*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 10 Issue 1 ǁ 2022 ǁ PP. 30-36*

# Heart disease prediction using Hybrid machine learning Model

Ms. Yamini Ratawal[1], Sabya[2], Saurav[3], Swati[4]
*Department of Computer Science and Engineering Dr. Akhilesh Das Gupta Institute of Technology and Management*

***Abstract—*** *Around the world, heart disease is the leading cause of mortality. The study's main focus is on predicting cardiovascular disease in the real world. The prediction of heart disease is influenced by a number of risk factors. ML is proving its ability to aid decision-making, and it is based on the large quantity of data available in the medical services market. Machine Learning methods are used to predict heart infection using a variety of assays. We present a new method for detecting essential features using machine learning approaches to improve cardiovascular expectation accuracy in this paper. A number of highlights, as well as a number of well-known classification methods, are included in the early model.*

***Keywords—****Heart disease, Machine learning algorithms, hybrid random forest algorithms*

---------------------------------------------------------------------------------------------------------------------------------------

Date of Submission: 15-01-2022 Date of acceptance: 30-01-2022

---------------------------------------------------------------------------------------------------------------------------------------

## I. INTRODUCTION

Cardiovascular disease (CVD) is the most serious medical problem today. It is one of the world's most fatal and chronic illnesses, responsible for the majority of deaths. According to the World Health Organization (WHO), cardiovascular disease kills nearly 20.5 million people each year, accounting for roughly 31.5 percent of all deaths globally. The number of fatalities each year is predicted to increase to 24.2 million by 2030. About 85% of cardiovascular disease deaths are caused by heart attacks and strokes.

Plaque build-up in the arteries obstructs the blood flow to the heart, resulting in a heart attack. A blood clot in a cerebral artery blocks blood flow to the brain, resulting in a stroke. When the heart is unable to provide enough blood to the body's numerous organs, heart disease develops. Some of the early indicators include an irregular heartbeat, shortness of breath, chest pain, abrupt disorientation, nausea, swollen feet, and a cold perspiration. To improve patient survival rates, doctors must be able to properly foresee and identify cardiac illness in a timely way. CVD is caused by high blood pressure, cholesterol, alcohol, and cigarette usage, as well as obesity, physical inactivity, and hereditary changes. Early detection of symptoms, as well as lifestyle changes such as increased physical activity and quitting smoking, as well as expert medical examinations, can all help to reduce mortality. The medical sector demands an automated intelligent solution for accurate heart illness prediction. This may be achieved by using mixing machine learning algorithms with the large quantity of patient data now accessible in the medical area.

Disease prediction has gotten a lot of interest from data science research organisations in recent years. Machine learning gives the most frequent prediction modelling tools to overcome the current constraints. It has a lot of promise when it comes to translating large amounts of data and generating prediction algorithms. It employs a computer to learn intricate and non-linear correlations between features by minimising the discrepancy between the predicted and actual output. To predict the outcome, the computer learns patterns from the existing dataset's features and applies them to the unknown dataset.

For prediction, classification is a strong machine learning method. Classification is a supervised machine learning strategy that is successful in diagnosing the ailment when taught with enough data. This study made a significant contribution by employing contemporary machine learning algorithms to create an intuitive medical prediction system for the identification of heart disease. A number of algorithms may be used to accomplish data mining techniques. Python programming is gaining popularity because to these algorithms created using scikit learn packages. As a consequence, real-time implementation of data mining methods is more reliable than ever before. In the medical field, classification methods like as decision trees, naive Bayes, and SVM (Support Vector Machine) are accessible; similarly, regression techniques such as Random forest and logistic regressions are available. Machine learning algorithms are frequently utilised in the field of medical diagnostics.

