Machine Learning Classifying Approach for Identifying Heart Disease in Medical Field

Kavitha B S[1]
Sri Siddhartha Institute of Technology
Department of Computer Science and Engineering
Tumakuru, Karnataka

Dr. M Siddappa[2]
Sri Siddhartha Institute of Technology
Department of Computer Science and Engineering
Tumakuru, Karnataka

***

Abstract- *Heart disease is one of the leading causes of death in the modern world. In the field of clinical data analysis, predicting heart disease is a major difficulty. In healthcare, especially in the area of cardiovascular, timely and accurate diagnosis of cardiac disease is crucial. The fatality rate can be minimized through early identification of heart disorders. There are few efficient analysis tools for uncovering hidden correlations and patterns in medical information from clinical records. As anticipating cardiac conditions is a challenging job, it is necessary to automate the process in order to avoid potential risks and to inform the patient ahead of time. In this field, machine learning algorithms are extremely significant. Machine learning (ML) has been found to be useful in helping with the decision-making and predicting of enormous amounts of data generated by the medical industry. Our goal is to find the most appropriate machine learning technique for diagnosing cardiac disease that is both computationally efficient and accurate.*

Keywords- Machine learning, Supervised learning, ROC curve, Confusion matrix, Cross-validation.

1. INTRODUCTION

Machine learning (ML) is an artificial intelligence subfield (AI) that is widely being used in the domain of cardiology. Nowadays, we use machine learning in our daily lives. Software programs that use machine learning can access data and use this to learn for themselves. It implies that in ML, prior knowledge is used to predict the future. It is simply how computers, with or without human supervision, make sense of data and decide or classify a task.

There are four types of machine learning algorithms: The boundaries of the algorithm are constrained by supervised learning [11], in which the developer labels the dataset.

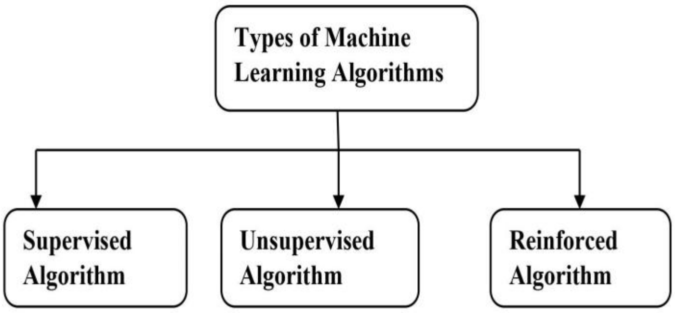


Fig -1: Machine Learning Types

Unsupervised learning does not require supervision, and semi-supervised machine learning, both supervised and unsupervised, is utilised in a combined manner. Reinforcement Learning means learning to explore things one by one, with the first occurrence serving as an input for the next.

Machine learning [15] is a modern technique that is divided into two stages: training and testing. After the system has gained experience and learned from data, the testing data is examined in compliance with the application's parameters.

The heart is a muscle that control blood flow into the body and is the most important component of the cardiovascular system. A network of blood vessels, such as veins, arteries, and capillaries, is also part of the cardiovascular system. These blood veins transport blood throughout the body. Heart disorders, often known as coronary heart disease, are caused by disruptions in normal blood flow from the heart (CVD). Heart disease is the leading cause of mortality in the globe. As per a World Health Organization (WHO) survey, heart attacks and strokes account for 17.5 million deaths worldwide. As a result, early diagnosis of heart irregularities and techniques for heart disease prediction can save a lot of lives and assist doctors in designing an appropriate treatment strategy, lowering the rate of death due to cardiovascular diseases.

Objective:

1. Find appropriate Machine Learning technique to diagnose cardiac disease.

2. LITERATURE REVIEW

Ankur Gupta et al. [1] proposed Machine intelligence framework (MIFH) to diagnose heart disease. The characteristics from the UCI Cleveland dataset are chosen using factor analysis of mixed data (FAMD). Some missing values have been found in the Cleveland dataset. Data imputation was utilised to fill in the missing values of the features in the new value. To convert the data into a suitable format, data standardisation was employed. The data is divided into two categories: training data and testing data. The resulting features are fed into the LR, KNN, RF, SVM, and DT machine learning classifiers. Performance indicators like as sensitivity, precision, accuracy, and f-measure are used to assess the effectiveness of machine learning algorithms. The results demonstrate that when using the RF machine learning classifier in conjunction with the FAMD, the accuracy is 93.44 percent. The proposed MIFH can efficiently distinguish between normal people and people who have heart problems.

