CINEFORUM
ISSN: 0009-7039
Vol. 66. No. 1, 2026

Disease Diagnosis for Healthcare System using Random Forest Machine Learning

[1]Prof. (Dr.) Rashmi Jha, [2]Lakshmi Kumari
Professor, Department of Computer Science and Engineering, MERI College of Engineering, Janakpuri, New Delhi[1]
Assistant Professor, Department of Computer Science, Institute of Information Technology and Management, New Delhi[2]
[1]rashmi.jha@meri.edu.in, [2]singhlakshmik@gmail.com

Abstract
Over one billion people worldwide suffer from high blood pressure, which is defined as having a diastolic pressure of 90 or higher and a systolic pressure of 140 or higher. Tobacco usage is estimated to cause about 10% of all cardiac problems, with approximately one billion people currently smoking worldwide. Elevated blood sugar levels, particularly in people with diabetes, significantly increase susceptibility to HD, a leading contributor to more than 60% fatalities among those with diabetes. The presence of elevated cholesterol levels in the bloodstream also increases the risk of developing coronary artery disease and stroke, accounting for 29% of cases of ischemic coronary disease. Despite the availability of tools and methodologies for the prediction of cardiac illnesses, there are still no effective models that can identify the disease. However, the increasing volume of data presents an opportunity for ML to play a pivotal role. As a subset of Artificial Intelligence (AI), ML has gained increasing popularity and is anticipated to become even more prominent. In the current digital era, remote monitoring devices collect vast amounts of patient data, contributing to 30% of global data generated by the healthcare industry annually, with an expected increase of 36% by 2025. This paper is predicting heart disease (HD) using different machine learning (ML) technique. The 10552 HD dataset is used for this paper and all technique is implement python software and calculate precision (P), accuracy (A) and recall (R). In all ML technique, the random forests (RF) provide good result compared to other technique.

Keywords – Heart Disease (HD), Random Forests (RF), Machine Learning (ML), Accuracy

I. INTRODUCTION
Heart disease is one of the WHO's biggest problems in medical innovation. Global causes of death from heart disease have changed dramatically, with life expectancy increasing significantly in the 21st century. Currently, there is an expected decline of about 30% globally, with the highest-wage countries experiencing a decline of about 40% and low- and middle-wage countries experiencing a decline of about 28%. Economic growth, suburbanization, and the associated changes in circadian rhythms are all contributing to this ongoing transformation, which is occurring at an even faster pace than in the last century, across all races, ethnicities, and countries around the world. Modern lifestyle changes have led to a significant increase in the number of heart failures in recent years. According to new research, the incidence of heart failure has increased over the past 25 years [1, 2].

Persistent and incurable diseases such as heart disease are the leading cause of death worldwide, according to new reports. Dramatic changes in health conditions around the world have led to an increase in heart disease worldwide. The leading cause of mortality worldwide every day is now heart illness. The substantial shift in people's health condition throughout the world is what causes the global rise in heart illness. Over the past 20 years, heart disease has alarmingly increased day by day and it is now one of the leading causes of death in most countries around the world. A new study focusing on cardiovascular health found that about 1.2 billion people die from heart disease every year. With huge variations in sociological, ethnic and economic backgrounds, there is no single solution to the growing burden of heart disease. Predicting heart failure has always been difficult due to its high cost. Modern imaging and clinical approaches to diagnose heart disease are too expensive. Palpitations, shortness of breath, fatigue, swelling, and chest pain are all symptoms of a heart attack or stroke. Most often, a blockage that stops blood flow to the heart or brain is the cause of these abrupt and dangerous events. Fatty deposits that develop on the inside walls of blood vessels that supply blood to the heart or brain are the most frequent cause of this. Bleeding or a blood clot from a brain artery causes a stroke. Rheumatic heart disease is still a major cause of morbidity and mortality in low- and middle-income nations, despite being uncommon in high-income nations like the US. The structure of cardiac mortality in 2017 is shown in Figure 1.

