Heart Disease Prediction using Hybrid Machine Learning

Shakir Ali[1], Mohammad Saifuddin[2], Pankaj Mishra[3]

[1,3]Assistant Professor, ACE Engineering College Hyderabad
[2]Research Scholer, Jamia Millia Islamia Central University New-Delhi

Abstract:

Heart disease usually refers to conditions such as narrowed or blocked blood vessels leading to heart failure, pain due to decreased blood flow to the heart (angina), or stroke. Heart disease is one of the leading causes of mortality in the world today. It contributes to about 30% of all deaths worldwide. Predicting cardiovascular disease is a critical challenge in the field of clinical data analytics. Machine learning (ML) has been shown to be effective in making decisions and predictions from the large amount of data generated in the healthcare industry. Machine learning is being used in many fields around the world. The healthcare industry is not exempt. Machine learning can play a significant role in predicting musculoskeletal conditions, heart disease and more. Such information, when predicted in a timely manner, can provide physicians with important clues so they can tailor their diagnosis and treatment to each patient. We are working on predicting potential heart disease in humans using machine learning algorithms. In this work, we present a novel method for identifying important variables that improve the precision of "cardiovascular disease" prediction. Along with a number of feature combinations and well-known classification techniques, the "prediction" model is introduced. When preprocessing the data in an efficient way we can achieve better performance. Preprocessing means handling missing values, removing outliers, balancing the data, scaling the feature, and optimizing the hyperparameter for the tree, selecting the important feature using the feature score technique. Experiment conducted shows that preprocessing technique helps us to get better performances. The experiment result also shows that hybrid random forest with linear model and feature score (HRFLMFS) prediction model for heart disease gives us a better result.

Keywords: Cleveland Heart Disease Database, Decision Trees, Random Forest, Hybrid algorithm, Machine learning

I. INTRODUCTION

According to the World Health Organization, 12 million people worldwide die each year from heart disease. Heart disease is a major cause of morbidity and mortality in the world population. The prediction of cardiovascular disease is considered one of the most important topics in the field of data analysis.

The burden of cardiovascular disease is rapidly increasing worldwide in recent years. Numerous studies have been conducted to identify the most influential factors for heart disease and accurately predict the overall risk. Heart disease has even been referred to as a silent killer, leading to the death of the affected

person without obvious symptoms. Early diagnosis of heart disease plays a critical role in lifestyle modification decisions for high-risk patients, reducing complications.

Machine learning is proving to be an effective aid in decision making and predictions from the large amount of data produced by the healthcare industry. This project aims to predict future heart disease by analyzing data from patients and using machine learning algorithms to classify whether or not they have heart disease.

Machine learning techniques can be a boon in this regard. Although heart disease can come in a variety of forms, there are a number of key risk factors that influence whether or not someone is ultimately at risk for heart disease. By collecting data from various sources, classifying them into appropriate headings, and finally analyzing them to extract the desired data, we can say that this technique is very well suited for predicting heart disease. It is difficult to detect heart disease because of several contributing risk factors such as diabetes, hypertension, high cholesterol, abnormal pulse, and many others. Various data mining and neural network techniques have been used to determine the severity of heart disease in humans. Disease severity is classified based on various methods such as K-nearest neighbor (KNN) algorithm, Decision Trees (DT), Genetic Algorithm (GA), and Naïve Bayes (NB) [11], [13]. The nature of heart disease is complex and therefore the disease must be treated carefully. Otherwise, the heart may be damaged or premature death may occur. The perspective of medical science and data mining are used for the discovery of different types of metabolic syndromes. Data mining with classification plays an important role in heart disease prediction and data investigation.

II. Related work

There are many current works studied by the researchers about heart disease prediction and analysis. Some of such works are addressed below.

The author studies Heart Disease Prediction Using Hybrid Machine Learning. They used Cleveland dataset and preprocessing of data before using classification algorithms [1]. The Knowledge extraction is based on Evolutionary Learning (KEEL), an open-source data mining tool that fills in the missing values in the dataset. A decision tree follows a top-down order. For each actual node selected by the hill-climbing algorithm, a node is selected at each level by a test. The parameters used and their values are confidence. The minimum confidence value is 0.25. The accuracy of the system is 88.4%.