Dakun lai et al. [2] suggested an arrhythmic risk marker-based technique for detecting sudden cardiac death (SCD) [10]. Sudden cardiac arrest (SCA) occurs when the heart suffers from an arrhythmia and heart stops. The authors of this research proposed a method for detecting sudden cardiac death using arrhythmic risk factors extracted directly from ECG signals. Machine learning classifiers KNN, DT, NB, SVM, RF take the risk markers as input and classify them as SCD or normal. It can detect SCD 30 minutes before symptoms appear. In comparison to other models, the results demonstrate that the RF classifier has a 99.49 percent accuracy. It is capable of distinguishing between SCD and normal.

Devansh Shah et al. [3] investigated a number of machine learning methods that can accurately and effectively predict whether or not a person will acquire heart disease. The information comes from a database in Cleveland. Pre-processing and feature selection are performed on the data. These features are used to train ML models such as KNN, SVM, NB, DT, and RF. To forecast cardiac disease, the ML model with the highest accuracy is chosen. The results demonstrate that at k=7, the K-nearest neighbour has the maximum accuracy.

Avinash Golande et al. [4] spoke about the effectiveness of several machine learning algorithms in predicting cardiac disease. Pre-processing, splitting, and classification are all parts of the proposed model. The data is pre-processed in the first stage, and any information that isn't a number is turned to one. Training and testing data are separated from the rest of the information. The training data is fed into four machine learning models: KNN, K-mean, Adaboost, and DT. The ML models learn from the training data and classify the test data into normal and heart disease categories.

Martin Gjoreski et al. [5] recommended using heart sounds to determine chronic heart failure (CHF). CHF is a condition in which the heart is unable to properly pump blood to the whole body. Both machine learning and deep learning techniques are used in this process. A phonocardiogram is used to record the heart sounds (PCG). In a healthy person, two heart sounds are routinely captured; however, additional heart sounds are recorded, that is not deemed acceptable. The existence of these extra sounds implies the presence of cardiac disease. From the raw PCG signal, the machine learning module features are extracted. The raw PCG signal is used by the deep learning module. The ML module examines features defined by experts and retrieves them using the opensmile tool. The recording also included sounds that were not related to the heart. It's time to get rid of these obnoxious noises. The DL is made up of multiple layers. The output of one layer is fed into the next layer as input. Using recording-based ML, the outputs of both the DL and ML modules are integrated. The recording-based ML is trained using the RF method. The output of recording-based ML predicts whether the sound comes from a healthy person or a cardiac patient, as well as the phases of CHF. The data reveal that utilising the aforesaid strategy yields a 93.2 percent accuracy.

Norma Latif Fitriyani et al. [6] suggested heart disease prediction model (HDPM) utilising statlog and Cleveland datasets. The HDPM model includes DBSCAN (density-based spatial clustering of noisy applications), SMOTE-ENN (synthetic minority over sampling technique edited closest neighbour), and XGBOOST (synthetic minority over sampling technique edited nearest neighbour). The HDPM model's result is evaluated to that of various machine learning algorithms such as LR, NB, RF, DT, and SVM. When compared to previous machine learning models, the suggested HDPM model achieved an accuracy of 98.40 percent and 95.90 percent for Cleveland and statlog, respectively.

Senthil Kumar Mohan et al. [7] suggested a novel heart disease prediction model. One or more strategies are combined to create the prediction model. The hybrid method is the name given to this novel technology. They suggested a new HRFLM method by combining the Random Forest and Linear methods in this study. The primary goal of this study is to improve heart disease prediction accuracy. The proposed hybrid random forest and linear technique may choose all of the features (HRFLM). There are no restrictions on the features available. The study suggests that the RF and LM approaches forecast cardiac disease with excellent accuracy. For patient personal identification, the attributes age and sex are used. The remaining qualities offer crucial data for predicting heart disease. The absence of cardiac disease is shown by the value of the num attribute being ZERO. The existence of cardiac disease is indicated by a number ranging from one to four. The number four implies that you are at a high risk of developing heart disease.