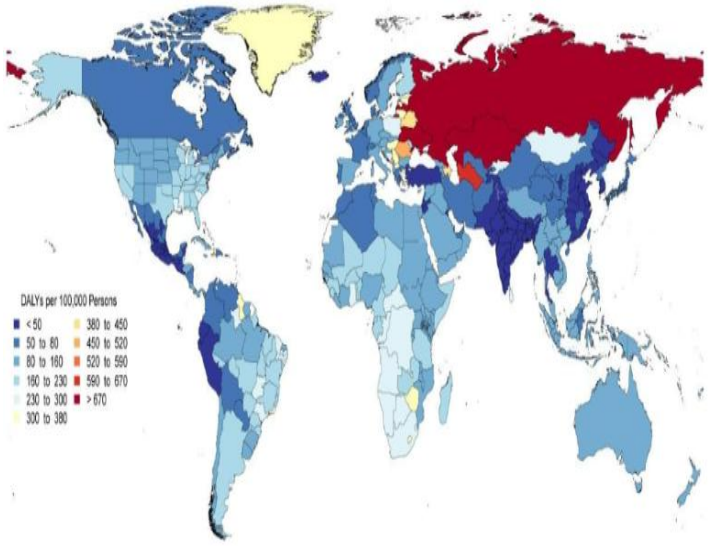

Figure 1: Structure of the Heart Disease Death Rate for the Year 2017

II. LITEATURE REVIEW

Advances in ML and computing capabilities have opened up many new opportunities in health research (O. Taylan et al. [1]). Various analysts have proposed ML calculations to improve the accuracy of disease prediction (R. Bhavani et al. [2]). To improve the accuracy of the results, many of the studies carefully checked for the presence of missing data in the dataset, which is an important aspect in the data pre-processing phase. I. Sutedja et al. [3] used different ML classifiers and the Pearson correlation coefficient to replace missing values in the Cleveland dataset. Karuna

Vishnu et al. [4] studied a large number of attributes using combined criteria (MICE) technique to solve the problem of missing properties, where other variables are assigned to each dataset variable during each iteration of a series of iterative predictive models to impute missing values. In another study [5], P. Rani et al. proposed an ANN imputation approach for predicting continuous variables (nearest neighbor mean) and categorical variables (most frequent). M. Diwakar et al. [6] used an LR model to classify cardiovascular diseases with an accuracy of 87.1% after cleaning the dataset and identifying missing features during preprocessing. M. Ganesan et al. [7] corrected the incorrect values. The feature selection method uses DT, LR, and Gaussian Naive Bayes (NB) algorithms to reduce the features from 13 to 4, achieving an accuracy of 82.75 percent. Priyan Malarvizhi et al. [8] improved the HD prediction accuracy of 297 datasets and 13 features of the Cleveland dataset by developing a hybrid RF model and a linear model. By applying UCI Cleveland dataset, used the XGBoost ML classifier to predict HD. Cervantes et al. [9] presents an HD prediction model developed using a classical LR algorithm. They evaluated the model on the Z-Alizadeh-Sani dataset. The algorithm LR achieved precision of 94.70% and 94.8% for the F1 score. It developed coronary HD diagnosis using the coronary arteriography technique with two-level stacking procedures. Using meta-classifiers, a final prediction is selected from base-level classifiers. The model was evaluated on 2020 patient data from the A. Javeed et al. [10] CHD dataset, achieving an accuracy of 95.3%.

III. SYMPTOMS OF HD

Since dyspnea is frequently brought on by intense exercise in healthy, well-trained individuals and by prolonged exercise in normal people who are not used to exercising, it should be regarded as abnormal if this symptom suddenly appears during physical work breaks. It is associated with all Huntington's diseases. It occurs only when resting, and almost always not when working hard. The effective onset of dyspnea is when there is increased awareness of breathing, caused by increased pain in the cardiovascular peak area, or a delay of more than 2 hours, a dull chest pain, and difficulty in taking enough air into the lungs [9].

