International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Cardiac Disease Prediction Using Machine Learning

Priyanka Yadav, Dr. Anita Pal
Computer Science & Engineering
Goel Institute of Technology and Management, Lucknow
1119priyanka@gmail.com, anita.pal@goel.edu.in

Abstract: Since the heart is one of the body's largest and most important organs, heart care is crucial. Since the majority of diseases are heart-related, it is necessary to predict heart diseases. To this end, a comparative study is required in this field. Since the majority of patients today pass away from diseases that were discovered too late due to inaccurate instrumentation, it is also necessary to understand more about more effective algorithms for predicting diseases. One effective technique for testing that is based on training and testing is machine learning. Machine learning is a subset of artificial intelligence (AI), a large field of study that focuses on creating computers that mimic human skills. However, machine learning systems are trained to process and utilise data; for this reason, the combination of these two technologies is sometimes known as machine intelligence. Since machine learning, by definition, learns from natural phenomena and objects, we employ biological parameters—such as blood pressure, cholesterol, sex, age, and so on—as testing data in our study. and based on these, a comparison is made on the accuracy of the algorithms. For example, in this project, four algorithms were used: SVM, k-neighbor, linear regression, and decision trees. In this research, we evaluate the accuracy of four distinct machine learning techniques and determine which is the best based on our calculations.

1. Introduction:

One of the most complicated and deadly illnesses affecting people today is heart disease (HD). With this illness, the heart often cannot pump enough blood to other areas of the body to maintain normal bodily functions. As a result, heart failure eventually sets in [1]. The United States has an extremely high rate of heart disease [2]. The symptoms of heart disease symptoms include peripheral edoema and increased jugular venous pressure brought on by functional cardiac or non-cardiac problems, as well as tiredness, swollen feet, and shortness of breath [3]. One of the main factors affecting the level of living is the intricacy of the early detection investigation procedures used to diagnose heart disease [4]. The diagnosis and treatment of heart illness are extremely complicated, particularly in underdeveloped nations where resources such as doctors and other medical professionals are scarce and diagnostic equipment is rarely available. These factors make it difficult to properly anticipate and treat heart patients [5]. Reducing the related risks of serious heart problems and enhancing heart security in patients requires an accurate and thorough identification of their heart disease risk [6]. According to the European Society of Cardiology (ESC), 3.6 million individuals worldwide receive a heart disease diagnosis each year, making 26 million diagnoses total. About 50% of patients with HD who have heart disease pass away during the first one to two years of their condition, and heart disease management expenses account for about 3% of the total cost of healthcare [7].The assessment of the patient's medical history, the results of the physical examination, and the evaluation of concerning symptoms by medical professionals form the foundation of invasive methods for detecting heart disease. Because of human error, most of these approaches result in imprecise diagnosis and frequently create delays in the diagnosis findings. In addition, evaluations require more time and are more costly and computationally difficult [8]. The development of a non-invasive medical decision support system based on machine learning predictive models, such as logistic regression (LR), AdaBoost (AB), fuzzy logic (FL), k-nearest neighbour (K-NN), artificial neural network (ANN), decision tree (DT), support vector machine (SVM), Naïve Bayes (NB), and rough set, was necessary to address the difficulties in invasive heart disease diagnosis. These machine learning-based expert medical decision systems have been widely used for heart disease diagnosis, and as a result, the ratio of heart disease death has decreased [11]. Numerous research papers have reported on the identification of heart disease using a machine learning-based approach. The literature review reports on the classification performance of several machine learning algorithms on the Cleveland heart disease dataset. Numerous researchers have used the Cleveland heart disease dataset, which is accessible online through the University of California, Irvine (UCI) data mining repository [12, 13]. This dataset has been utilised by several researchers to explore various classification problems associated with cardiac disorders using various machine learning classification techniques. Detrano and associates. [13] suggested a decision support system for heart disease categorization based on a logistic regression classifier, and it achieved a 77% classification accuracy. Using global evolutionary techniques with the Cleveland dataset, [14] produced highly accurate predictions. The characteristics in the research were chosen using feature selection techniques. As a result, some factors determine how well the technique performs in terms of categorization. Using multilayer perceptron (MLP) and support vector machine algorithms, Gudadhe et al. [15] developed a suggested classification system for heart illness and achieved an accuracy of 80.41%. In order to create a system for classifying cardiac diseases, Kahramanli and Allahverdi [16] employed a hybrid neural network approach that combined artificial and fuzzy neural networks. Additionally, 87.4% classification accuracy was

