approaches, we could take the advantage of both random-forest and linear models and could predict the model accurately.

II. Materials and Methods

In order to predict heart diseases efficiently and timely, proper features selection or generation are required.

A. Dataset & its Features

Dataset is taken from the UCI machine learning repository and our dataset belongs to cleveland database. It comprises of data of 303 patients With 13 total features. Age and sex features were not given any importance as they doesn't provide any help in clinical data or content. But in some diseases and models, age and sex of patients plays very significant role.So we have considered two approaches, one with and other without considering these features.

Various features of data-set are as follow:-

- Age :- it is measured in years.
- sex :- 1 for male and 0 for female
- thal :- 3 for normal, 6 for fixed defect and 7 for reversable defect
- ca :- it is number of major vessels coloured after flouroscopy (0-3)
- slope :- slope of peak exercise ST segment
- fbs :- fasting blood sugar
- chol :- serum cholesterol
- trestbps :- blood pressure (resting)
- cp :- chest pain type

B. Features Selection

For implementing our proposed model, i.e HRFLM, we have to combine the advantage of random-forest and linear model.We have used lasso as a linear model here. Lasso basically provides a set of features which are best for the classification models. It returns the coefficient for each feature, Negative coefficient means that the feature is not useful for classification and rest of the features are then selected to implement the classification model or task.

In our data-set there are two features, i.e. age and sex, which are non-clinical in nature, that's why we have implemented each model once with and other without using these features so that we can achieve better results.

C. k Fold Cross Validation and evaluation parameters

For each model described in below section, 5 fold cross validation technique is adopted. Cross validation is required to produce more accurate output.It evaluates the model on the given sample data and reduces the error rate. And for each train-data we split data into train-test with test size=0.2

To evaluation performance of different models we basically two evaluation parameters in our analysis that are:-

- Accuracy : It is the measure of all correctly identified cases.It is the ratio of correctly predicted observation to the total observations:-

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

Fig. 2. Accuracy evaluation parameter

- F1 score : It is harmonic mean of Precision and Recall and give better result for incorrectly classified than accuracy :-

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Fig. 3. F1 score evaluation parameter

D. Methods

To predict the Heart Disease, various data mining algorithms are used after pre-processing the data. Some of the methods used are as follow:-

1) Decision Tree: Decision-tree is the classification model which could predict the labels of test data target using input-features provided. Here each internal node is labelled with input features that are being provided and outgoing edge from that nodes contains label of target.In Decision-Tree, we have used Entropy as a distance metric. It selects attribute with lowest entropy value. Train data is split into train and test data with test size of 20%. After that model is trained i.e. fitted and test labels are predicted.The best accuracy obtained is 81.5% and 78%, respectively for data without sex and age attributes and with sex and age attributes.

2) K-Nearest Neighbour: In this model, the vectors are formed for each record of train and test data and then the K-Nearest Neighbour model is applied. Firstly, the euclidean distance of a particular node is calculated from all the other nodes. From all the nodes only k nodes are selected which have the least distance from the test node. For assigning the label to the node, the labels of all the k nearest node is taken into consideration. Whichever label has a higher count, that label is assigned to the test node. For implementing this model, the sklearn library is used. The data is split into the test and the train part. In the training phase, all the training nodes are plotted into the coordinate graph. In the prediction phase, the node for which the label is to be predicted is plotted into the same coordinate plane. The Euclidean distance is calculated from all the training nodes and k-nodes are refined. Finally, the label having a higher count of nodes is assigned to the test node. The accuracy is calculated using sklearn's function. This model generates predictions with 68% accuracy. The same process is repeated after removing non-medical features namely "Age","sex".

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
71	0	1	160	302	0	1	162	0	0.4	2	2	2	1

Fig. 1. Features Available

Accuracy achieved with this approach is equals to 76%.

3) Naive Bayes: In the Naive Bayes approach Bayes rule is used for applying the classification model. Firstly, by using the training data prior probability is calculated for both the labels which are (0,1). The Gaussian is used for finding out the prior probability which may lie in the range of 0 to 1. After performing the fit operation finally the classification for test data is performed. While doing the classification for test data, Naive Bayes uses the prior probability and class probability which is equal to 1/2 in this case. Gaussian Naive Bayes calculates the probability of test data instance for both the classes and then assign the higher probability class label to the data. For implementing, the GaussianNB model from sklearn is used. The data is divided into train and test sets for calculating the accuracy of the model. When the labels are generated for the test data set then accuracy_score function from sklearn is used. The accuracy achieved from GaussianNB is approx 75-80% depending upon the test-train split. Also the same model is applied after removing the non-medical feature from the data such as ”Age”,”sex”. It do not shows that much difference in the accuracy as the accuracy in both the cases are nearly equal.

