How to Cite:
Vardhan, G. H., Reddy, N. S. S., & Umamaheswari, K. M. (2022). Heart disease prediction using machine learning. *International Journal of Health Sciences*, *6*(S2), 7804–7813.
https://doi.org/10.53730/ijhs.v6nS2.6955

Heart disease prediction using machine learning

Goud Harsha Vardhan
B.Tech Final year student, Department of Computer science and Engineering, SRM Institute of Science and Technology, Chengalpattu, India

Nallamilli Sneha Sisir Reddy
B.Tech Final year student, Department of Computer science and Engineering, SRM Institute of Science and Technology, Chengalpattu, India

Dr. K.M. Umamaheswari
Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Chengalpattu, India
Corresponding author email: umamahek@srmist.edu.in

***Abstract*---**Today, heart disease is a severe hazard to one's health that may result in death or a long-term impairment because of the complicated blend of clinical and pathological evidence required to diagnose it. Despite the fact that medical diagnosis is a complex activity that plays a critical role in saving human lives, there is a dearth of appropriate tools to detect hidden linkages and patterns in electronic health data. Computer-based automated decision support systems are needed to lower the costs of clinical testing because of this complexity. " In this study, we provide a method for predicting the presence of cardiac disease based on clinical data collected from patients. In this study, the primary goal is to develop a predictive model for heart disease based on a combination of characteristics (risk factors). Different machine learning classification strategies will be deployed and evaluated based on conventional performance metrics such as accuracy in order to compare different machine learning algorithms.

***Keywords*---**machine learning, diagnosis, heart disease, classification, prediction.

Introduction

Heart illness encompasses a wide spectrum of disorders. These include coronary artery disease, cardiac rhythm disorders (arrhythmias), and congenital heart defects. In the world's population, heart disease is the leading cause of illness and death. Prediction of cardiovascular illness is recognized as one of the most significant topics in clinical data analysis [1]. In the healthcare business, there is a massive quantity of data. Using data mining, a big collection of raw healthcare data may be transformed into useful information that can be used to make more educated choices and forecasts [8].

According to the World Health Organization, more than 10 million people die each year as a result of heart disease. Healthy living and early detection are the only ways to avoid heart-related illnesses [7]. Today's healthcare system has two major challenges: delivering high-quality services and making accurate diagnoses. Even while heart disease has been identified as the leading cause of mortality around the globe in recent years, it is also one of the most treatable.

Experts in the medical field have collected a massive amount of data that may be examined to get new insights. Extraction of significant and hidden data from massive volumes of data may be done using data mining methods. Typically, a medical database contains discrete data. Because of this, making decisions based on discrete data is difficult [2-3]. Machine Learning (ML) is a branch of data mining that is adept at dealing with huge, well-formatted datasets.

Invasive-based procedures for diagnosing heart disease are based on medical specialists' evaluation of the patient's medical history, physical examination report, and any symptoms that seem to be problematic. All of these procedures often result in inaccurate diagnoses and frequently cause diagnostic findings to be delayed owing to human error. In addition, it is more costly and computationally difficult, and it takes longer to evaluate [5-6].

“Noninvasive medical decision support systems based on machine learning prediction models such as support vector machine (SVM), k-nearest neighbor (K-NN), artificial neural network (ANN), decision tree (DT), logistic regression (LR), AdaBoost (AB), Naive Bayes (NB), fuzzy logic (FL) and rough set [9-10] have been created by different researchers to overcome these problems in invasive-based diagnosis of heart disease”. Our main objective is to diagnose heart disease in the early stage considering different factors like diabetes, blood pressure, age…etc. So that we can cure the disease or make it better.

Background

The use of a machine-learning-based system for heart disease diagnostics has been shown in several studies. The classification performance of several machine learning algorithms on the heart disease dataset has been presented in the literature review. El-Hasnony et al. [13] suggested a logistic regression classifier-based decision support system for heart disease categorization and achieved a classification accuracy of 77%. The use of global evolutionary techniques on the Cleveland dataset resulted in very accurate predictions. Feature selection strategies were utilized in the research to pick features. Because of this, it is important to pick the right characteristics in order to get the best classification results. MLP and SVM algorithms were employed [15] for the categorization of cardiac disease. The suggested categorization method was able to achieve an accuracy of 80%. A method for classifying cardiac illness was developed to use a hybrid neural network that combines a fuzzy neural network with an artificial neural network [16-17]. Classification accuracy of 87.4 percent was attained by the suggested classification method, which was tested.