2.1 Chest pain or discomfort

One of the most prevalent signs of Huntington's disease is chest pain, but it can be challenging to identify whether the heart is the cause or something else. One of the symptoms is also chest pain. Chest pain is often mistaken for heart failure. Heart disease is the most prevalent and dangerous condition that can cause chest pain if it is caused by other conditions, but heart pain itself is treatable. It describes pain in the jaw, head, arms, and other parts of the body as well as irritation, pressure, tightness, numbness, and other discomfort in the chest, neck, and upper trunk. It may continue for several minutes, a few days, or even a week, depending on your expectations. The meaning of chest pain is very unclear, and there are some medical cases of HD that cause symptoms [10].

Syncope is described as "a loss of consciousness that causes a decrease in blood flow to the brain". Since around 2017, syncope has been defined as "a sudden, temporary loss of consciousness" and is a common symptom that many people experience only once in their lifetime and is not a serious

illness. However, syncope can represent a dangerous, life-threatening condition. If syncope occurs, it must be diagnosed and treated. The causes of syncope are divided into two levels: neurological, metabolic, vasomotor, and cardiac. Syncope is a disease that causes sudden death. It can be caused by a variety of conditions that change the heartbeat. Palpitations, also known as "missing beats" or a fast, irregular heartbeat, are a common symptom. Some people with palpitations have arrhythmias, which are considered abnormal heart rhythms. Regardless of the cause, there are many different types of arrhythmias that can cause palpitations [11]. These side effects are normal for any illness. Weakness may be expected as an inability to function normally. Somnolence indicates lack of sleep or insomnia. Patients often fall asleep suddenly during the day. Indicative of HD, slurred displays indicate dysfunction of various organs of the body. Similarly, dizziness, lack of energy, fatigue, and lethargy usually require treatment to identify the specific cause. Disorders that cause nighttime sleep, such as restless legs syndrome and insomnia, are called somnolence [12].

IV. ML

ML proves invaluable, offering reproducible outcomes and the ability to learn from previous computations.

3.1 Supervised Learning

Supervised learning utilizes labeled data for classifying and solving problems, with regression and classification techniques as its two main branches. The regression analysis determines relationships among variables, indicating whether changes in explanatory variables are linked to changes in the dependent variable. In contrast, classification techniques assign objects to specific classes based on predefined criteria. Supervised learning methods represent the predominant approach in ML for predicting HD. These algorithms undergo training using a dataset comprising historical patient information, with each patient possessing a known label indicating the presence or absence of HD.

3.2 Unsupervised Learning

On the other hand, unsupervised learning lacks labeled data and introduces biases about the input's structure. When addressing CVD risk, regression techniques are essential to calculate an individual's risk based on actual numerical values associated with various risk factors [13]. In contrast, unsupervised learning algorithms analyze HD data without predefined labels, enabling them to uncover inherent patterns and relationships autonomously.

3.3 Reinforcement Learning

Within this framework, an agent is tasked with performing actions, and its effectiveness is contingent on its ability to comprehend the environment in which these actions occur. The agent maintains an internal state and interacts with the environment to achieve this understanding. A crucial aspect of this learning process involves using a reward function. The agent acquires knowledge about its environment by receiving positive or negative rewards based on its actions. The objective is to maximize positive rewards and minimize negative ones, encouraging the agent

to learn and adapt over time. It is noteworthy that in reinforcement learning, there is no obligatory reliance on human experts possessing domain-specific knowledge. Applying this concept to healthcare, particularly in the context of HD management, reinforcement learning could prove valuable. For instance, an intelligent system could adapt its decision-making processes to optimize patient care by continuously learning from the patient's health data and treatment outcomes.

