A Different Methodologies to Predict Heart Disease Using Machine Learning Techniques.

Kavitha B S
Sri Siddhartha Institute of Technology
Tumakuru
Karnataka

Dr. M Siddappa
Sri Siddhartha Institute of Technology
Tumakuru
Karnataka

Abstract

Heart disease is the severe human disease, and it has a major impact on people's health. Heart disease occurs when the heart fails to supply blood to every part of the body. Early detection of heart disease, can help to reduce mortality. Heart disease diagnosis must be accurate and timely in order to prevent and cure heart failure. In many ways, the diagnosis of heart disease based on standard medical history has been regarded unreliable. For differentiating healthy people and people with heart problems, non-invasive strategies like machine learning are accurate and effective.

The heart disease dataset can be used to build a machine learning system to predict heart disease. The 7 machine learning algorithms, 7 performance assessment parameters like specificity, sensitivity, Matthews corelation coefficient, accuracy, and cross-validation strategy, 3 feature selection methodology are used to find the people with heart problems.

The suggested methodology can quickly distinguish between those with heart disease and those who are healthy. The main objective is to identify heart illness using different methodologies of machine learning.

***Keywords*- Machine learning, Supervised learning, Non-invasive method, Confusion matrix, Cross-validation.**

I. INTRODUCTION

Heart disease (HD) is the catastrophic human disease [15]. The heart cannot circulate the sufficient blood to other parts of the human body to perform the body's regular functions, and as a result, heart failure develops.

Shortness of breath, physical weakness, swollen feet, and exhaustion are all indicators of heart illness.

In underdeveloped countries, it is difficult for early detection and treatment of heart illness due to shortage of specialists and equipment. The exact and proper evaluation of a patient's heart disease risk is essential for lowering their related risks of serious heart problems and increasing heart security.

The analysis of patient's medical history, examining physical report, and analysing symptoms are the common methods used by the doctors in invasive diagnostic approach. All of these procedures lead to erroneous diagnosis and, in many cases, delays in diagnosis findings owing to human error. Furthermore, it is more expensive, computationally complex, and time consuming.

The 7 machine learning classifiers like support vector machine (SVM), K-Nearest Neighbour (K-NN), Artificial Neural Network (ANN), Decision Tree (DT), Logistic Regression (LR), AdaBoost (AB), Naïve Bayes (NB) used by non-invasive method to identify heart illness.

A non-invasive system has been developed by a number of researchers and is widely utilised for heart disease diagnosis, and the ratio of heart disease deaths has dropped as a result of these machine-learning-based expert medical decision systems.

The university of California Irvine (UCI) data mining repository has a Cleveland heart disease dataset that has been used by many researchers. This is the dataset that various researchers have utilised to investigate various machine learning classification challenges linked to cardiac illness using various machine learning classification techniques.

II. LITERATURE SURVEY

Chieh-Chen Wu et al. [2] implemented an algorithm to identify MACE in patients with chest pains. Invasive and non-invasive procedures were used to generate the two models. A complete risk categorization model was constructed using these characteristics. Only non-invasive factors were used to build the reduced risk stratification model. The results reveal that both the full and reduced models function well and can predict MACE development within 90 days.

Ankur Gupta et al. [1] proposed a Machine intelligence framework to diagnose heart disease. The features from the UCI Cleveland dataset are chosen using factor analysis of mixed data (FAMD). 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.

Detrano et al. [3] proposed a decision support system based on logistic regression classifiers and an accuracy of 77% obtained. The Cleveland dataset was utilised in conjunction with global evolutionary techniques to obtain high prediction accuracy. The selection of features in the study was done using features selection methods. As a result, the approach's Classification performance is dependent on some features.

Gudadhe et al. [4] implemented SVM algorithms and multi-layer perceptron to identify heart illness and obtained an accuracy of 80.41%.

Kahramanli and Allahverdi et al. [5] implemented a hybrid technique that combines a fuzzy neural network and an

artificial neural network in a neural network to classify heart patients with normal patients. The proposed classification system also attained an accuracy of 87.4 percent in categorization.

Palaniappan and Awang et al. [6] proposed an expert medical diagnosis system for heart disease. The performance accuracy of 86.12% obtained by using Naïve Bayes model. The accuracy of 88.12% obtained by using ANN model and the decision tree classifier had an accuracy of 80.4 percent.

Olaniyi and Oyedotun et al. [7] proposed a three-phase model to identify heart disease in angina by using ANN, with an 88.89 percent classification accuracy. The proposed method might be simply integrated into existing healthcare data systems.

III. Materials and Methods

A. Dataset

The Cleveland heart disease dataset [9] is used by a number of researchers and is available via the University of California, Irvine's online data mining repository. This dataset was utilised in this study to develop a machine-learning-based method for diagnosing cardiac disease. There are 303 patients in the dataset, 76 features, and few missing values.