Millions of people worldwide suffer from heart disease, which is still the leading risk factor for death. In order to lower the cost of diagnostic testing, doctors should use efficient, dependable, and computer-aided methods [14]. Using data mining software, computers may create and categorize different properties. Heart disease may be predicted using classification methods in this study. Machine learning and its methodologies, as well as data pre-processing, evaluation metrics, and a description of the dataset utilized in this study.

Machine Learning algorithms are very constructive algorithms in medical image analysis to diagnose a specific disease [7][11]. There are many divisions in machine learning like (i). supervised, (ii). semi-supervised, (iii).unsupervised, (iv).reinforcement and (v).active learning algorithms.

i. Supervised Learning: This algorithm generates a cluster of training instances with a respective objective to make the computer system learn accurately. Once the input is received as training sets, it is classified using classifications, and the outcomes of the machine are continuous instead of discrete. It is used for perceiving the root mean square error to acquire the accuracy by means of regression.
ii. Unsupervised Learning: Unlike supervised learning unsupervised learning took self-decision. It restores the hidden pattern by feature learning. The inputs are divided in form of clusters. The clustering relays on the group of resemblance.
iii. Semi-Supervised Learning: As the name obliques semi-supervised learning uses a partial training set of data. Hence semi-supervised learning is trained on both supervised and unsupervised features based on labeled and unlabeled sets of data.
iv. Active Learning: An active learning algorithm trains for only restricted datasets to enhance the optimality.
v. Reinforcement Learning: In this reinforcement learning algorithm, it acts as a response to the trained data based on the activity of the respective programs.

Heart disease is now being referred to as a silent killer, since it often results in death without evident signs. The illness's nature has led to a rise in fear about the sickness and its long-term effects. As a result, ongoing research is being conducted to try to forecast the occurrence of this devastating illness. As a result, new instruments and approaches are being tested on a regular basis to meet today's healthcare demands. Machine Learning methods may be quite useful in this situation. Even though heart disease may manifest itself in a range of ways, there is indeed a consistent set of key risk factors that impact whether someone is at risk for heart disease or not. By gathering information from many sources and taking it through a systematic classification and analysis process, we are able to draw a conclusion. This method is well-suited for use in the field of cardiovascular disease prognosis. Since "Prevention is better than cure," as the term indicates, it is possible to reduce the number of deaths caused by heart disease through early detection and management of the condition.

Methods and Materials

Data Collection

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI [4]. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system as shown in table 1.

Table 1. Feature selected in a dataset

Sl. No.	Attribute Description	Distinct Values ofAttribute
1.	Age- represents the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of the person (0-Female, 1-Male)	0,1
3.	CP- represents the severity of chestpain the patient is suffering.	0,1,2,3
4.	RestBP-It represents the patient's BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level ofthe patient.	Multiple values between 126 & 564
6.	FBS-It represents the fasting bloodsugar in the patient.	0,1
7.	Resting ECG-It shows the result ofECG	0,1,2
8.	Heartbeat- shows the max heartbeatof a patient	Multiple values from71 to 202
9.	Exang- used to identify if there is exercise-induced angina. If yes=1 or else no=0	0,1
10.	OldPeak- describes the patient's depression level.	Multiple values between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments (Unsloping, Flat, Down sloping)	1,2,3.
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for a patient suffering from pain in the chest or difficulty in breathing. There are 4 kinds of values that represent the Thallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. It is class or label, Colum. It represents the number of classes in the dataset. This dataset has binary classification i.e. two classes (0,1) In class "0" represent there is less possibility of heart disease whereas "1" represents high chances of heart disease. The value "0" Or "1" depends on the other 13 attributes.	0,1

Data Pre-processing

Prior to training and testing a machine learning classifier, data must be preprocessed to ensure an efficient representation of the data. The dataset has been preprocessed using methods such as eliminating missing values, standard scalar, and MinMax Scalar. Because the standard scalar assures that all features have a mean and variance of 0, it yields the same coefficient for all variables. In MinMax Scalar, the data is shifted such that all features fall between the range of 0 and 1.