V. PROPSOED METHODOLOGY

ML is a subject in which algorithms are used to instruct machines without the need for human intervention. We can train them to accomplish a specific work, and then use that training to handle similar jobs without having to explicitly program them. Training the machine with an algorithm and feeding it the dataset leads results in the construction of a classifier, which we test in the testing phase to determine its accuracy. Accuracy is always a major concern in the field of medical science, wireless communication, and different algorithms can provide varying degrees of accuracy when applied to the same data set. It is essential to determine which algorithm yields the best results in order to develop a better classifier and achieve better results when categorizing things. In today's world, virtually every industry makes use of machine learning in some capacity.

Random Forest: - Random forest machine learning classification methods are supervised machine learning classifiers that are applicable to both classification and regression. A random forest is made up of many decision trees working together. This algorithm uses bagging techniques to generate random features. Every decision tree in a random forest makes predictions, and the best answer is determined by voting. Average of all the decision trees is taken and the result is taken as prediction. Problem of over-fitting is also reduced by average of decision trees. Since, we have used weak tool for prediction purpose, so firstly dataset is loaded and the dataset consists of 8 features which represents the dataset's behaviour. Few samples from the dataset are selected randomly. Bagging technique of the random forest classifier is used to select n random features from total number of m features. Training algorithm trains the random sample and out of bag error is determined using those random samples.

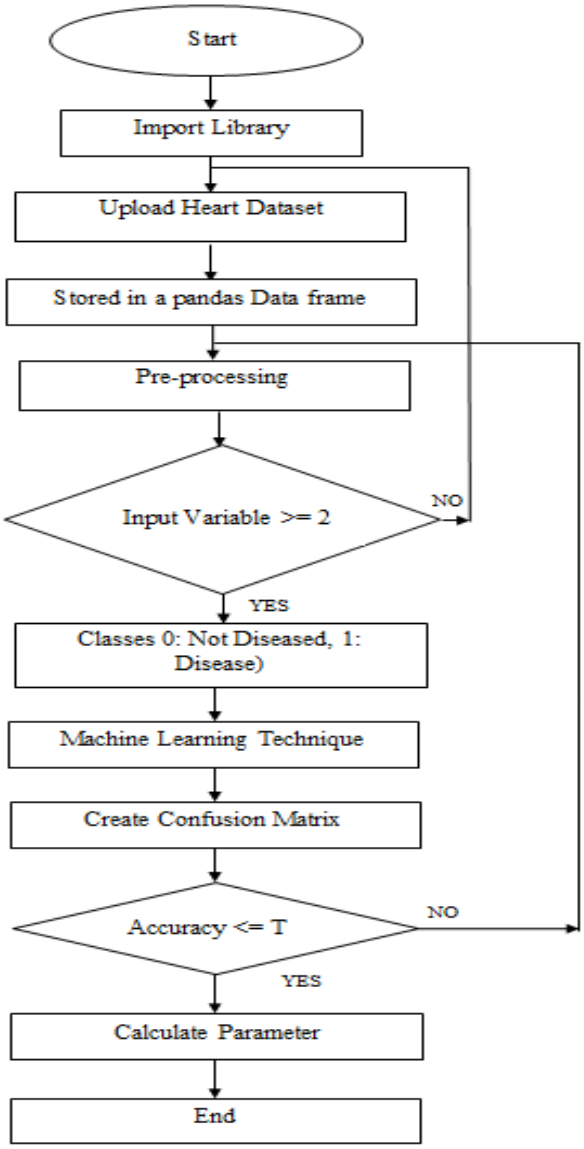

Fig. 2: Flow Chart of proposed model

This Classifier, on the other hand, by painstakingly fitting several noisy sets of data until every data point in the training set matched without error, it defies simple statistical norms. Perhaps more amazing, at least to statisticians, is that it will continue to improve a strategy that already removes generalization mistake.