There are 297 samples with thirteen independent features, as well as a target output label that was extracted and used to diagnose heart disease. The six samples were removed due to missing values in the dataset. To depict a cardiac patient or a healthy subject, the target output label contains two classes. The extracted dataset has 297*13 features.

B. Methodology of the Proposed System

The proposed approach was created with the goal of distinguishing between person who have cardiac disease and those who are healthy. The performance of ML models was evaluated based on selected and full features. The important features were selected by using feature selection algorithms such as mRMR, LASSO, Relief.

The Cleveland heart disease dataset has been used in a number of studies, including ours. The system uses the machine learning classifiers [11] like logistic regression, K-NN, SVM, DT and NB. The suggested methodology is divided into 5 stages: (1) Dataset pre-processing, (2) Feature selection, (3) Machine learning classifiers (4) Cross-validation method, and (5) Performance evaluation methods for classifiers.

1) Data Pre-processing:

Data Pre-processing [12] is required to remove inconsistent data. The pre-processed data is used by the ML classifiers. This pre-processed data is used to train and test the ML classifiers. The dataset was subjected to pre-processing strategies such as missing value removal, standard scalar, and MinMax scalar.

Each feature has the same mean and variance in the standard scalar which results in same coefficient for all features. In MinMax Scalar, all the features having values between 0 and 1. The row with missing value is removed from the dataset.

2) Feature selection Algorithms:

Feature selection [13] is required to remove irrelevant features which impair the machine learning classifier's classification performance. The feature selection algorithms improve the accuracy of classification and reduces the execution time.

1. Relief Feature Selection Algorithm

All features in a dataset were assigned weights in relief algorithm and can adjust these weights over time. The weights of the important features to target are high, while the weights of the remaining features are low.

RELIEF Algorithm

Require: for each training instance set S, a vector of feature values and the class value

$n \leftarrow$ number of training instances

$a \leftarrow$ number of features

Parameter: $m \leftarrow$ number of random training instances out of n used to update W

Initialize all feature weights $W[A]$: 0.0

For k: =1 to m do

Randomly select a "target" instance R_k

Find a nearest hit "H" and nearest miss (instances)

For A: = 1 to a do

$W[A]$: = $W[A]$ – diff $(A, R_k, H)/m$ + diff $(A, R_k, M)/m$

End for

End for

Return the weight vector W of feature scores that compute the quality of features

Algorithm 1: Relief Algorithm

2.Minimal-Redundancy-Maximal-Relevance Feature Selection Algorithm.

The features which are linked to the target label were selected by mRMR [14]. In mRMR, the heuristic search strategy is utilised to find the best features with the most relevance and the least redundancy. It computes pairwise redundancy by checking one characteristic at a time. The mRMR is unconcerned with the joint association of features.

mRMR Algorithm

Input: initial features, reduced features
The initial feature is the number of features in original features set;
reduced feature is the required number of features
Output: selected features; // numbers of selected features
Relevance = mutual info (f_i, class);
Redundancy = 0;
For feature f_j in initial feature do
Redundancy ± mutual info (f_i, f_j);
End For
mrmrValue[f_i] = relevance − redundancy;
End For
Selected features sort (mrmrValues) take (reduced features);

ALGORITHM 2: MRMR

3. Least Absolute Shrinkage and Selection Operator.

The least absolute shrinkage and the selection operator choose features work by changing the absolute value of the features coefficient. The coefficient of some features becomes 0 and the features with zero coefficient are removed from the set.

With low coefficients feature values, the LASSO performs effectively. In selected feature subsets, features with high coefficient values will be included. Some irrelevant features can be selected in LASSO, and a subset of those features can be included.

3) Machine learning classifiers:

Machine learning [11] methods are used to classify cardiac patients and healthy persons. This paper goes through a few popular classification algorithms in detail.

1.Logistic regression.

A classification algorithm is a logistic regression where predicting the value of the predictive variable y is [0,1], 0 is the negative class and 1 is the positive class in a binary classification task. Multiclassification is also used to forecast the value of y when y is [0,1,2,3].

2. Support vector machine.

SVM is a machine learning classification technique [13] that has mostly been used to solve classification problems. SVM employed a maximum margin method, which resulted in the solution of a difficult quadratic programming issue. SVM is frequently used in numerous applications due to its good classification performance.

3. Naïve Bayes

The NB is a classification algorithm based on supervised learning [10]. The class of a new feature vector is determined using the conditional probability theorem. The training dataset is used by the NB to determine the conditional probability value of vectors for a particular class. The new vectors class is produced based on the conditional probabilities of each vector after computing the probability conditional value of each vector.