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate. Pre-processing of data is required for improving the accuracy of the model.

System Architecture

Predicting cardiac disease using the four classification methods described above and analyzing performance. The study's goal is to accurately identify whether or not a patient has a cardiac disease. The health care provider inputs the patient's health report values into the system. Using this information, an algorithm is created that predicts the likelihood of developing heart disease. Figure. 1 illustrates the system architecture of the model.

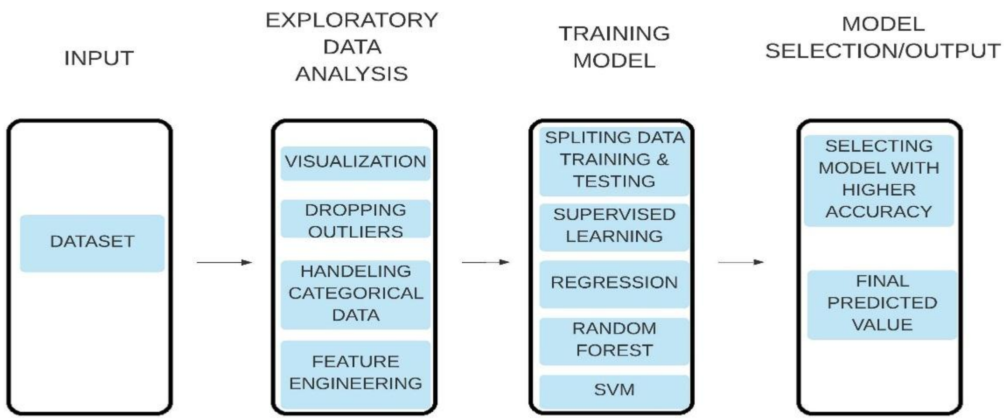

Figure 1. System Architecture

The attributes listed in Table 1 are used as input to many machine learning algorithms, including Random Forest, SVM, K-NN, Decision Tree, ANN, Logistic Regression, and Naive Bayes. The input dataset is divided into 80 percent of the training dataset and the remaining 20 percent of the testing dataset. Various measures, including accuracy, precision, recall, and F-measure scores, are used to calculate and analyze the performance of each method. The algorithms investigated in this study are mentioned below.

Naive Bayes

The NB is a classification method that uses supervised learning. Probability theory is used to decide which feature vectors belong in which class. Conditional probability values of vectors for a particular class are determined using the training dataset. Conditional probabilities are used to create a new vectors class after calculating the probability conditional value of each vector. NB is used to categorize text-related issues [12].

Logistic Regression

Predicting the value of the predictive variable y when it falls within the range [0, 1] is a binary classification issue in which the solution is either 0 or 1. It also employs multi-classification to predict y for y = [0,1,2,3]. There are 13 independent variables that make logistic regression good for classification [5].

Random Forest

In addition to classification and regression, Random Forest techniques are employed in a variety of contexts. It builds a tree out of the data and uses it to generate predictions. It is possible to utilize the Random Forest technique on huge datasets, even if substantial sets of record values are missing. The decision tree's produced samples may be preserved and used for new datasets. There are two steps in a random forest: first, generate a random forest, and then make a prediction using the random forest classifier that was built in the first stage [11].

Decision Tree

In a Decision Tree method, the central node reflects a dataset's properties, while the outside branches indicate the results [18]. Because of their pace, reliability, and ease of interpretation, decision trees are often used. In the Decision Tree, the prediction of the class label comes from the tree's root. Record values are evaluated from the perspective of root attribute values.

Support Vector Machine

SVM has pulled in a lot of consideration in the most recent decade and effectively tried for different space applications [18-19]. SVMs are generally utilized for learning characterization, relapse, or positioning capacities. SVM depends on measurable learning hypothesis and underlying danger minimization standard and has the aim of deciding the area of choice limits otherwise called hyperplane that produce the ideal partition of classes. SVM is the most vigorous and careful grouping strategy, there are numerous issues. The information examination in SVM depends on arched quadratic programming, and it is computationally exorbitant, as settling quadratic programming techniques requires huge network tasks just as tedious mathematical calculations.