Amin et al. [8] suggested hybrid intelligent machine-learning-based predictive method for the diagnosis of heart disease. The Cleveland heart disease dataset was used to evaluate the system. Three feature selection algorithms were utilised with seven well-known classifiers, including logistic regression, K-NN, ANN, SVM, NB, DT, and random forest. The key characteristics were chosen using relief, mRMR, and LASSO. The system was validated using the K-fold cross-validation method. Different assessment measures were also used to check the performance of classifiers. When selected by the FS algorithm Relief, the classifiers logistic regression with 10-fold cross-validation exhibited the best accuracy of 89 percent.

Mohan et al. [9] used mixed machine learning approaches to create an HD prediction solution. He additionally suggested a new strategy for extracting essential features from data for machine learning classifier training and testing. They were classified correctly 88.07 percent of the time.

Samuel et al. [10] suggested, for the diagnosis of HD created an extensive healthcare decision support system based on neural networks and fuzzy AHP. In terms of accuracy, the suggested methodology performed at 91.10 percent.

3. MATERIALS AND METHODS

1.1 Dataset

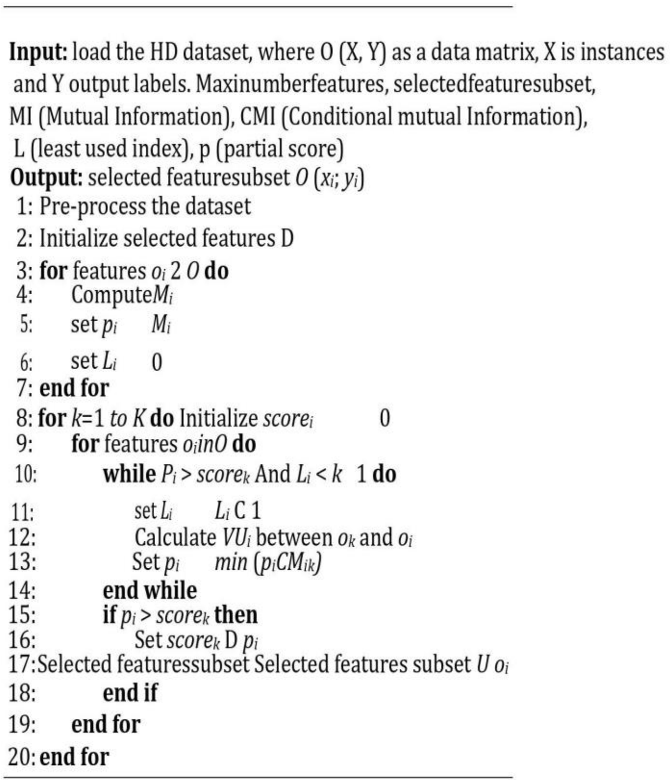
The dataset is taken from the Cleveland repository. There are 16 features with one output label which is used to predict whether the person is having heart disease or not. There are 4241 patient's details with some missing values.

1.2 Data Pre-processing

The dataset contains missing values and noise. The missing values and noise, if not removed gives incorrect result which leads to incorrect classification. So, pre-processing of dataset is required to remove missing values and noise.

1.3 Feature selection algorithms

Feature selection is a very important technique. The number of features can be minimized and appropriate features can be selected by using feature selection technique. If features are not selected properly then ML model fails to correctly classify between heart disease and normal healthy person. In this study, education feature is removed from the dataset, since education is not required to decide whether the person is having heart disease or not.



```
Input: load the HD dataset, where O (X, Y) as a data matrix, X is instances
 and Y output labels. Maxinumberfeatures, selectedfeaturesubset,
MI (Mutual Information), CMI (Conditional mutual Information),
 L (least used index), p (partial score)
Output: selected featuresubset $O\ (x_i; y_i)$
Pre-process the dataset
Initialize selected features D
for features $o_i$ 2 $O$ do
    Compute$M_i$
    set $p_i$    $M_i$
    set $L_i$    0
end for
for $k$=1 to $K$ do Initialize $score_i$         0
    for features $o_i in O$ do
        while $P_i > score_k$ And $L_i < k$  1 do
            set $L_i$      $L_i$ C 1
            Calculate $VU_i$ between $o_k$ and $o_i$
            Set $p_i$     $min\ (p_i CM_{ik})$
        end while
        if $p_i > score_k$ then
            Set $score_k$ D $p_i$
Selected featuressubset Selected features subset $U\ o_i$
        end if
    end for
end for
```