The author studies Prediction of heart disease using Machine Learning Algorithms using decision trees and Naive Bayes algorithms for heart disease prediction. In the decision tree algorithm, the tree is created based on certain conditions that lead to correct or incorrect decisions [2]. The algorithms like SVM and KNN give results based on vertical or horizontal division conditions and depend on dependent variables. But decision trees have a tree-like structure with root nodes, leaves, and branches based on decisions made in each tree. They also used the Cleveland data set. The dataset was split into 70% training and 30% test using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can deal with complicated, nonlinear and dependent data, so it is suitable for the heart disease dataset because this dataset is also complicated, dependent and nonlinear. This algorithm provides an accuracy of 87%.

The author proposed Prediction of Heart Disease using Machine Learning Algorithms, in which they explain point by point Naive Bayes and decision tree classifiers, which are particularly used in heart disease prediction [3]. Some analysis was performed to investigate the execution of predictive data

mining strategies on the same dataset. The result was that the decision tree has the highest accuracy compared to the Bayesian classifier.

The author proposed a paper "Prediction of Heart Disease Using Machine Learning" in which training and testing of datasets is performed using Multi-Layer Perceptron neural network algorithm. In this algorithm, there is an input layer and an output layer and one or more hidden layers between these two layers [4]. Through the hidden layers, each input node is connected to the output layer. Some random weights are assigned to this connection. The other input is called bias, which is assigned a weight depending on the requirement. The connection between the nodes can be feedforwarded or feedback.
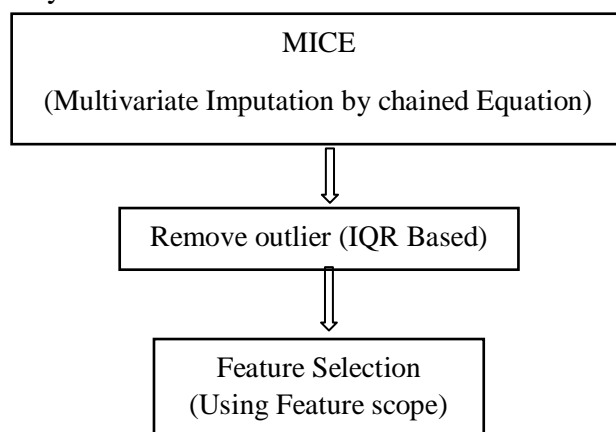
The author proposed "Heart Disease Prediction Using Effective Machine Learning Techniques," in which some data mining techniques are used to assist physicians in discriminating heart diseases [5]. The commonly used methods are k-nearest neighbor, decision tree, and Naïve Bayes. Other unique characterization-based strategies are packing calculation, partial strength,[6] consecutive negligible streamlining and neural systems, straight kernel self-arranging guide and SVM (Bolster Vector Machine). Lakshmana Rao et al. proposed "Machine Learning Tech[Clerical Data]t Disease Prediction" in which the factors contributing to heart disease are numerous [6]. various neural systems and data mining techniques are used to determine the severity of heart disease in humans.
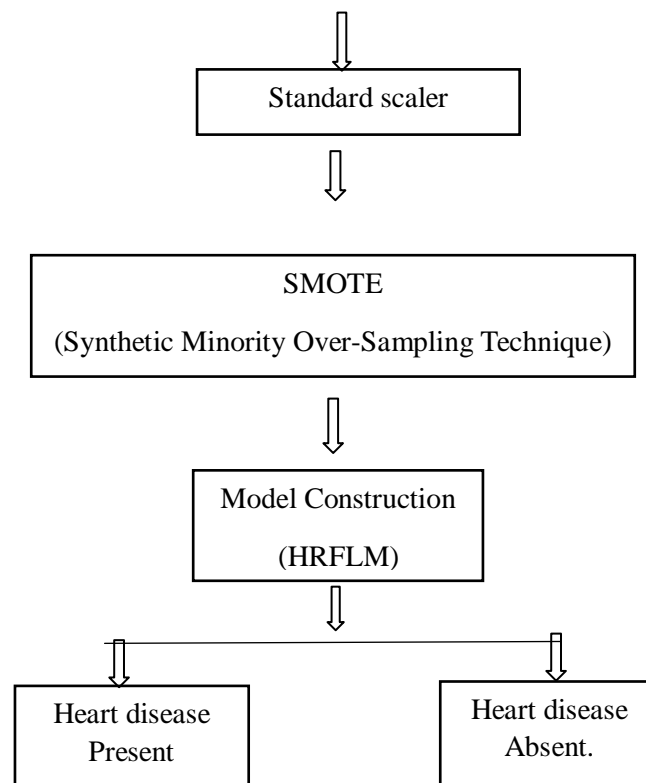
In this research author proposed "Heart Attack Prediction Using Deep Learning", in which heart attack prediction system using deep learning techniques and to predict the probable aspects of heart related infections of patient Recurrent Neural System is used [7]. This model uses Deep Learning and data mining to create the best and most accurate model with the least errors. This work serves as a strong reference model for other heart attack prediction models.