Algorithm for Proposed Methodology: -

Step 1: Start

Step 2: Import Library: define all loading dataset, visualization, data preprocessing, data splitting, confusion matrix, machine learning and accuracy library

Step 3: Upload Dataset (files. upload command)

Step 4: Stored in a Pandas Data frame (pd.read_csv command)

Step 5: Pre-processing
Step 6: Variable (input variable i.e. age, sex, cp, chol, bp)
(6.1) if
(6.2) (Input variable >= 2)
(6.3) end if;
Step 7: Classes; 0 for not disease and 1 for disease
Step 8: Applied Machine learning Technique
Step 9: Create Confusion Matrix
Step 10: Threshold (T), maximum value of T is 100
(10.1) if
(10.2) (Accuracy >= T)
(10.3) end if;
Step 11: Calculate Parameter
Step 12: End

VI. SIMULATION RESULTS

Simulation Parameter

The accuracy of each fold indicates how well the model has learned from the training data and how well it can predict new data. The high accuracy of a fold suggests that the model can produce precise predictions and has successfully learned the data's underlying patterns. Therefore, Eq. can be used to measure the accuracy 1.

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

For a HD classification problem, its measures include Precision-Recall and accuracy. The formula to derive these measures is given in Eq. 2 and Eq. 3.

$$\Pr ecision = \frac{TP}{TP + FP} \quad (2)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (3)$$