Artificial Neural Network

As a supervised machine learning technique, the artificial neural network is a mathematical model that integrates neurons that convey signals [11]. The ANN is made up of three parts: inputs, outputs, and transfer functions. The input units receive remarkable values and weights that are updated throughout the network's training phase. The output of the artificial neural network is determined for the known class; the weight is recalculated based on the difference between the expected and actual output of the known class. Neurons are used to build an ANN. This various combination of neurons from various structures is analogous to multilayer perception.

K- Nearest Neighbor

One of the most popular supervised learning algorithms, K-NN, is used to classify data. With the K-NN technique, a new input may be predicted as to what class it belongs to based on how closely it resembles previous examples of that input in the training data set [20]. For example, if a training set contains examples that are identical to fresh data. Let (x, y) be the training observations, and let h: X Y be the learning function, such that given an observation x, h(x) can determine the value of the observation y.

Experimental Results

Similarity scores were utilized to compare the performance characteristics of several machine learning models that were assessed for the purpose of comparing their respective accuracy rates. Logistic regression has an accuracy of 0.86 for a given set of test data, whereas Random Forest has an accuracy of 0.90. As compared to other machine learning algorithms, the Random Forest model predicts Heart Disease with a greater accuracy.

In addition, the computing time was lowered, which is beneficial for deploying a model. An outlier identification method known as Random Forest may help identify data points with a non-Gaussian distribution. The problem here is that the dataset's sample size is quite small. Machine learning and other optimization approaches may also be applied to improve assessment outcomes. It is possible to employ a variety of various methods of normalizing the data and comparing the outcomes. More techniques might be explored to combine heart-disease-trained ML and DL models with specific multimedia for the benefit of patients and clinicians. In summary, inventory is an unpredictable mechanism that follows a chain segment, and its dependencies are unpredictable By doing Model fitting for the classification based algorithms, based on the accuracy percentages of all the algorithms, we are more likely to decide to use a random forest algorithm which suits the most for predicting whether the patient is likely to have heart disease or not. Figure 2 depicts the evaluation result of a different algorithm for the prediction of heart disease.

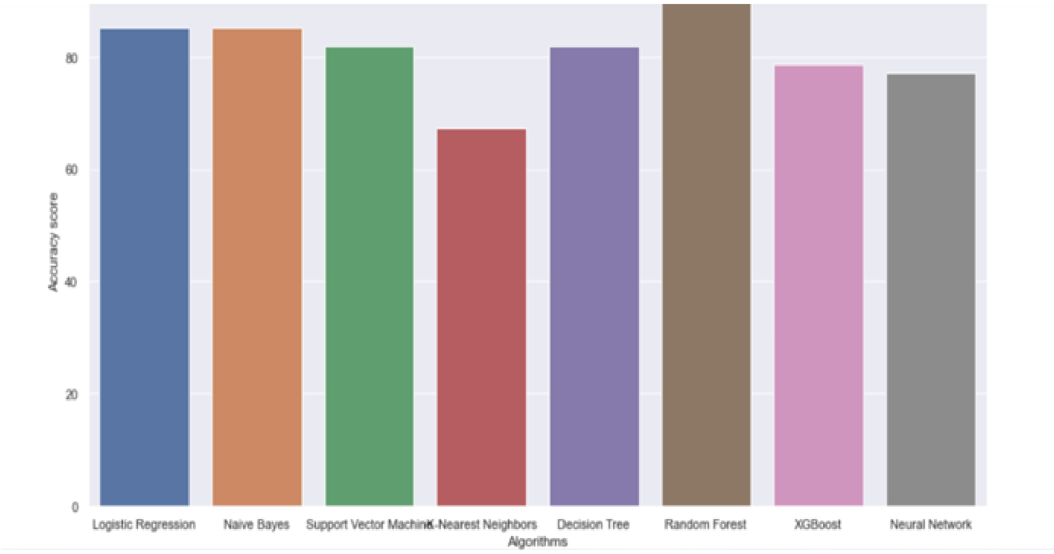

Figure 2. Performance evaluation of a different algorithm for the prediction of heart disease