In this research author proposed "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," whose main goal is to improve accuracy in cardiovascular problems [8]. The algorithms used are KNN, LR, SVM, NN to achieve an improved exhibition level with a precision level of 88.7% through the heart disease prediction model using Hybrid Random Forest with Linear Model (HRFLM).

III. PROPOSED WORK

An overview of the system's operation is provided by the system architecture. The following is a description of how this system functions. Data sets containing patient data are gathered through data set gathering. The useful qualities for heart disease prediction are chosen by attribute selection. The available data sources are located, then further picked, cleaned, and changed into the form that is required. As previously indicated, HRFLM is used to preprocessed data to assess the precision of heart disease prediction. The accuracy measure evaluates how accurate certain classifiers are.

```
┌─────────────────────────────────────────┐
│                 MICE                      │
│ (Multivariate Imputation by chained       │
│  Equation)                                │
└─────────────────────────────────────────┘
                    ⇓
        ┌───────────────────────────┐
        │ Remove outlier (IQR Based) │
        └───────────────────────────┘
                    ⇓
          ┌─────────────────────┐
          │  Feature Selection   │
          │ (Using Feature scope)│
          └─────────────────────┘
```

```
            Standard scaler

                  ⇓

         SMOTE
(Synthetic Minority Over-Sampling Technique)

                  ⇓

      Model Construction
          (HRFLM)

           ⇓              ⇓

   Heart disease      Heart disease
    Present            Absent.
```
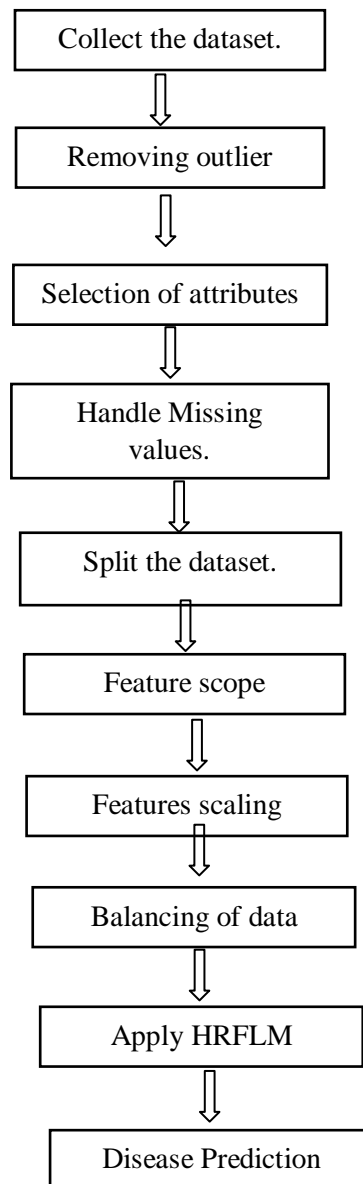
System Architecture

Proposed Algorithm (HRFLM) To increase the precision of predictions for cardiac datasets, hybrid machine learning is an approach that blends linear models and random forest classification algorithms. It will be feasible to create a hybrid algorithm that combines a linear model and a Random Forest to create an internal voting classifier, assessing the quality of both algorithms' predictions and selecting the one that delivers the most accurate results. As a result, when utilizing a hybrid model, we will always have a more precise heart disease prediction method.

The Hybrid technique has three steps:

I. We do this by using the probability function, which returns the target's probabilities in an array, to identify the output probabilities of each model. The target variable has an equal number of categories and probabilities in each row.

II. To achieve the lowest classification error possible, use the log loss function to determine the ideal weight for combining the two models. The degree to which your forecast deviates from the actual designation is gauged by a metric known as the log loss function.