Algorithm 1: FCMIM Algorithm

In this study, we developed the Fast Conditional Mutual Information (FCMIM) feature selection method. It is a strategy for selecting features that is based on conditional mutual information (CMI). The following operations are included in the "FCMIM" algorithm design. Consider the collection O (X; Y), where X represents instances and Y represents output labels. To estimate the value of feature relevance and redundancy in the dataset, the FCMIM uses the CMI. The FCMIM [13] algorithm selects features that enhance their mutual information with the class label, subject to the outcome of any previous feature selection (O). The cmi criterion selects features that are significantly associated with the class whether they're least correlated with any other feature that has previously been chosen. Because it lacks information about the class to predict, cmim does not choose a characteristic that is comparable to those already selected, even if it is individually powerful.

1.4 Leave one out cross validation

In cross validation the dataset is divided into training and testing set. The 70% of data is used to train the model and 30% of data is used to test the model. In LOSO technique, first data is used to test the model and the remaining data is used to train the model in the first iteration. In second iteration, the second data is used to test the model and remaining data is used to train the model. This process continuous until all the data points are covered.

1.5 Performance evaluation metrics

The metrics [14] are used to check whether the model is predicting the disease properly or not. If correct kind of metrics is not used to check the performance of the model, then it is difficult to get correct predictions. The confusion matrix can be used to calculate the metrics. A confusion matrix is a table pattern that helps visualise the different outcomes of predictions and results of a classification issue.

Table 1:

Confusion matrix

	Predicted HD patient (1)	Predicted healthy person (0)
Actual HD Patient (1)	TP	FN
Actual healthy person (0)	FP	TN

Accuracy=((TP+TN)/(TP+TN+FP+FN)) *100

Precision=((TP)/(TP+FP)) *100

Specificity=((TN)/(TN+FP)) *100

Sensitivity=((TP)/(TP+FN)) *100

TP: these are the cases which correctly get classified as true and are true.

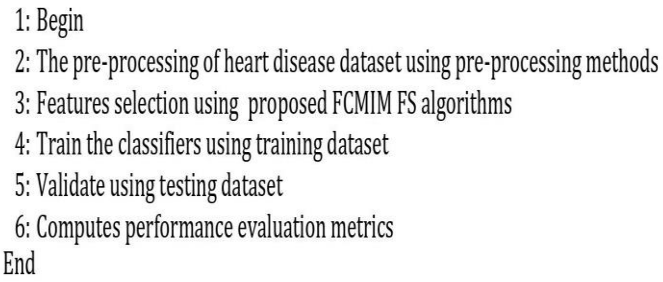
FP: are those cases which wrongly get classified as true but are false.

TN: are those cases which correctly get classified as false and are false.

FN: are those cases which wrongly get classified as false but are true.

1.6 Proposed Heart Disease Detection Methodology

The system was created with the goal of detecting cardiac disease. On a set of features, the performance of various machine learning classifiers for HD identification was evaluated. For feature selection, we presented the FCMIM algorithm. The performance of the predictors was assessed using feature sets chosen using proposed FCMIM algorithm. For the optimal model assessment, the LOSO approach of cross-validation was also applied. The model's effectiveness is evaluated by accuracy, specificity, sensitivity and computation time, which is determined for classifier evaluation.



```
 1: Begin
 2: The pre-processing of heart disease dataset using pre-processing methods
 3: Features selection using  proposed FCMIM FS algorithms
 4: Train the classifiers using training dataset
 5: Validate using testing dataset
 6: Computes performance evaluation metrics
End
```