4) Neural Network: A Multi-layer perceptron model is applied using TensorFlow and Keras for the classification of data. The model is consist of 4 different layers. Each layer has a predefined output size and an activation function. First layer has the output size as 512 and the activation function is relu. For the second layer, the output size is defined as 256 and the activation function is relu. For the third layer, the output size is defined as 64 and the activation function is relu. For the fourth and final layer, the output size is defined as 512 and the activation function is softmax. For the compilation of model, the adam optimizer is used which try to optimize the accuracy after each epoch. As the loss function sparse categorical cross-entropy is used. After each epoch model tries to decrease this loss rate. When all the layers are well defined and compiled then fit operation is performed on the train data set. The model learns it’s parameter which can be used for the prediction. Finally, the class label for the test data set is calculated. For finding the accuracy of this multi-layer perceptron model, the accuracy_score function is used from sklearn library. The accuracy achieved by this model is approx 70-72% depending on the test-train split. The model is also used for classification after removing the non-medical features from the dataset such as ”Age”,”sex”. Here, the accuracy gets increased after removing these two features from the dataset.

5) Random Forest: Random Forest is a tree based algorithm. The random forest algorithm aggregates the results of various decision tree algorithms running over randomly selected subsets from the training set. For example : if random forest runs decision tree algorithm over three different randomly selected subsets, and two of them are having label ”A”,

and one is having label "B", then Random Forest algorithm votes for the class labels, and the class label with maximum votes (occurances) wins which is assigned as predicted class label.

Advantages

- It produces a highly accurate classifier.
- It gives estimates of what variables that are important in the classification.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

Disadvantages

- It may tends to overfit for soem datasets.
- In case of Categorical variables with different levels, random forest may tends to bias towrads the categorical variable with more levels.

We have splitted our original dataset into train and test set in the ratio of 70:30. we have applied the random forest algorithm from sklearn predefined library with parameters max_depth=2, random_state=0. And we are able to achieve the following scores :

6) Support Vector Machine: Support vector machine which is abbreviated as SVM cab be used either as classification or as regression algorithm, but nowdays it is used widely as classification algorithm because of high accuracy with less computation.

Support vector machine basically finds hyper planes between the data points so that it can classify the data points. To seperate the data points there are multiple hyper planes possible, SVM finds out the hyper plane with having maximum margin with their data points. Margin is the distance between data points and the hyper plane.

Hyper planes are the decision boundaries which helps to classify the data points. Dimension of the hyper plane depends on the number of features. If we have only two features, the hyper plane is just a line, data point of one side is classified as one class, while data points on other side is classified as other class. If we have three features then hyper plane is like a plane, and if we the number of features is greater than 3, then it is difficult to imagine the hyper plane in those cases.

Here We have splitted our original dataset into train and test set in the ratio of 70:30. we have applied the SVM from sklearn predefined library with parameters C=1000, gamma=0.01, kernel="rbf". And we are able to achieve the following scores :

7) Language Model: Language Model is generally used to predict the probability of a sequence, i.e how likely is a sequence based on the previous present sequences. A sequence can be of size 1 or more. This probability can be evaluated using the formula given in the fig below.

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

Fig. 4. Language Model

Here we are using bigram language model, i.e we are evaluating the probability of two words which is calculated by the given formula :

P(A—B) = count(A, B)/Count(A)

Here for a particular test record we will find the probability of each feature value with respect to class label, for example we are finding P("target" = 1 — "Age" = 50). Thats how we have evaluated the probabilities of each feature value with respect to particular class label for a particular test record, and then sum up all the probabilities for all the features to get the value for particular class label. And the class label with having higher sum of probabilities is assigned as the predicted class for that particular test record.

We have divided the data into train and test in the ratio of 70:30. And find the class labels for all the test records using the bigram model as explained above.

8) Logistic Regression: Logistic regression is suitable for the classification problem where the target value is categorical. Suppose that we have to predict that a particular mail is spam or not, and if we use linear regression, it gives a continuous value as output and we have to put some threshold. Lets say, our threshold value is 0.6 and for a spam mail linear regression gives the output value 0.5. So it is classified as non-spam mail which can lead to serrious consequences in real time. Thats why we use logistic regression. Here are the parameters of Logistic Regression :

Output : 0 or 1

Hypothesis : Z = WX + B (where W is weight matrix, X is the input feature matrix and B is constant)

Predicted value = Sigmoid(Z)

The hypothesis gives us the estimated probability, i.e how confident is that the predicted class label is the actual class label.