Conclusion

Heart disease deaths are on the rise, making it necessary to develop a system that accurately predicts heart disease. One of the most important aims of the inquiry is the detection of cardiovascular disease. “Random Forest, SVM, K-NN, Decision Tree, ANN, Logistic Regression, and Naive Bayes algorithms are compared for their accuracy in predicting heart disease using the UCI machine learning repository dataset in this work”. “The Random Forest algorithm was shown to be the most accurate method for predicting heart disease, with an accuracy rate of 90.16 percent”. In the future, the development of a web-based framework depends on the Random Forest strategy, including the use of a huge sample than that employed in this experiment, which will help to increase the accuracy of the results and aid health professionals in better forecasting heart disease.

References

1. Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." In 2020 international conference on electrical and electronics engineering (ICE3), pp. 452-457. IEEE, 2020.
2. Rajdhan, Apurb, Avi Agarwal, Milan Sai, Dundigalla Ravi, and Poonam Ghuli. "Heart disease prediction using machine learning." International Journal of Research and Technology 9, no. 04 (2020): 659-662.
3. Haq, Amin Ul, 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." Mobile Information Systems 2018 (2018).
4. Gavhane, Aditi, Gouthami Kokkula, Isha Pandya, and Kailas Devadkar. "Prediction of heart disease using machine learning." In 2018 second international conference on electronics, communication and aerospace technology (ICECA), pp. 1275-1278. IEEE, 2018.

5. Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7, no. 2.8 (2018): 684-687.
6. Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1, no. 6 (2020): 1-6.
7. Dinesh, Kumar G., K. Arumugaraj, Kumar D. Santhosh, and V. Mareeswari. "Prediction of cardiovascular disease using machine learning algorithms." In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7. IEEE, 2018.
8. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.
9. Dwivedi, Ashok Kumar. "Performance evaluation of different machine learning techniques for prediction of heart disease." Neural Computing and Applications 29, no. 10 (2018): 685-693.
10. Nashif, Shadman, Md Rakib Raihan, Md Rasedul Islam, and Mohammad Hasan Imam. "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system." World Journal of Engineering and Technology 6, no. 4 (2018): 854-873.
11. Maji, Srabanti, and Srishti Arora. "Decision tree algorithms for prediction of heart disease." In Information and communication technology for competitive strategies, pp. 447-454. Springer, Singapore, 2019.
12. Ghosh, Pronab, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso De Boer. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." IEEE Access 9 (2021): 19304-19326.
13. El-Hasnony, Ibrahim M., Omar M. Elzeki, Ali Alshehri, and Hanaa Salem. "Multi-label active learning-based machine learning model for heart disease prediction." Sensors 22, no. 3 (2022): 1184.
14. Repaka, Anjan Nikhil, Sai Deepak Ravikanti, and Ramya G. Franklin. "Design and implementing heart disease prediction using naives Bayesian." In 2019 3rd International conference on trends in electronics and informatics (ICOEI), pp. 292-297. IEEE, 2019.
15. Buettner, Ricardo, and Marc Schunter. "Efficient machine learning based detection of heart disease." In 2019 IEEE international conference on E-health networking, application & services (HealthCom), pp. 1-6. IEEE, 2019.
16. Li, Jian Ping, 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 8 (2020): 107562-107582.
17. Safdar, Saima, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review." Artificial Intelligence Review 50, no. 4 (2018): 597-623.
18. Alotaibi, Fahd Saleh. "Implementation of machine learning model to predict heart failure disease." International Journal of Advanced Computer Science and Applications 10, no. 6 (2019): 261-268.
19. Reddy, G. Thippa, and Neelu Khare. "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model." Journal of Circuits, Systems and Computers 26, no. 04 (2017): 1750061.

20. Nilashi, Mehrbakhsh, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. "An analytical method for diseases prediction using machine learning techniques." Computers & Chemical Engineering 106 (2017): 212-223.