Algorithm 2: Proposed Heart Disease Identification Methodology

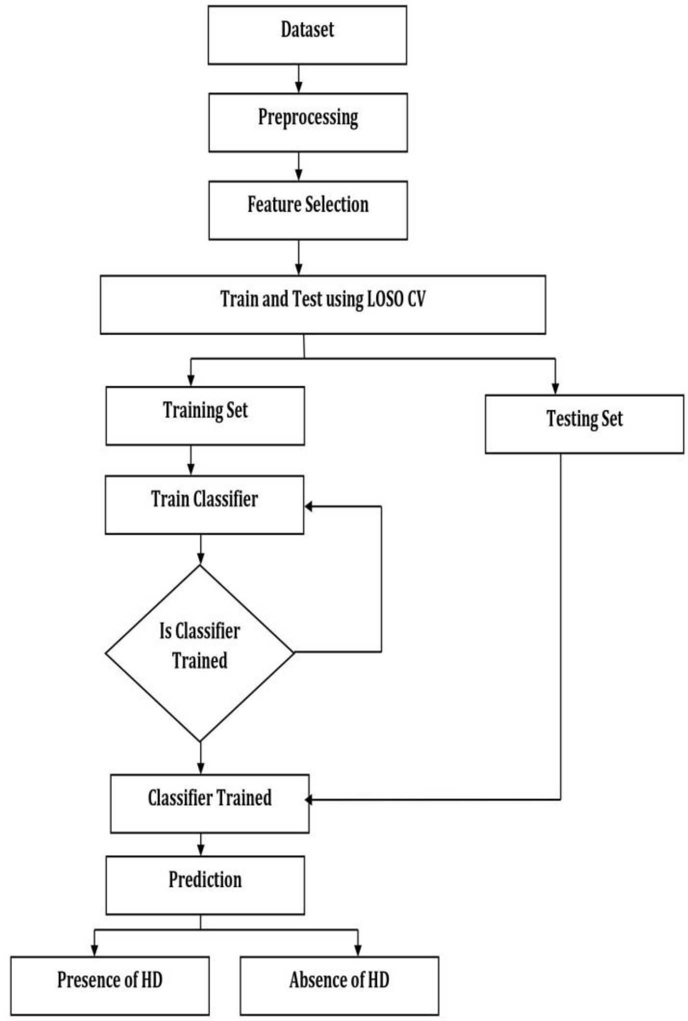


Fig -2: Proposed Heart Disease Identification Methodology.

4. EXPERIMENTAL RESULTS

1.1 Experimental Design setup

The dataset is taken from repository. The data is pre-processed to remove missing values and noise. The features are selected using FCMIM algorithm. The logistic regression classifier is used to detect heart disease. The LOSO cross validation method is used to train and test the classifier. The performance of the classifier is validated by using metrics. On an Intel(R) i7-

CPU all of the experiments were run in a python environment.

1.2 Experimental Results

The total number of rows with missing values is 489. Since, it is only 12 percent of the entire dataset, the rows with missing values are excluded from the dataset. The features taken to predict heart disease are sex, age, current smoker, cigarette per day, blood pressure, prevalent stroke, prevalent hyp, diabetes, total cholesterol, systolic BP, diastolic BP, body mass index, heart rate, glucose. The input graph is plotted in the chart 1.