P(Y=1—X;theta) : Probability that predicted class label is 1, given input variable X which is parameterized by theta.

P(Y=0—X;theta) : Probability that predicted class label is 0, given input variable X which is parameterized by theta.

P(Y=0—X;theta) + P(Y=0—X;theta) = 1

If the value of "Z" is +ve infinity, it will give the class label "1", and if the value of "Z" is -ve infinity, it gives the class label "0".

Here We have splitted our original dataset into train and test set in the ratio of 70:30. we have applied the Logistic Regression from sklearn predefined library with parameters C=10.0, penalty="l2". And we are able to achieve the following scores :

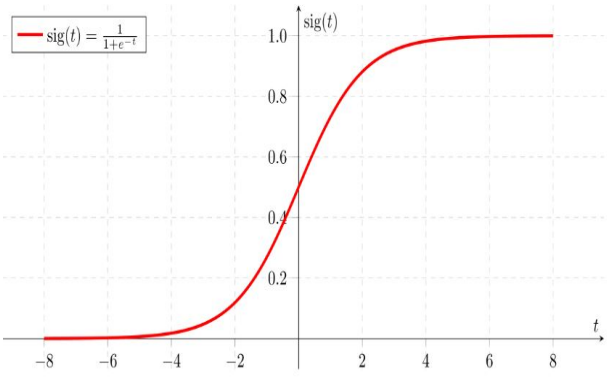

Fig. 5. Sigmoid Function

9) Hybrid Random Forest Linear Model (HRFLM) : A hybrid approach which is consist of the advantage of random forest and linear model is defined in this section. This model can be named as HRFLM(Hybrid Random Forest Linear Model). For implementing the linear model, the lasso model from sklearn is used. Lasso model tries to fit the dataset and find out the best features out of the features passed to this model. The lasso model returns the coefficient for each of the feature as output. If the coefficient for any feature is negative then that feature is considered to be less useful for classification purpose and if any feature has a positive coefficient then that feature is considered to be good for classification purposes. Here, before applying the random forest model lasso is applied and then the best features according to lasso are used further for random forest classification model. Random forest makes the multiple decision trees and uses the best one for the prediction. It removes the limitations of decision tree such as overfiting for training data. Random forest uses the concepts such as bootstrap, bagging for the tree learning. For implementation, Firstly the data is passed from the lasso and then the features are refined as per the lasso's output. On this refined data, the random forest is applied. Accuracy for this model is achieved as 86% also the same process is repeated after removing two non-medical features namely "Age" and "sex" and the accuracy remains a bit same as the previous approach.

III. Results

In our analysis we find accuracy and F1-score of prediction using different models in both cases i.e. including features age and sex and without including these two features.The results that we found in both cases are shown in figure 6 and 7:-

MODELS APPLIED	ACCURACY	F1-SCORE
Decision Tree	60.65%	0.7868
Naïve Bayes	78.48%	0.8324
KNN	63.93%	0.6272
Neural-network	80.32%	0.8019
SVM	57.44%	0.7750
Random-Forest	82.25%	0.7939
Logistic-Regression	78.90%	0.9180
Language Model	80.20%	0.7910
HRFLM	83.10%	0.7245

Fig. 6. Results of different models without including features age and sex.

Here we can see that in case when Non-medical features are excluded we get best accuracy for HRFLM and highest f1 score in logistic regression model.

MODELS APPLIED	ACCURACY	F1-SCORE
Decision Tree	73.77%	0.8356
Naïve Bayes	83.08%	0.7538
KNN	65.63%	0.6345
Neural-network	75.40%	0.7530
SVM	58.67%	0.8659
Random-Forest	82.22%	0.8032
Logistic-Regression	85.53%	0.8305
Language Model	80.21%	0.7950
HRFLM	83.89%	0.8196

Fig. 7. Results of different models including features age and sex.

Here we can see that in case when Non-medical features are included we get best accuracy for logistic regression and high f1_score for SVM model.
We have done graphical analysis of accuracy and f1_score for different models in both the cases (including and excluding non-medical features) also and the resulting graphs that we obtain are shown in figures 8,9,10 and 11:-