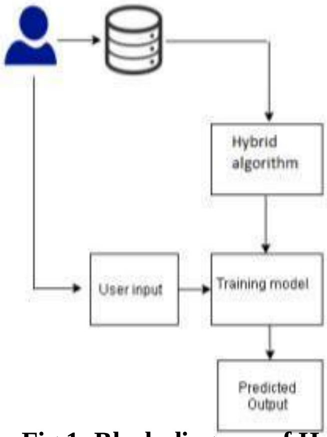


**Fig 1: Block diagram of Heart Disease prediction**

This study's purpose is to apply machine learning to predict heart illness using an automated medical diagnostic method.

Because it is the most accurate categorization strategy for predicting heart disease, the hybrid model is utilised disease. A hybrid model is a game-changing technique. which are the probability derived from a single machine learning algorithm. As input, the model is passed into another machine learning model.

Based on both machine learning and implementation concerns, this hybrid approach offers us with better-optimized outputs. To forecast cardiac illness, the proposed technique combines an automated machine learning diagnosis model with a high novelty-based hybrid model. This hybrid model can predict heart disease. This study makes use of the Kaggle dataset. This dataset is commonly used by machine learning researchers. In total, there are 303 instances in this collection, each with a unique attribute..

## II. RELATED WORK

Many recent studies on heart disease prediction and analysis are being conducted by researchers. Some of these works are discussed farther down.

In domains that are directly relevant to this project report, a lot of work is being done. ANN was developed to deliver the best level of prediction accuracy in the medical field. The back Propagation multilayer perception (MLP) of ANN is utilised to anticipate cardiac state. The acquired data is compared to current model findings in a similar location, and the results are found to be in need of improvement.

NN, DT, Support Vector Machines SVM, and Naive Bayes are utilised to uncover patterns in the UCI laboratory's database of heart disease victims. The outcomes of these algorithms are dependent on their execution and precision. By achieving an 85 percent conclusion rate, the suggested hybrid strategy exceeds the other present methods. The Sequence without Segmentation of Convolutional Neural Networks (CNN) is well-known.

N. K. Kumar et al. used a UCI dataset with 303 records and 10 characteristics to train five machine learning classifiers to predict cardiovascular disease: LR, SVM, DT, RF, and KNN. In comparison to the other classifiers, the RF classifier had the greatest accuracy of 85.71 percent and a ROC AUC of 0.8675. A. Gupta et al. used the factor analysis of mixed data (FAMD) approach to train LR, KNN, SVM, DT, and RF classifiers by replacing missing values with the majority label and produced 28 features using the Pearson correlation coefficient from the Cleveland dataset.; The accuracy of the results based on a weight matrix was 93.44 percent. M. Sultana et al. used Weka to test KStar, J48, sequential minimum optimization (SMO), BN, and MLP classifiers on a typical heart disease dataset from the UCA repository with 270 records and 13 characteristics; they found SMO to be the most accurate, with an accuracy of 84.07 percent. The author of [1] developed particular rules based on the PSO algorithm and compared them to get a more reliable rule for detecting heart disease. C 5.0 is used for illness classification using binary classification once the rules have been evaluated. For implementation, the author used data from the UCI repository and rated it as excellent using PSO and the Decision Tree Algorithm, we can improve accuracy. Using the Cleveland dataset with 297 records and 13 characteristics, S. Mohan et al. [5] created an efficient hybrid random forest with a linear model (HRFLM) to improve the accuracy of heart disease prediction. They came to the conclusion that the RF and LM approaches had the lowest error rates. S. Kodati et al. [6] used Orange and Weka data mining tools to construct a heart disease prediction system (HDPS) with the Cleveland dataset of 297 occurrences and 13 characteristics, evaluating the accuracy and recall metrics for the naive Bayes, SMO, RF, and KNN classifiers. In [7], the author described how data mining techniques may be used to predict cardiac disease. They used approaches including the KNN algorithm, neural network classifications, decision tree algorithm, and Bayesian classification algorithms to research and assess. The application of the genetic algorithm in feature selection for heart disease important traits was also investigated by the author. and experimented with the study and assessed good accuracy with the decision tree model analyses the computation of cardiac sickness using a variety of machine learning methods. The decision tree, KNN algorithm, SVM, and linear regression method are among the classification and regression models used in the study for prediction. According to the results of the trials, the KNN algorithm had the highest accuracy. On the other hand, this paradigm may be employed in real-time or in apps.