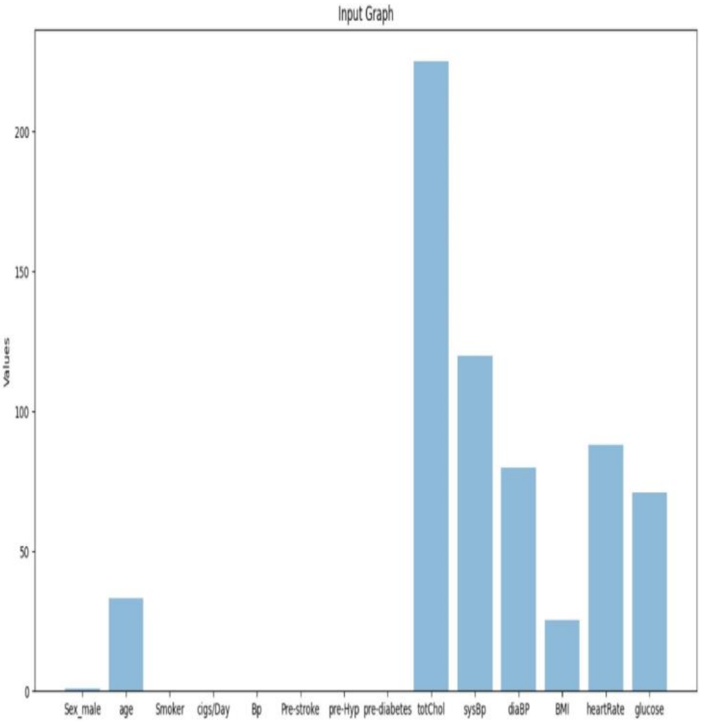


Chart -1: Input Graph

The logistic regression [12] ML model is trained by using 15 features along with one target class label. The results of logistic regression are as shown in the figure 3.

Table 2:

Result of LR Classifier

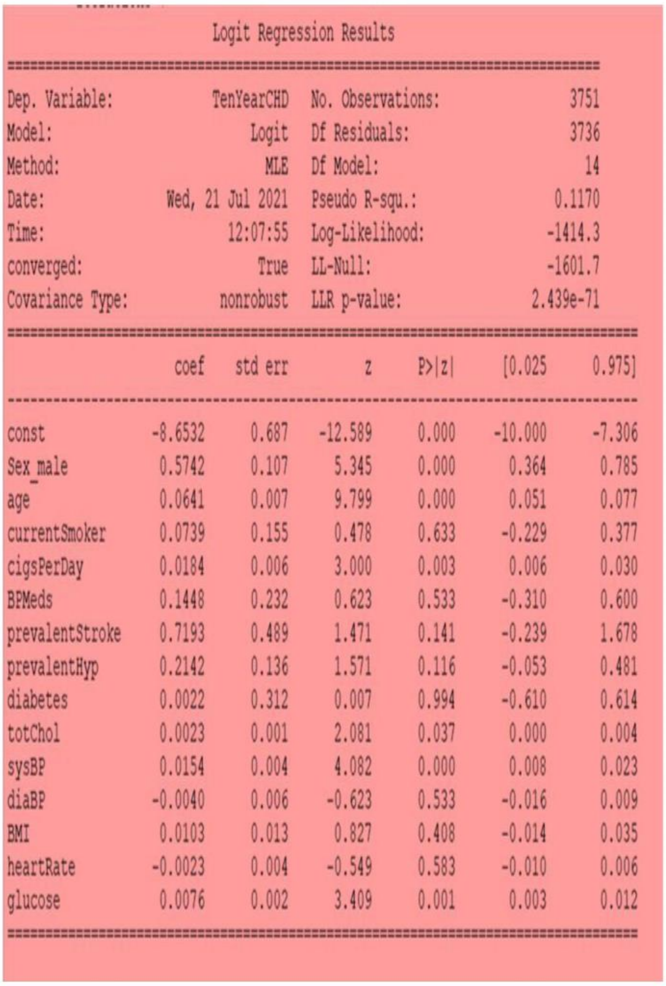
ML MODEL	Accuracy of the model	Sensitivity or True Positive Rate	Specificity or True Negative Rate	Positive Predictive Value	Negative Predictive Value	Misclassification Rate
Logistic regression	87.48	5.434	98.93	41.66	88.22	12.51

The confusion matrix with 0.1 threshold is [313 346] [12 80] and with 393 correct predictions and 12 Type II errors (False Negatives). The confusion matrix with 0.2 threshold is [519 140] [44 48] and with 567 correct predictions and 44 Type II errors (False Negatives). The confusion matrix with 0.3 threshold is [600 59] [64 28] and with 628 correct predictions and 64 Type II errors (False Negatives). The confusion matrix with 0.4 threshold is [640 19] [80 12] and with 652 correct predictions and 80 Type II errors (False Negatives).