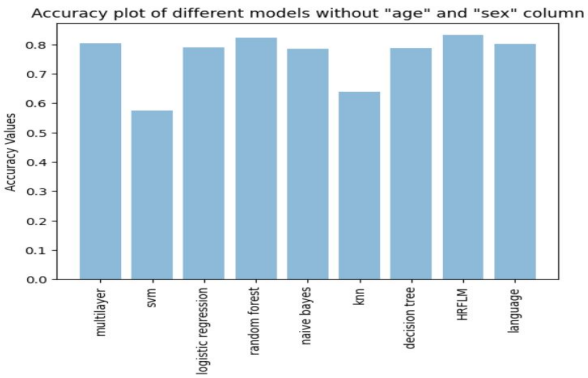

Fig. 8. Accuracy of different models without age and sex columns

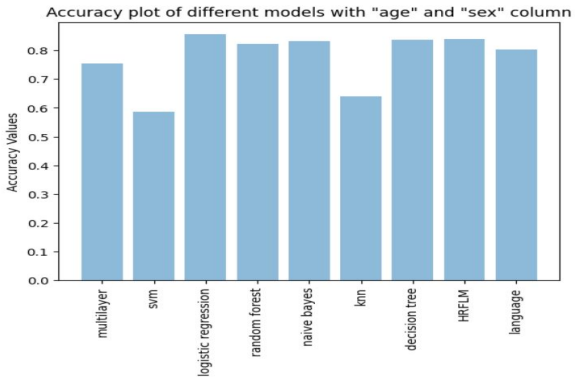

Fig. 9. Accuracy of different models with age and sex columns

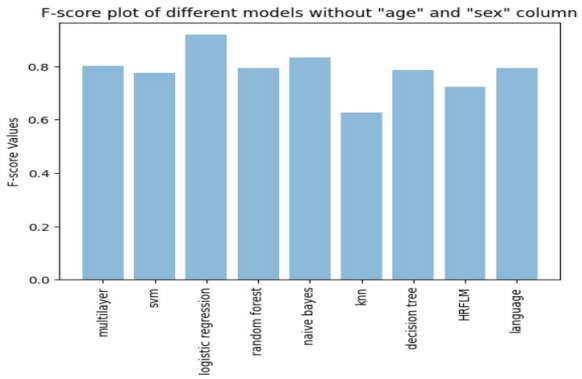

Fig. 10. F1-Score of different models without age and sex columns

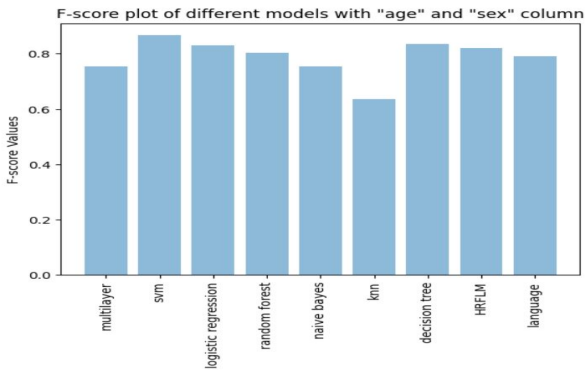

Fig. 11. F1-Score of different models with age and sex columns

In case of decision tree we found accuracy at different depth and chose the highest accuracy value of k as shown in figure 12 :-

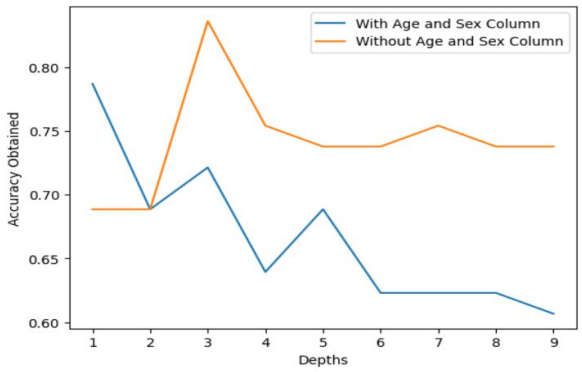

Fig. 12. Decision-Tree accuracy at different depth

Similarly in case of KNN we find accuracy at different values of K and result is shown through figure 13:-

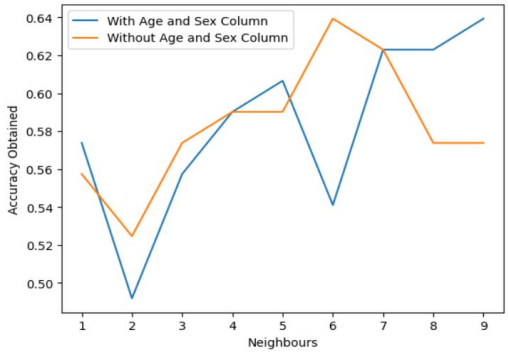

Fig. 13. KNN accuracy with different Neighbours

.