TUsing the Cleveland dataset with 297 observations and 13 features, ougui et al. [8] compared the performances of LR, SVM, KNN, ANN, NB, and RF models to identify heart disease using six data mining tools: Orange, Weka, RapidMiner, Knime, MATLAB, and Scikit-Learn. V. Pavithra et al. [9] used the UCI dataset of 280 cases to present a new hybrid feature selection technique integrating random forest, AdaBoost, and linear correlation (HRFLC) to predict heart disease.. After eleven (11) features were chosen using filter, wrapper, and embedding techniques, the hybrid model's accuracy increased by 2%. C. Gazeloglu et al. [10] projected 18 machine learning models using three feature selection procedures (correlation-based FS, chi-square, and fuzzy rough set) using the Cleveland dataset of 303 cases and 13 variables to find the best prediction combination for heart disease diagnosis.

The HD classification system was constructed by Detrano et al. [11] using machine learning classification techniques, the system's accuracy was 77 percent. With the global evolutionary technique and features selection method, the Cleveland dataset was employed. Gudadhe and his colleagues[12]. Using Perceptron and support vector machine (SVM) algorithms, they built a diagnostic system for HD categorisation that had an accuracy of 80.41 percent. Humar et al. [13] created a high-definition classification system based on a neural network with Fuzzy logic integration. The categorization system was 87.4 percent accurate. Resul et al. [14] developed an ANN ensemble-based diagnostic system for HD, obtaining an accuracy of 89.01 percent using enterprise miner (5.2) as a statistical measuring technique.

## III. PROPOSED WORK

The probabilities acquired from one machine learning model are fed into another machine learning model as input in a hybrid model, which is a new method. This hybrid model offers us with better-optimized outcomes because it is based on both machine learning algorithms that are chosen for execution.

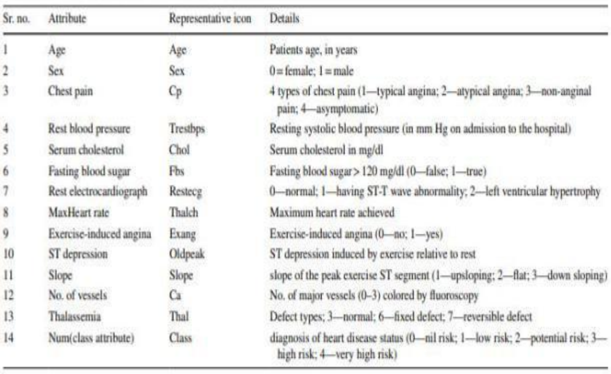
The given job is implemented using Scikit - learn libraries, pandas, matplotlib, and other essential libraries. The data was collected from the Kaggle repository. There are binary groups of cardiac disease in the retrieved data. The machine learning algorithm is used in conjunction with a hybrid model, such as a decision tree or a random forest..

***Data set details***

*A. Data set*

For the research, we used a dataset from the Kaggle library. It contains a real dataset of 304 data samples with 14 various characteristics, such as blood pressure, chest pain, ECG result, and so on (Fig 2). In this work, we used four algorithms to assess the causes of heart illness and build a model that was as precise as possible.