Table 3:

Results of Sensitivity and Specificity with Different Threshold

Confusion Matrix	0.1 Threshold	0.2 Threshold	0.3 Threshold	0.4 Threshold
Sensitivity	86.95	52.17	30.43	13.04
Specificity	47.49	78.75	91.04	97.11



```
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 3751
Model:                          Logit   Df Residuals:                     3736
Method:                           MLE   Df Model:                           14
Date:                Wed, 21 Jul 2021   Pseudo R-squ.:                  0.1170
Time:                        12:07:55   Log-Likelihood:                -1414.3
converged:                       True   LL-Null:                       -1601.7
Covariance Type:            nonrobust   LLR p-value:                 2.439e-71
===================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
const             -8.6532      0.687    -12.589      0.000     -10.000      -7.306
Sex_male           0.5742      0.107      5.345      0.000       0.364       0.785
age                0.0641      0.007      9.799      0.000       0.051       0.077
currentSmoker      0.0739      0.155      0.478      0.633      -0.229       0.377
cigsPerDay         0.0184      0.006      3.000      0.003       0.006       0.030
BPMeds             0.1448      0.232      0.623      0.533      -0.310       0.600
prevalentStroke    0.7193      0.489      1.471      0.141      -0.239       1.678
prevalentHyp       0.2142      0.136      1.571      0.116      -0.053       0.481
diabetes           0.0022      0.312      0.007      0.994      -0.610       0.614
totChol            0.0023      0.001      2.081      0.037       0.000       0.004
sysBP              0.0154      0.004      4.082      0.000       0.008       0.023
diaBP             -0.0040      0.006     -0.623      0.533      -0.016       0.009
BMI                0.0103      0.013      0.827      0.408      -0.014       0.035
heartRate         -0.0023      0.004     -0.549      0.583      -0.010       0.006
glucose            0.0076      0.002      3.409      0.001       0.003       0.012
===================================================================================
```

Fig -3: Logistic Regression Result.

The comparison of existing system with proposed LR system is as shown in chart 2. The accuracy of the classifier is plotted on the x-axis and samples are plotted on y-axis.

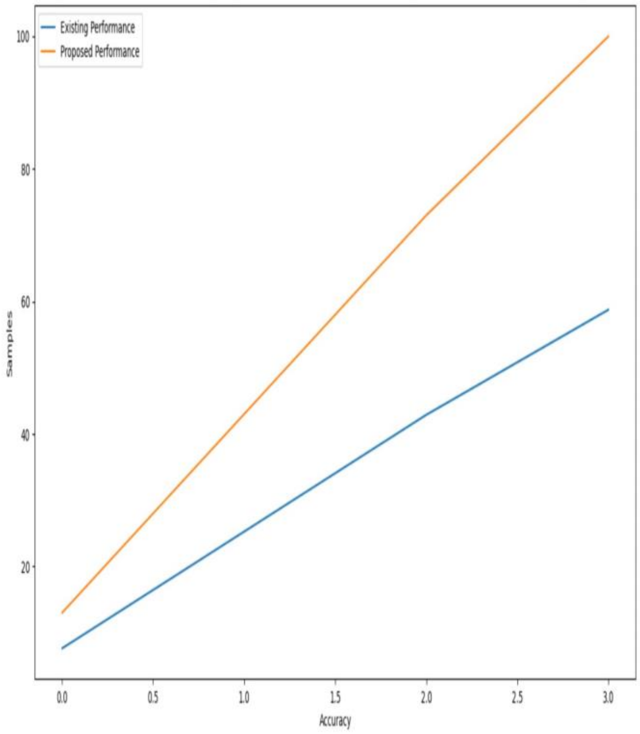


Chart-2: The comparison of Existing system with Logistic regression

An ROC curve is a graph that depicts a classifier's performance over all classification thresholds. The two parameters are plotted on the curve- TPR and FPR. An ROC curve depicts all of a classifier's points at different thresholds. The values of FPR are plotted along the horizontal axis (x-axis) and the values of TPR are plotted along the vertical axis (y-axis) in ROC space. It is shown in the figure 4.

TPR (True positive rate) It's the percentage of correctly categorised positive examples.

FPR (False positive rate) It's the percentage of negative examples that have been classified erroneously.