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 3: HD Dataset

1. Histograms

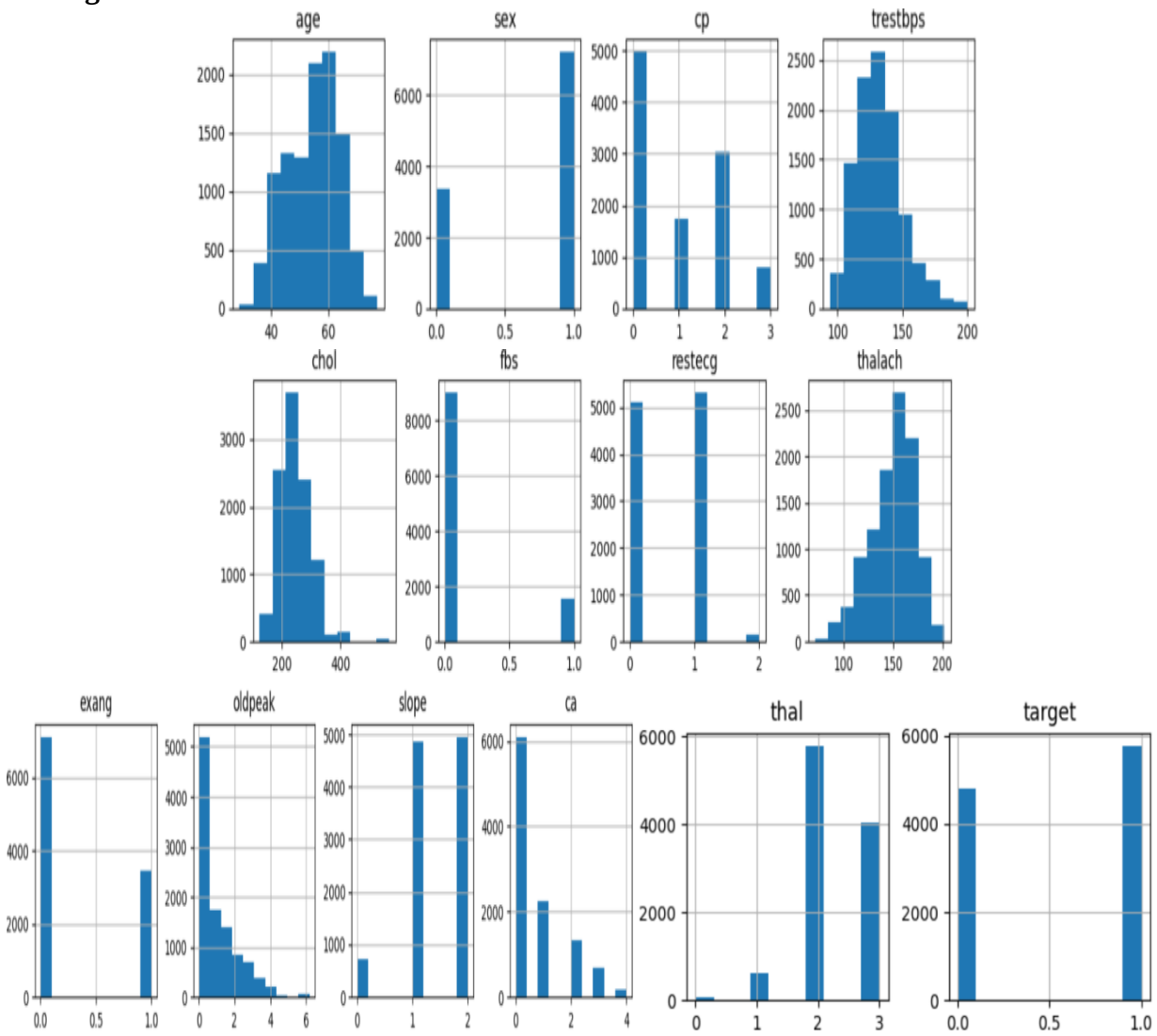

Fig. 4: Histograms of HD Dataset

2.Heart Disease Frequency for Age

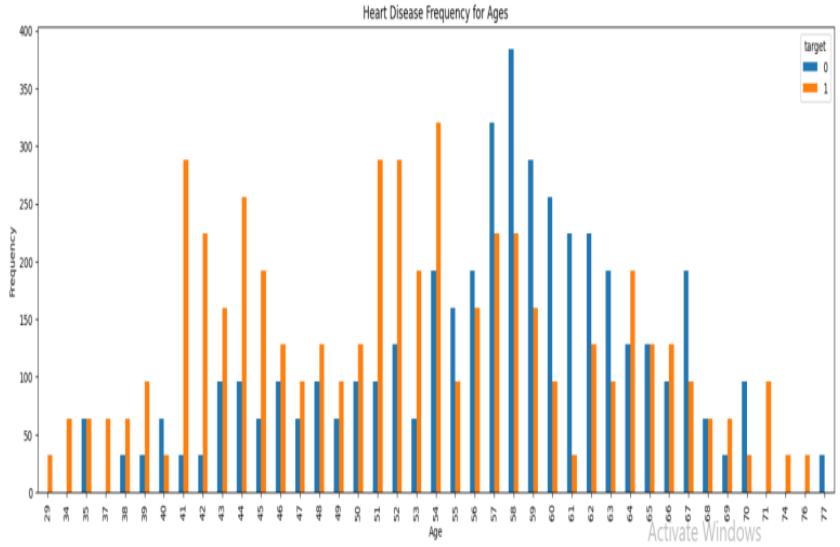

Fig. 5: Frequency of HD Dataset

3.Define Machine learning model

Table 1: Different ML Technique

Model	Precision	Recall	Accuracy	Misclassification
Logistic Regression (LR)	84.66%	90.28%	84.26%	15.74%
Naïve Bayes (NB)	86.11%	90.88%	85.88%	14.12%
K-NN	87.32%	92.88%	86.88%	13.12%
Decision Tree (DT)	91.32%	94.45%	92.92%	7.08%
Random Forest (RF)	98.49%	98.88%	97.76%	2.24%