III. Finally, combine the two models using the weighted average from the previous step, and make a prediction.

Proposed system: We have emphasized machine learning more in our study activity. There are techniques that enable two approaches to interpreting empirical data around machine learning. First, use the data's properties to find the complex relationships, then apply patterns to make predictions. Using methods akin to a technology that examines a data sample or training data to identify properties that the probability distribution is unable to detect, it is feasible to determine the link between the observed variables in the data. Since fresh information, it is feasible to apply the newly acquired knowledge to make a wiser decision. Typically, we can divide machine learning algorithms into various groups based on the outcomes. Unsupervised learning and supervised learning are a couple of these divisions. High

dimensionality could be a difficulty if we wish to evaluate big variables. There are numerous classification techniques available to prevent such issues. A single-stage selection strategy, for instance, can perform better when used before another approach. Each of the defined tactics has the following common traits: 1. A dimension-reduction technique 2. A method of variable selection.

Collect the dataset.

⇩

Removing outlier

⇩

Selection of attributes

⇩

Handle Missing values.

⇩

Split the dataset.

⇩

Feature scope

⇩

Features scaling

⇩

Balancing of data

⇩

Apply HRFLM

⇩

Disease Prediction

Proposed Methodology Workflow diagram

The first step in our model is to collect the dataset. We can download the dataset from **Kaggle.com**. In our dataset, there are 14 attributes and 1025 instances of the datapoints. We will split the data for the "purpose" of training and testing. We will take 75% of the data for training purposes and 25% of data used for training.

Selecting attributes re very important. If there are more attributes, then our model will not perform well. And hence we need to select only relevant attributes from the dataset. And we can eliminate attributes from dataset if the two attributes are correlated to each other.

Preprocessing of data Building a "machine learning model" requires a number of steps, including data preprocessing. Initial data may not be accurate or in the model's needed format, which can produce false results. The data is transformed into the desired format during preprocessing. It is applied to the data collection to eliminate noise, doubling, and misplaced values. Importing datasets, dividing datasets, scaling properties, and other operations are included in data preprocessing. Preprocessing of the data is necessary to increase the model's accuracy.

Handle Missing Values In many different fields of data science andmachine learning, missing data is a typical occurrence. Unfortunately, it can be difficult to handle successfully, and there is frequently no perfect solution.

Types of Missing data: There are three types of missing data that are commonly categorized. Completely random missing data (MCAR), random missing data and non- random missing data. Each type can occur in your data or even be a combination of multiple missing data types. MICE ALGORITHM: Data is assumed to be MAR, or missing at random, while performing multiple imputation using a chained equation. Occasionally, data from a data collection that are connected to the other features and can be anticipated given the values of the other characteristics are absent. The common techniques of employing mean, mode, or median cannot be used to infer them. For instance, a person may or may not have diabetes if their weight is missing. This value must be compared to other factors, such as height, BMI, and obesity, in order to estimate the right value.

Removing outliers are the ones which do not lie between some ranges. Sometimes outliers are useful and sometimes outliers are not useful and hence we need to remove the outliers. The Best Way to Spot an Outlier in a Dataset: An outlier must meet one of the two requirements listed below:

Outlier less than $Q1 - 1.5(IQR)$

Outlier greater than $Q3 + 1.5(IQR)$

Balancing data There are two approaches to balance out unbalanced data sets. Both inadequate and excessive sampling occurs.

(a) Under Sampling: When below sampler, the size of the vast class is decreased in order to maintain the balance of the data set. When there is enough data, this process is taken into account.

(b) Over Sampling: By enlarging the small, limited samples, oversampling balances the data set. When there is not enough data, this process is taken into account.

Smote Algorithm: - A numerical method called Synthetic Minority Oversampling Technique can be used to evenly growth the number of suitcases in your dataset. In this what we do, we add the number of rows in class which has the lowest number of data points, but we do not reduce the number of rows which is having the greater number of rows.

Feature Scaling, we know that there are number of independent features in the dataset and the feature scaling is a process by which we can distribute our attributes uniformly in range of [-1,1].

Necessity of feature scaling: Age, salary, and BHK dwelling are the three factors included in the dataset. Think about the age range of 10 to 60, the salary range of 1 to 40 lacs, and the number of bedrooms (1 to 5). These qualities stand one another. Assume the data point to be projected is [57, 33 Lacs, 2] and the centroid of class 1 is [40, 22 Lacs, 3].

So, feature scaling is the straightforward answer to this issue. Age, salary, and BHK are scaled by feature scaling algorithms in a specified range, such as [-1, 1] or [0, 1]. Then, none of the features can outweigh the others.