**Table: Data Set Description**



| Sr. no. | Attribute | Representative icon | Details |
|---|---|---|---|
| 1 | Age | Age | Patients age, in years |
| 2 | Sex | Sex | 0 = female; 1 = male |
| 3 | Chest pain | Cp | 4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic) |
| 4 | Rest blood pressure | Trestbps | Resting systolic blood pressure (in mm Hg on admission to the hospital) |
| 5 | Serum cholesterol | Chol | Serum cholesterol in mg/dl |
| 6 | Fasting blood sugar | Fbs | Fasting blood sugar > 120 mg/dl (0—false; 1—true) |
| 7 | Rest electrocardiograph | Restecg | 0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy |
| 8 | MaxHeart rate | Thalch | Maximum heart rate achieved |
| 9 | Exercise-induced angina | Exang | Exercise-induced angina (0—no; 1—yes) |
| 10 | ST depression | Oldpeak | ST depression induced by exercise relative to rest |
| 11 | Slope | Slope | slope of the peak exercise ST segment (1—upsloping; 2—flat; 3—down sloping) |
| 12 | No. of vessels | Ca | No. of major vessels (0–3) colored by fluoroscopy |
| 13 | Thalassemia | Thal | Defect types; 3—normal; 6—fixed defect; 7—reversible defect |
| 14 | Num(class attribute) | Class | diagnosis of heart disease status (0—nil risk; 1—low risk; 2—potential risk; 3—high risk; 4—very high risk) |

**Fig 2: Various attributes used for prediction**

The data is then categorized and divided into training and test data sets, which are subsequently subjected to various algorithms in order to get accuracy score results.

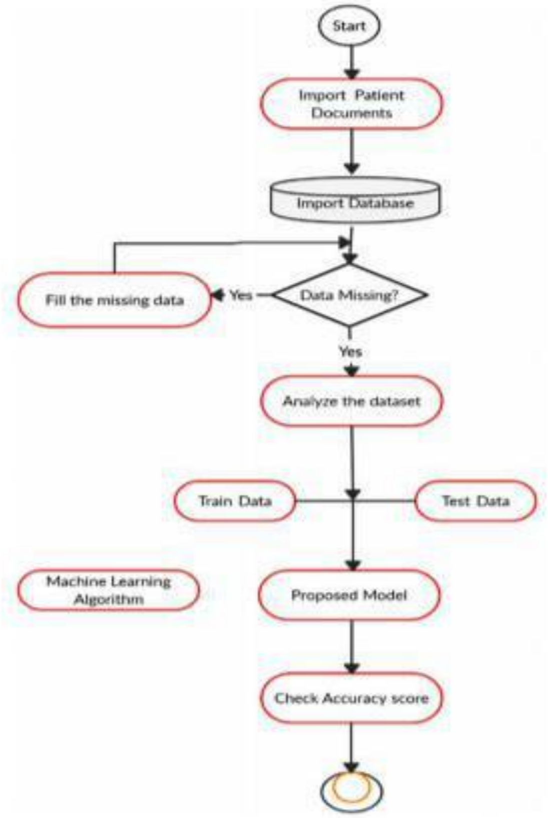


**Fig 3: Work Flow**

*B. Data pre-processing*

Large amounts of missing and noisy data are present in real-life data. These data are pre-processed to eliminate these flaws and produce confident forecasts. The sequential chart of our suggested model is depicted in Figure 1.

Noise and missing values are common while cleaning gathered data. The data must be cleansed of noise and missing values must be supplied in order to provide an accurate and effective result.

Transformation is the process of changing the format of data from one form to another in order to make it more understandable. It entails activities including as smoothing, standardization, and aggregation.

Integration **-** The data may not come from a single source but from several, and it must be combined before being processed.

Reduction **-** The information gathered is complicated, and it must be presented in order to yield useful results.

*C . Algorithms*

- ***Naïve Bayes classifier :-***

Naïve Bayes classifier is a supervised algorithm. It is a simple classification method using Bayes theorem. It assumes strong (Naive) self-dependence among attributes. All the attributes single-handed contribute to the probability to maximize it. It is suitable to work with the Naïve Bayes model and does not use Bayesian approaches. Many complex real-world problems use Naive Bayes classifiers :

P(X/Y) = P(Y/X) × P(X)

**P(Y)**

P(X/Y) is the posterior probability, P(X) is the class prior probability, P(Y) is the predictor prior probability, P(Y/X) is the probability of predictor.