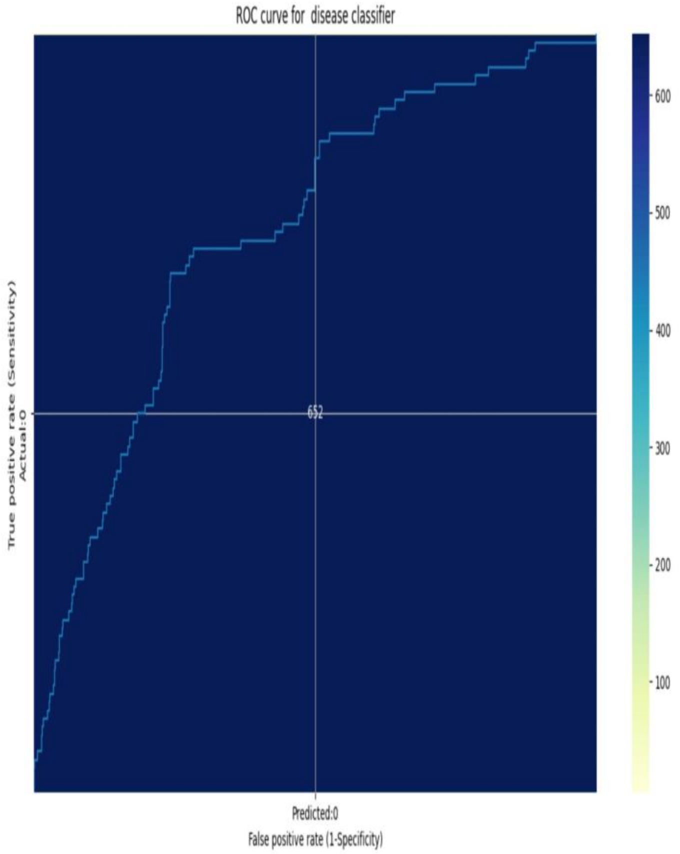


Fig -4: The ROC curve

5. CONCLUSIONS

In this work, an effective machine learning-based diagnosis system for the diagnosis of heart disease was built. The system is designed using machine learning classifier LR. FCMIM, is suggested to handle the feature selection problem. The system uses the LOSO cross-validation approach to determine the optimum model parameters. The technique is being evaluated on a dataset of Cleveland cardiac diseases. Additionally, performance evaluation measures are applied to quantify the system's performance.

We believe that establishing a decision model based on machine learning algorithms will be better appropriate for heart disease detection. We realize that incorrect features reduce diagnosis system performance and lengthen processing time. As a result, another novel aspect of our research was the use of feature selection technique to identify the right features, which improved accuracy rate while also minimizing the assessment system's process time.

REFERENCES

[1] G. Ankur, K. Rahul, Harkirat Singh Arora and Balasubramanian Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," IEEE Access, vol. 8, pp. 14659-14674, 2020.

[2] L. Dakun , Z. Yifei , . Z. Xinshu, S. Ye And Md Belal Bin Heyat, "An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers," IEEE Access, vol. 7, pp. 94701-94716, 2019.

[3] Devansh Shah, Samir Pate and Santosh Kumar Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, pp. 1-6, 2020.

[4] A. Golande and P. K. T, "Heart Disease Prediction Using Effective Machine Learning Techniques," IJRTE, vol. 8, no. 1S4, pp. 944-950, 2019.

[5] Martin Gjoreski, Anton Gradišek, Borut Budna, Matjaž Gams And Gregor Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure From Heart Sounds," IEEE Access, vol. 8, pp. 20313-20324, 2020.

[6] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian And Jongtae Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," IEEE Access, vol. 8, pp. 133034 -133050, 2020.

[7] M. Senthilkumar , T. Chandrasegar And S. Gautam , "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[8] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon, Shah Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," pp. 2-21, 2018.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, ''Effective heart disease prediction using hybrid machine learning techniques,'' IEEE Access, vol. 7, pp. 81542–81554, 2019.

[10] 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.

[11] Shakti Chourasiya and Suvrat Jain, "A Study Review On Supervised Machine Learning Algorithms," (SSRG-IJCSE), vol. 6, no. 8, 2019.

[12] Rajesh N, T Maneesha, Shaik Hafeez and Hari Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," International Journal of Engineering & Technology, vol. 7, pp. 364-366, 2018.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226 1238, Aug. 2005.

[14] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, arXiv:1811.12808. [Online]. Available: http://arxiv.org/abs/1811.12808

[15] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Inf. Syst., vol. 2018, pp. 1 21, Dec. 2018.