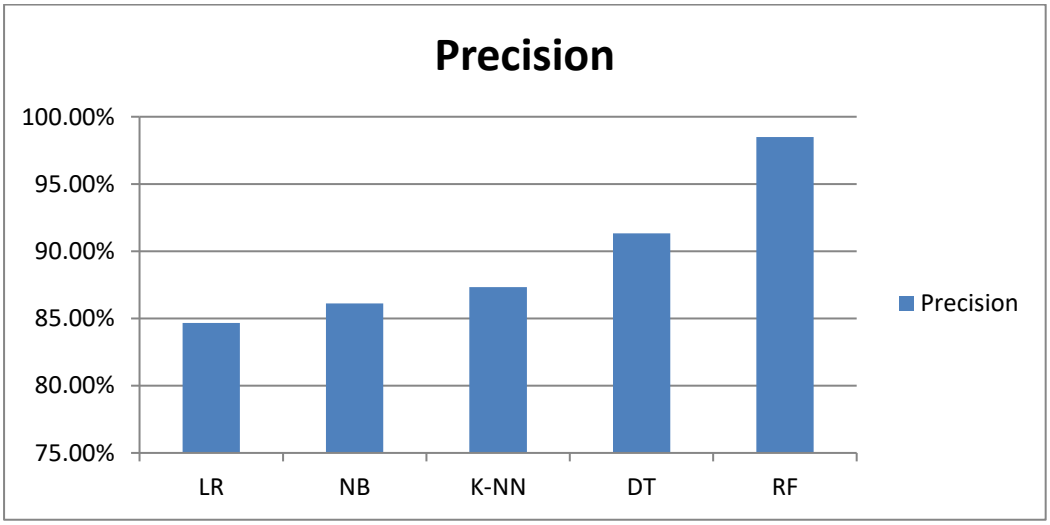

Fig. 6: Graphical Precision

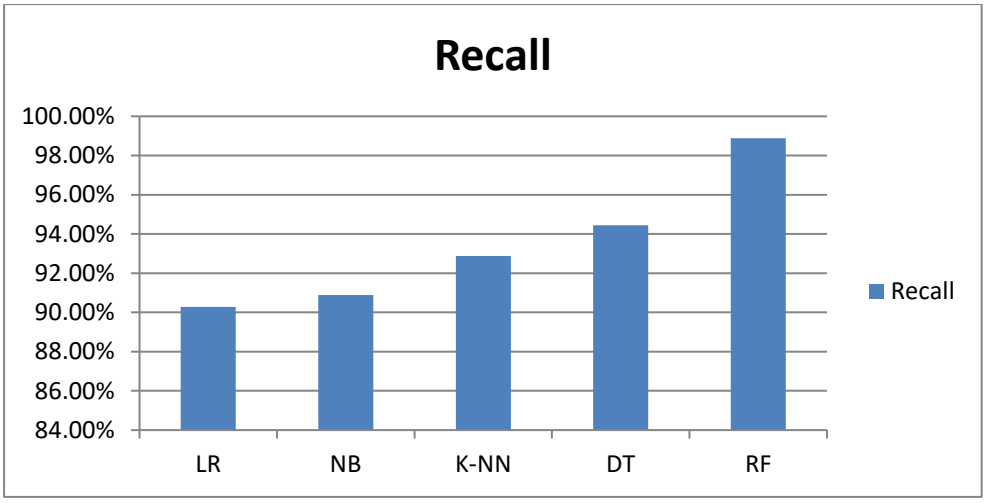

Fig. 7: Graphical Recall

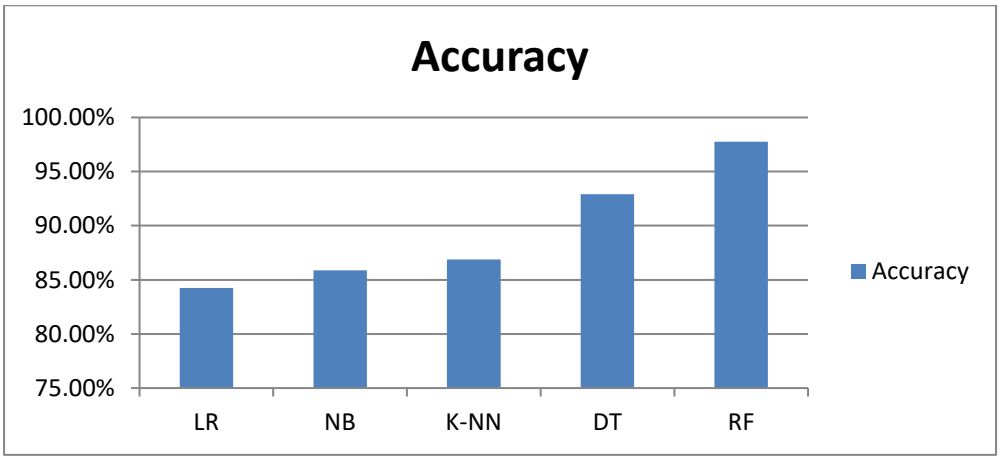

Fig. 8: Graphical Accuracy

Table II: Comparison Result

	Model	**Accuracy**	**Misclassification Rate**
Previous Technique	ANN-BFG	91.50	8.50%
	ANN-LM	96.20	3.80%
Proposed Technique	Random Forest	98.28%	1.72%

VII. CONCLUSION

Several machine learning methods based on DT, LR, KNN, NB, and RF are being studied in light of current research that demands for the automatic identification of risks and statistical numerical features results are age, cholesterol, resting blood pressure, ST-depression, and number of major

blood vessels. This work proposed a RF strategy utilized in prior investigations. The proposed RF is utilized to implement the techniques that worked the best.

REFRENCES

[1] O. Taylan, A. Alkabaa, H. Alqabbaa, E. Pamukçu and V. Leiva, "Early prediction in classification of cardiovascular diseases with machine learning neuro-fuzzy and statistical methods", *Biology*, vol. 12, no. 1, pp. 117, 2023.

[2] R. Bhavani, V. Ramkumar, V. Ravindran, R. Sindhuja and K. Swaminathan, "An efficient SAR image detection based on deep dense-mobile net method", *7th International Conference on Computing in Engineering & Technology (ICCET 2022)*, vol. 2022, pp. 92-95, 2022, February.

[3] I. Sutedja, "Descriptive and predictive analysis on heart disease with machine learning and deep learning", *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1-6, 2021, October.

[4] Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam and Hui Na Chua, "Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis", International Conference on Intelligent and Advanced Systems (ICIAS), pp. 01-05, IEEE 2021.

[5] P. Rani, R. Kumar, N. Ahmed and A. Jain, "A decision support system for heart disease prediction based upon machine learning", *Journal of Reliable Intelligent Environments*, vol. 7, pp. 263-275, 2021.