3. Artificial Neural Network.

The artificial neural network [8] is a mathematical model that combines neurons that pass signals. It is a supervised machine learning approach. Inputs, outputs and transfer functions are the three components of the ANN. The input units are given unusual values and weights, which are changed during the network's training process.

4. Decision Tree Classifier

A decision tree consists of leaf nodes, parent nodes or decision nodes. The method used by the decision trees are simple and straightforward in terms of how to make decision. Internal and external nodes were interconnected in a decision tree. The internal nodes are the parts of the system that make decisions. The leaf node has no child node and it consists of a label.

5. K-Nearest Neighbor.

K-NN is a supervised learning model which is used to classify the objects. The training samples are considered as points in the graph. The new sample is compared with training samples which is close to the new sample. The training samples which are close to the new sample becomes nearest neighbor of new sample. Let (x,y) be the training samples and h:X→Y be the learning function, so that h(x) can determine the y value given an observation x.

4) Validation Method of Classifiers:

The 4 performance assessment indicators and k-fold cross-validation are used in this study. The more information can be found in the subsections below.

1. K-fold cross-validation

The data set is partitioned into k groups of equal size in k-fold cross-validation, with k-1 groups used to train the model and the remaining part is used to test the model in every step. The above method is repeated k times. The performance of the classifier is calculated using k results. Before selecting training and testing new sets for the new cycle, for each fold of the process, the technique was repeated ten times, and all instances in the training and test groups were randomly distributed across the whole dataset. Finally, averages of all performance measures are computed at the end of the 10-step process.

2. Performance Evaluation Metrics

Various performance evaluation measures were employed in this study to evaluate the classifiers performance. The confusion matrix is used to predict in each step. The confusion matrix gives two correct and two wrong outputs.

TP: It is true positive (TP), indicating that the heart illness was accurately diagnosed and that the person has heart disease.

TN: A healthy person was appropriately categorised and that the subject is healthy because the projected output was true negative (TN).

FP: A healthy patient was wrongly labelled as having cardiac disease because the estimated output was false positive (FP) (a type 1 error).

FN: The heart illness was mistakenly labelled as false negative (FN) means the person is not having heart disease because the person is healthy (a type 2 error)

TABLE I:
CONFUSION MATRIX

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

A positive case indicates that it is unhealthy, whereas a negative case indicates that it is healthy.

Classification Accuracy: It depicts the classification system's total performance:

$$\text{Classification accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

Classification Error: It is the classification model's wrong classification:

$$\text{classification err} = \frac{FP+FN}{TP+TN+FP+FN} * 100\%$$

Sensitivity: It is the proportion of newly diagnosed heart patients to the overall number of heart patients. The "true positive rate" refers to the classifier's sensitivity for spotting positive cases. In other words, sensitivity verifies that test result is positive and the person has the disease.

$$\text{sensitivity (Sn)/recall/true positive rate} = \frac{TP}{TP+FP} * 100\%$$

Specificity: when the test result comes back negative and the subject not having heart disease.

$$\text{specificity (Sp)} = \frac{TN}{TN+FP} * 100\%$$

Precision: The formula of precision is

$$\text{precision} = \frac{TP}{TP+FP} * 100\%$$

MCC: It reflects a classifiers ability to predict with values between [-1, +1]. If the MCC classifier's result is +1, the classifier's predictions are perfect. The value -1 denotes that classifier's make utterly incorrect predictions. If the MCC value is nearer to 0, the classifier makes random predictions.

$$\text{MCC} = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} * 100\%$$

ROC and AUC:

The machine learning models ability for classification is examined using the receiver optimistic curves. ROC analysis is a graphical representation that contrasts the true positive rate and false positive rate in machine learning algorithm classification results. AUC is a measure of a classifiers ROC. The higher the AUC value, the more effective the classifiers performance will be.