Disease prediction in this part we are going to apply different machine learning algorithm but specially HRFLM (proposed). We will apply a new dataset to the already trained model. And our prediction model will give output as yes means the person is having heart disease or the other output it will give no means the person is not having the heart disease.

Prediction of Disease for classification, a variety of machine learning algorithms are employed, including SVM, Naive Bayes, Decision Trees, Random Forests, Logistic Regression, and HRFLM. The algorithm with the highest accuracy is employed to forecast heart disease after a comparative examination of the algorithms is conducted.
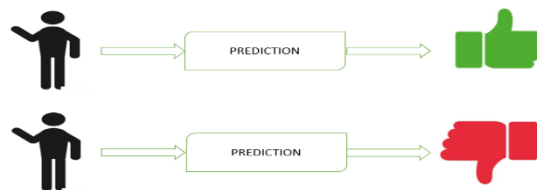
In this figure we can see that if we give some input data to our model then our model will predict whether the person has heart disease or not.

i.   Experimental Analysis

a) Dataset Detail:

In the Cleveland dataset there are 76 attributes. But most of the work has been conducted only on 14 attributes since other attributes are not so relevant. The specific dataset which is utilized by machine learning scholars is the "Cleveland dataset". The "target" field indicates the status of patient illness having two values as 0(zero) and 1(one). If the target is 0(zero) it means patient is having heart disease, while the target value 1 indicates that patient is not having heart disease when the target is 1(one) it means patient is not having heart disease.

In the hungry dataset, there are 276 datapoints or rows. The target field indicates the status of patient illness. This consists of two integer values as 0(zero) and 1(one). If the target is 0(zero) it means the patient is having heart disease, while when the target is 1(one) it means patient is not having heart disease.

b) Performance analysis:

In this research, heart disease is predicted using a variety of machine learning techniques, including SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and HRFLM. There are 76 attributes in the UCI data set for heart illness, however only 14 of them are considered for heart disease prediction. For this project, several patient characteristics are considered, including gender, the type of chest pain, fasting blood pressure, serum cholesterol, exang, etc. Each algorithm's accuracy must be evaluated, and the algorithm with the highest accuracy will be chosen to forecast heart disease. The experiment is evaluated using a number of metrics, including accuracy, confusion matrix, precision, recall, and f1 score.

Accuracy - The ratio of the number of accurate predictions to all the data set's inputs is known as accuracy. It is written as:

$$Accuracy = (TP + TN) / (TP + FP + FN + TN)$$

TP (True positive):A True Positive is the outcome where the model correctly predicts the positive class.

FP (False Positive): A False Positive is the outcome where the model incorrectly predicts the positive class.

FN (False Negative): And a false negative is an outcome where the model incorrectly predicts the negative class.

TN (True Negative): A true negative is an outcome where the model correctly predicts the negative class. A presentation indicator for a machine learning cataloging task where the productivity could include two or more classes is the confusion matrix. There are four likely anticipated and actual value groupings in the table.
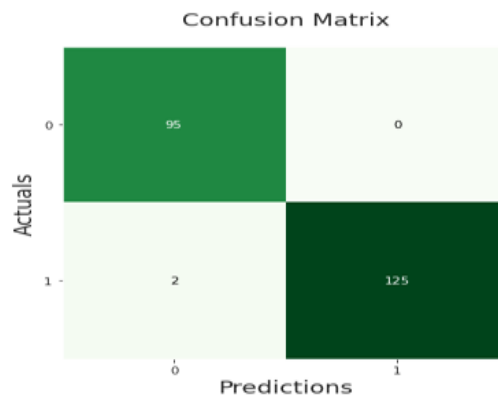
Confusion Matrix

A performance indicator for a machine learning classification task where the output could include two or more classes is the confusion matrix. There are four possible anticipated and actual value combinations in the table.

Precision:A classification model's capacity to isolate only the pertinent data points. Precision is calculated by dividing the total number of true positives by the total number of true positives + false positives.

$$Precision = (TP) /(TP+FP)$$

Recall: The capacity of a model to locate all pertinent instances in a data source. Recall is calculated mathematically as the product of the number of true positives divided by the sum of the true positives and false negatives.

$$Recall = (TP) / (TP + FN)$$

F1 Score: By calculating its harmonic mean, the F1 score condenses a classifier's precision and recall into a single metric. It is mainly used to compare the effectiveness of two classifiers. If classifier B has a better precision and classifier A has a higher recall, what will happen? Which classifier produces superior results in this situation may be determined using the F1 values for both classifiers.