[6] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion", *Materials Today: Proceedings*, vol. 37, pp. 3213-3218, 2021.

[7] M. Ganesan and Dr. N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models", International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 01-05, IEEE 2019.

[8] Priyan Malarvizhi Kumar, Usha Devi Gandhi, "A novel Internet of Things architecture with machine learning algorithm for early detection of heart diseases", Computers and Electrical Engineering, Vol.65, pp. 222–235, 2018.

[9] Cervantes, J., García-Lamont, F., Rodríguez, L., Lopez-Chau, A., "A comprehensive survey on support vector machine classification: Applications, challenges and trends", Neurocomputing, Vol. 408, pp. 189–215, 2020.

[10] A. Javeed, S. Rizvi, S. Zhou, R. Riaz, S. Khan and S. Kwon, "Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification", *Mobile Information Systems*, pp. 1-11, 2020.

[11] C. Latah and S. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked 16*, pp. 100203, 2019.

[12] C. S. M. Wu, M. Badshah and V. Bhagwat, "Heart disease prediction using data mining techniques", *Proceedings of the 2019 2nd international conference on data science and information technology*, pp. 7-11, 2019, July.

[13] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of heart disease using machine learning", *2018 second international conference on electronics communication and aerospace technology (ICECA)*, pp. 1275-1278, 2018, March.

[14] Amin Khatami and Abbas Khosravi, “Medical image analysis using wavelet transform and deep belief networks”, Journal of Expert Systems with Applications, Vol. 3, Issue 4, pp. 190–198, 2017.

[15] T. Shu, B. Zhang and Y. Tang, "Effective Heart Disease Detection Based on Quantitative Computerized Traditional Chinese Medicine Using Representation Based Classifiers", *Evidence-Based Complementary and Alternative Medicine 2017*, pp. 1-10, 2017.