IV. GENERAL PROCEDURE TO IDENTIFY HEART DISEASE.

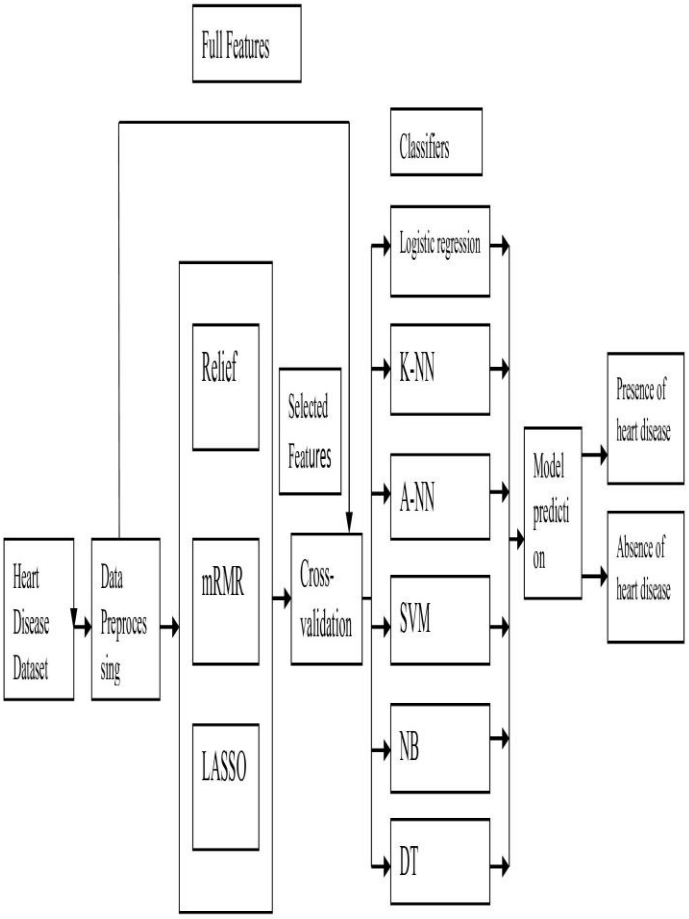

Fig. 1 The common technique to identify heart disease using machine learning technique.

The heart disease dataset taken from UCI Cleveland repository. The dataset may have noise, missing values, or be in an unusable format for machine learning classifiers. The data in the dataset needs to pre-processed to remove noise, replace missing values with the new values or remove missing values and convert data into an appropriate form that can be used by machine learning classifiers.

The features can be selected by using feature selection algorithms. The features play a very important role in predicting heart disease. These features are used to train the ML model. If the model is not trained by proper features, then model may not predict properly. The important features are selected by using relief, lasso, mrmr algorithms.

The features that separate the original data into training and testing samples are used in cross-validation procedures. The training data is used to train the machine learning model, while the testing data is used to determine whether the model has been properly trained. The selected features are fed into machine learning classifiers.

The classifiers learn from these features and predict the data whether the person having heart disease or not.

V. Conclusion

A machine-learning based predictive method for the diagnosis of heart disease was proposed in this research study. The method was put to test on a dataset of Cleveland heart disease patients. Three feature selection algorithms were utilised with 7 classifiers, including LR, K-NN, ANN, SVM, NB, DT and RF. The notable features were chosen using Relief, mRMR and LASSO.

Different assessment measures were also used to measure the performance of the ML model. The ML model performance is increased in terms of specificity, sensitivity, accuracy and MCC by using feature selection algorithm which is used to pick key features and it minimizes the algorithm processing time. The use of a machine-learning based technology to design a decision support system for heart disease detection will be more appropriate. Furthermore, certain non-essential features reduce the system's efficiency and lengthened the processing time.

The use of feature selection algorithms to identify the best features is another key part of this research, which improves classification accuracy while also reducing the diagnosis system's execution 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] Chieh-Chen Wu, Wen-Ding Hsu2, Yao-Chin Wang, Woon-Man Kung, I-Shiang Tzeng, Chih-Wei Huang, Chu-Ya Huang And Yu-Chuan Li1, *"An Innovative Scoring System for Predicting Major Adverse Cardiac Events in Patients With Chest Pain Based on Machine Learning,"* IEEE Access, vol. 8, pp. 124076 -124083, 2020.

[3] R. Detrano, A. Janosi, and W. Steinbrunn, *"International application of a new probability algorithm for the diagnosis of coronary artery disease,"* American Journal of Cardiology, vol.64, no. 5, pp.304-310,1989.

[4] M. Gudadhe, K. Wankhade, and S. Dongre, *"Decision support system for heart disease based on support vector machine and artificial neural network,"* in Proceedings of International Conference on Computer and Communication technology (ICCCT), pp. 741-745, Allahabad, India, September 2010.

[5] H. Kahramanli and N. Allahverdi, *"Design of a hybrid system for the diabetes and heart diseases,"* Expert Systems with Applications, vol. 35, no. 1-2, pp. 82-89,2008.

[6] S. Palaniappan and R. Awang, *"Intelligent heart disease prediction using data Mining techniques,"* in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008), pp. 108-115, Doha, Qatar, March-April 2008.

[7] E.O. Olaniyi and O.K. Oyedotun, *"Heart disease diagnosis using neural networks Arbitration,"* International Journal of Intelligent Systems and Applications, vol. 7,, no12, pp. 75-82, 2015.

[8] R. Das, I. Turkoglu, and A. Sengur, *"Effective diagnosis of heart disease through Neural networks ensembles,"* Expert Systems with Applications, vol. 36, no. 4, pp. 7675-7680,2009.

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

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

[11] 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.

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

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

[14] Jian Ping L, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan And Abdus Saboor, *"Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,"* IEEE Access, vol. 8, pp. 107562-107582, 2020.

[15] 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.