The accuracy score achieved using Naïve Bayes is 85.25 %. ***K Nearest Neighbor (K NN) :-***

The K-nearest neighbor's algorithm is a supervised classification algorithm approach. It classifies things dependent on the nearest neighbor. It is a type of instance-predicated learning. The computation of distance of an character from its neighbors is measured applying Euclidean distance. It uses a group of titled points and uses them on how to label another point. The data are clustered based on similarity amongst them, and it is achievable to fill the missing values of data using K-NN. Once the missing values are filled, varied prediction approaches apply to the data set. It is possible to gain better closeness by applying various combinations of these algorithms. The accurateness score achieved using KNN is 67.21 %.

***Support Vector Classifier:***

SVM (Support Vector Machine) is a supervised machine learning algorithm which can be used for division and regression problems as support vector classification (SVC) and support vector regression (SVR). This classifier separates data points employing a fluttery plane with the most important quantity of margin. Support vectors are the data points which are closest to the fluttery plane. There are several kernels on which the hyper plane are constantly decided. This paper altogether focuses on four kernels videlicet linear, polynomial (poly), radial base function (RBF) and sigmoid. This type of classifier uses lower memory because they use a subset of showing points within the decision phase. The accuracy achieved by SVM on our data set is 81.97 %.

***Random Forest :-***

This algorithm comes under ensemble learning. Basic unit of this algorithm is decision Tree. Here we split the dataset into multiple datasets then build one decision tree for each sub dataset, then output of all the DTs are considered and majority voting is done for doing the prediction . Ensemble method always gives good result compared to single DT. In this, multiple trees are generated using bagging method. There are two types of voting in random forest. They are hard voting and soft voting. Proposed work uses hard voting, means the class which uses maximum number of votes from trained with ensemble methods. When the new patient's data needs to be predicted it is passed to the trained model and majority voting by different classifiers is taken to do the final prediction. Final prediction result is stored in database for further report generation.

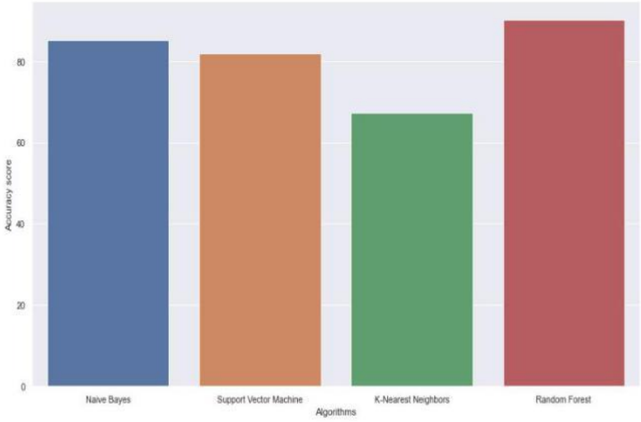
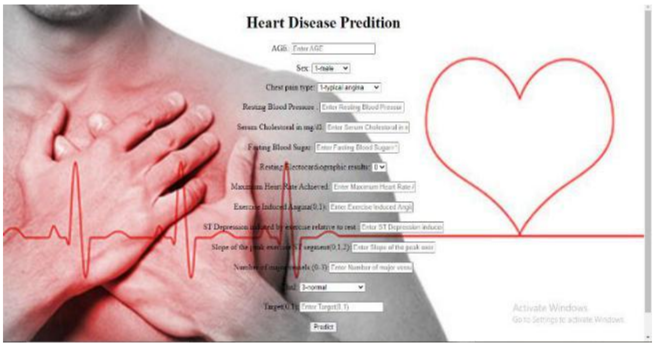


**Fig4 .Algorithms**

Markup Language which is employed to convert textbook into images, tables, links and other representations. CSS stands for Cascading Style Sheets which is a programming language that governs the presence of the webpage. The integration of the web operation created using HTML needs to be integrated with the model generated preliminarily to gain an operation which could successfully prognosticate the presence of heart disease in the individual. GCP is one of the primary options for cloud-based deployment of ML models, along with others similar as AWS, Microsoft Azure, etc. With Google App Engine using SDK installer who help to run the design on the home server.