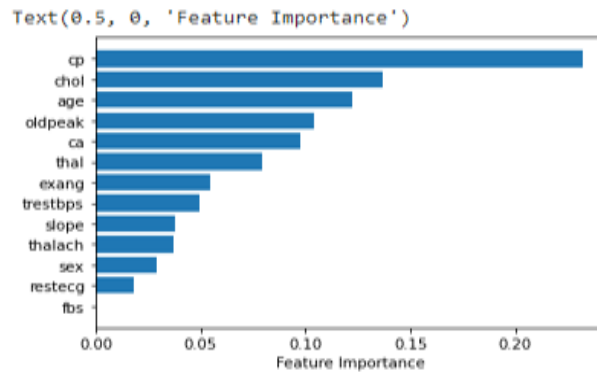The F1-score of a classification model is calculated as follows:

$$2 * (P * R) / (P + R)$$

Where P is the precision and R is Recall of the classification model.
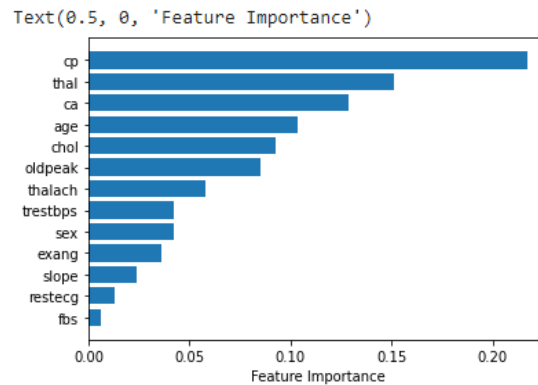
c) Result

In the below figure shows the feature score of every attribute using Decision Tree algorithm. Here we can see "restecg" has the lowest feature score while "cp" has the highest feature score.

Features scoping using decision tree

In the below figure shows the feature score of every attribute using Random Forest algorithm. Here we can see "fbs" has the lowest feature score while "cp" has the highest feature score.

Features scoping using decision tree

Classification result on heart disease dataset:

Classifier	Sensitivity	Specificity	Precision	Recall	F-Score	Accuracy
NV	79.80	60.00	90.50	79.24	84.50	75.80
LM	94.90	20.00	88.88	94.58	91.60	85.10
LR	91.10	25.00	89.60	90.80	90.20	82.90
DL	95.00	33.00	90.70	94.58	92.60	87.40
DT(FS)	100.00	100.00	100.00	100.00	100.00	100.00
HR	98.80	10.00	87.10	93.95	92.40	86.10
HRFLM	92.80	82.60	90.10	89.90	90.00	88.40
SVM	100.00	0.00	86.10	99.92	92.50	86.10
Proposed	100.00	100.00	100.00	100.00	100.00	100.00

In the above table, we can see that we got sensitivity, Specificity, Precision, Recall, F-Score and Accuracy 100 percent because we have trained our model only on the important Features which we have achieved by Random Forest Regressor.

Classification result inCleveland dataset

Classifier	Sensitivity	Specificity	Precision	Recall	F-Score	Accuracy
NV	82.31	84.75	84.37	82.31	83.33	83.50
SVM	81.09	87.80	86.92	81.10	83.91	84.46

LR	82.92	87.19	86.62	82.88	84.73	85.07
RF	84.14	89.02	88.46	84.14	86.25	86.60
HRFLM	92.80	82.60	90.10	95.03	92.50	88.40
Proposed	96.00	90.00	91.12	92.00	91.55	91.00

In the above table, we can see that we have achieved Sensitivity, Specificity, Precision, Recall, F-score, and Accuracy as 96.00, 90.00, 91.12, 92.00, 91.55 and 91.00 respectively because in this we have optimized the hyperparameter of random forest. But in this, our Recall and F-Score is not quite good but still it is acceptable.

Algorithm	Accuracy		
	Heart Dataset	Cleveland Dataset	Hungry Dataset
NB	75.80	83.5	82.50
SVM	86.10	84.46	84.60
LR	82.90	85.07	82.03
RF	86.10	86.6	------
HRFLM	88.40	88.40	86.00
DT	100.00	------	------
DL	87.40	------	83.05
LM	85.10	------	86.04
Proposed	100.00	91.00	87.50

In the above table gives the accuracy achieved using different machine learning algorithms. The experiment has been conducted on heart dataset, Cleveland dataset, and Hungary dataset. It can be seen from table 5.8, the proposed model outperforms and gives the highest accuracy among all the existing machine learning algorithms. This is due to the reason that the experiment has been conducted using top six selected features having highest feature score.