IV. Discussion

In this paper we use different techniques and models to predict Cardio-vascular diseases. As heart disease depends on various features, so we have used such features to predict it. The data that we have used also contains features such as age and sex that are thought to have less impact on heart disease prediction. So we have done analysis for both the cases i.e. prediction including features age and sex and prediction and without including these two features. The performance of different models is calculated using performance measures such as accuracy and F1-score. In case when non-medical features as age and sex are excluded, the best performing model based on accuracy is HRFLM and based on f1-score is logistic regression, respectively. In another case when non-medical features are included the best performance is obtained by logistic-regression model.The accuracy and F1-score in this case is 85.53% and 83.05% respectively which is highest among all models.

In our analysis we get better accuracy as well as better F1 score than the work which was done before. Here we have also taken care of the details, as we consider different depth in case of decision tree and also check for different value of

K in case of KNN. We also provided graphical representation of these two cases as in Fig.12 and Fig.13. We also have provided bar chart representation of different performance measures as accuracy and F1-score for both cases which is shown in Fig.8 to Fig.11 Here in these graph x axis represent different models while y axis represent score value of calculated performance measure.
Heart diseases are one of the major cause of high mortality rate.It can be controlled if we are able to detect or predict it in its early stage so that preventive measure can be taken on time. So using given method for prediction can play important role in such cases.Overall we can say that using given features we are able to predict good results which can be used to predict symptoms of heart disease in early stage and can play an important role to control mortality rate.

Future work can also be done in same field by including more features and using larger and real time data from different sources. As technology is improving and advancing day by day so further new methods can also be generated for the such prediction in favour of humanity.

V. Contributions

We have Worked in a group of 4 people and individual's contributions are as follow:-

- Abhishek Vyas - MT19086 :- Implemented SVM and logistic regression model with best parameters evaluation.
- Amrit Mohan Kaushik - MT18022 :- Implemented Naive-Bayes, KNN and HRFLM model.
- Chirag Chawla - MT19089 :- Implemented language model and random forest model.
- Megha Mathur - MT19104 :- Implemented Decision-tree and multi-layer perceptron models.

References

[1] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[2] L. Baccour, "Amended fused TOPSIS-VIKOR for classification(ATOVIC) applied to some UCI data sets," Expert Syst. Appl., vol. 99,pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[3] C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.(EMBC), Jul. 2017, pp. 2566–2569.

[4] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.

[5] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013. doi: 10.1016/j.eswa.2012.08.028.

[6] S. Zaman and R. Toufiq, "Codon based back propagation neural network approach to classify hypertension gene sequences," in Proc. Int. Conf. Elect., Comput. Commun. Eng. (ECCE), Feb. 2017, pp. 443–446.

[7] W. Zhang and J. Han, "Towards heart sound classification without segmentation using convolutional neural network,"inProc.Comput.Cardiol. (CinC), vol. 44, Sep. 2017, pp. 1–4.

[8] D.K.Ravish,K.J.Shanthi,N.R.Shenoy,andS.Nisargh,"Heartfunction monitoring, prediction and prevention of heart attacks: Using artificial neural networks," in Proc. Int. Conf. Contemp. Comput. Inform. (IC3I), Nov. 2014, pp. 1–6.

[9] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.

[10] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, "Analysis of neural networks based heart disease prediction system," in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[11] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," in Proc. IEEE 4th Int. Conf. Knowl.Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014

[12] M.GandhiandS.N.Singh,"Predictionsinheartdiseaseusingtechniques ofdatamining,"inProc.Int.Conf.FuturisticTrendsComput.Anal.Knowl. Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[13] M. S. Amin, Y. K. Chiam, K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics Inform., vol. 36, pp. 82–93, Mar. 2019. [Online]. Available: https:// linkinghub.elsevier.com/retrieve/pii/S0736585318308876

[14] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," in Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS), Mar. 2017, pp. 1–5.

[15] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network , " inProc.3rdInt.Conf.Comput.Sustain.GlobalDevelop. (INDIACom), New Delhi, India, Mar. 2016, pp. 3107–3111.

[16] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy-AHP for heart failure risk prediction," Expert Syst. Appl., vol. 68, pp. 163–172, Feb. 2017.

[17] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,"J.KingSaudUniv.-Comput.Inf. Sci.,vol.24,no.1,pp.27–40,Jan.2012.doi:10.1016/j.jksuci.2011.09.002.

[18] T.Mahboob,R. Irfan,andB.Ghaffar,"Evaluating ensemblepredictionof coronary heart disease using receiver operating characteristics," in Proc. Internet Technol. Appl. (ITA), Sep. 2017, pp. 110–115.

[19] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," in Proc. IEEE 4th Int. Conf. Knowl.Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[20] N. K. S. Banu and S. Swamy, "Prediction of heart disease at early stage using data mining and big data analytics: A survey," in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256–261.