Deployment is that the process of integrating a machine learning model into a longtime product system so as to form data-driven business decisions. It's one of the last way in the machine learning process, and it's also one of the most time-consuming.



Heart Disease Predition

AGE: Enter AGE

Sex: 1-male

Chest pain type: 1-typical angina

Resting Blood Pressure : Enter Resting Blood Pressur

Serum Cholestoral in mg/dl: Enter Serum Cholestoral in n

Fasting Blood Sugar: Enter Fasting Blood Sugar>

Resting Electrocardiographic results: 0

Maximum Heart Rate Achieved: Enter Maximum Heart Rate A

Exercise Induced Angina(0,1): Enter Exercise Induced Angi

ST Depression induced by exercise relative to rest : Enter ST Depression induce

Slope of the peak exercise ST segment(0,1,2): Enter Slope of the peak exer

Number of major vessels (0-3): Enter Number of major vess

Thal: 3-normal

Target(0,1): Enter Target(0,1)

**Fig 5. Deployed Home Page**

## IV. CONCLUSION

Among all the machine learning algorithms used, the highest closeness is achieved by Random Forest with 90.16%. This shows that the machine learning algorithms can be used to read heart disease effortlessly with different parameters and models. Machine learning is really useful in forecasting, cracking problems and other areas. Machine learning is an effective way to break problems in different areas too. Our ambition is to deliver every citizen with the availableness to identify pitfalls, if there are any, at an earlier stage so that they may be well set.

### *D.Model Deployment*

Front end development is that process of making a web operation using HTML and CSS languages which might be favorable for a stoner to interact with it and understand the contents of the webpage. HTML stands for the Hyper Text

## REFERENCES

[1]. Alkeshuosh, Azhar Hussein, et al. "Using PSO algorithm for producingbest rules in diagnosis of heart disease." 2017 international conference oncomputer and applications (ICCA). IEEE, 2017.

[2]. Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine LearningClassifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21.

[3]. Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access 2019, 8, 14659–14674. [CrossRef]

[4]. Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 20163rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka,Bangladesh, 22–24 September 2016; pp. 1–5.

[5]. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEEAccess 2019, 7, 81542–81554

[6]. Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai UniversityAnalysis of Heart Disease using in Data Mining Tools Orange and Weka. Glob. J. Comput. Sci. Technol. 2018, 18.

[7]. Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: Anoverview of heart disease prediction." International Journal of ComputerApplications 17.8 (2011): 43-48.

[8]. Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. HealthTechnol. 2020, 10, 1137–1144. [CrossRef]

[9]. Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. Mater. Today Proc. 2021,22, 660–670. [CrossRef]

[10]. Gazelo ˘glu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. Prog. Nutr. 2020,22, 660– 670.

[11]. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease", Amer. J. Cardiol., vol. 64, no. 5, pp. 304-310, Aug. 1989

[12]. M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network", Proc. Int. Conf. Comput. Commun. Technol. (ICCCT), pp. 741-745, Sep. 2010.

[13]. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", Expert Syst. Appl., vol. 35, no. 1, pp. 82-89, Jul. 2008

[14]. R. Das, I. Turkoglu and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Syst. Appl., vol. 36, no. 4, pp. 7675-7680, May 2009.

[15]. Ani, R., Krishna, S., Anju, N., Aslam, M. S., & Deepa.IOT based patient monitoring and diagnostic prediction tool using ensemble classifier. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).doi:10.1109/icacci.2017.8126068