IV. Conclusion

Heart disease is a serious condition that can cause severe difficulties including heart attacks. Heart disease estimation is difficult and crucial in the medicinal industry. Categorizing the processing of raw healthcare data from cardiac information will help save lives and detect "abnormalities" in heart disease early in the long term. In this study, preprocessing has been done using different machine learning techniques to identify cardiac problems in an efficient way. However, if the illness is identified initial and preventive movements are initiated as soon as reasonable, the death proportion can be significantly decreased. According to research. DM and ML methods are important because they have the capability to accurately anticipate the existence of diseases. Using a Decision Tree and Random Forest we can predict heart disease in a very efficient manner. In this proposed HRFLM method we hybrid the RF and LM characteristics. We also perform preprocessing the data in an efficient way that helps us to attain a better performance. Preprocessing includes handling missing values, removing outliers, balancing the data, scaling the feature, and optimizing the hyperparameter for the tree, selecting the important feature

using the feature score technique. The experiment result shows that the proposed method has predicted heart disease in an efficient way and giving much better performance.

V. Future Scope

Though ML techniques are motionless in their infancy, this demonstration that they may end up existence an eccentric addition to enduring care. The prediction approaches can be improved in the future by conducting this research with other combinations of machine learning algorithms. Furthermore, novel feature collection techniques can be formed in order to gain a more comprehensive empathetic of relevant features and improve the accuracy of heart sickness estimation.

Reference:

1. Pandiaraj A, Prakash SL, Kanna PR. Effective Heart Disease Prediction Using Hybrid Machine Learning. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) 2021 Feb 4 (pp. 731-738). IEEE.

2. Santhana Krishnan, Prediction of Heart Disease Using Machine Learning Algorithms, IEEE 2019 International Conference on Innovations in information and Communication Technologies (ICICT), pp.1-7.

3. A. Dewan, M. Sharma, Prediction of heart disease using a hybrid technique in data mining classification, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2015, pp. 704-706.

4. K. Pahwa, R. Kumar, Prediction of heart disease using hybrid technique for selecting features, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), IEEE, 2017, pp. 500-504.

5. N. Bhatla, K. Jyoti, an analysis of heart disease prediction using different data mining techniques, 1 (2012) 1-4.

6. I. Yekkala, S. Dixit, M. Jabbar, Prediction of heart disease using ensemble learning and Particle Swarm Optimization, 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), IEEE, 2017, pp. 691-698.

7. T.J. Peter, K.J.I.J.o.S. Somasundaram, R. Publications, Study and development of novel feature selection framework for heart disease prediction, 2 (2012) 1-7.

8. K.H. Miao, J.H. Miao, G.J.J.I.J.A.C.S.A. Miao, Diagnosing coronary heart disease using ensemble machine learning, DOI (2016).

9. J. Soni, U. Ansari, D. Sharma, S.J.I.J.o.C.S. Soni, Engineering, Intelligent and Effective Heart Disease Prediction System Using Weighted Classifiers, 3 (2011) 2385-2392.

10. D.S. Medhekar, M.P. Bote, S.D.J.I.J.E.R.S.T.E. Deshmukh, Heart disease prediction system using naive Bayes, 2 (2013).

11. S.A. Pattekari, A.J.I.J.o.A.C. Parveen, M. Sciences, Prediction system for heart disease using Naïve Bayes, 3 (2012) 290-294.

12. H.D. Masethe, M.A. Masethe, Prediction of heart disease using classification algorithms, Proceedings of the world Congress on Engineering and computer Science, 2014, pp. 22-24.

13. A.J.O.J.o.C.s. Taneja, technology, Heart disease prediction system using data mining techniques, 6 (2013) 457-466.

14. M. Sharma, R.J.I.J.o.C.A. Kaur, Data Mining in Healthcare using Hybrid Approach, 975 (2015) 8887.

15. R. Chitra, V.J.I.j.o.s.c. Seenivasagam, Review of heart disease prediction system using data mining and hybrid intelligent techniques, 3 (2013) 605-609.

16. M. Gandhi, S.N. Singh, Predictions in heart disease using techniques of data mining, 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), IEEE, 2015, pp. 520-525.