attained by the suggested categorization approach. An expert medical diagnosis system for cardiac illness was created by Palaniappan and Awang [17] using machine learning techniques such Naïve Bayes, decision trees, and artificial neural networks.The performance accuracy of the Naive Bayes prediction model was 86.12%. ANN, the second-best predictive model, with an accuracy of 88.12%, while the decision tree classifier produced an accurate forecast of 80.4%. In order to identify heart illness in angina, Olaniyi and Oyedotun [18] created a three-phase model based on the ANN and attained an 88.89% classification accuracy. Furthermore, implementing the suggested approach in healthcare information systems would not be difficult. Using the statistical analysis system enterprise miner 5.2 with the classification system, Das et al. [19] created an ANN ensemble-based prediction model that identifies cardiac disease and obtained 89.01% accuracy, 80.09% sensitivity, and 95.91% specificity. Jabbar and others. [20] created a heart disease detection system using a feature selection method and a machine learning classifier multilayer perceptron ANN-driven back propagation learning technique. The accuracy performance of the suggested system is outstanding. The authors in [12] developed an integrated decision support medical system based on ANN and Fuzzy AHP, which makes use of machine learning algorithms, artificial neural networks, and fuzzy analytical hierarchical processing to detect heart disease. Their suggested classification technique produced a 91.10% classification accuracy. The research's contribution is to provide a medical intelligent decision support system for heart disease detection that is based on machine learning. The present study classified individuals with heart disease and healthy individuals using a variety of machine learning prediction models, including logistic regression, k-nearest neighbour, ANN, SVM, decision tree, Naive Bayes, and random forest.Relief, minimum redundancy-maximum-relevance (mRMR), shrinkage, and selection operator (LASSO) are four more feature selection techniques that were utilised to identify the most significant and strongly correlated characteristics that had a significant impact on the target projected value. Also employed were cross-validation techniques like k-fold. Several performance assessment measures were employed to assess the classifier's performance, including receiver optimistic curves (ROC), Matthews' correlation coefficient (MCC), specificity, sensitivity, accuracy, and error.The computation of the model's execution time has also been done. Moreover, the heart disease dataset was pre-processed using several methods. The Cleveland heart disease dataset from 2016 was used to train and test the suggested algorithm. Cleveland heart disease dataset is accessible online via the UCI data-mining repository. Python was used for all calculations using an Intel(R) CoreTMi5-2400CPU @3.10GHz PC. The following are the main contributions of the suggested research project:

(a) The performance of each classifier has been examined in terms of execution time and classification accuracy for all characteristics.

(b) Using k-fold cross-validation, the classifiers' performances were evaluated on a subset of features chosen using the feature selection (FS) algorithms Relief, mRMR, and LASSO.

(c) The study makes recommendations for which feature algorithm works best with which classifier to create a high-level intelligent system for heart disease that can distinguish between individuals with heart disease and those who are healthy. The following sections of the document are organised as follows:

A brief overview of the theoretical and mathematical foundations of feature selection and classification techniques in machine learning is provided in Section 2's background information on the heart disease dataset. Performance evaluation metrics and the cross-validation approach are also covered. The experimental results are thoroughly detailed in Section 3. It ended

Section 4 is concerned with the conclusion of the paper.

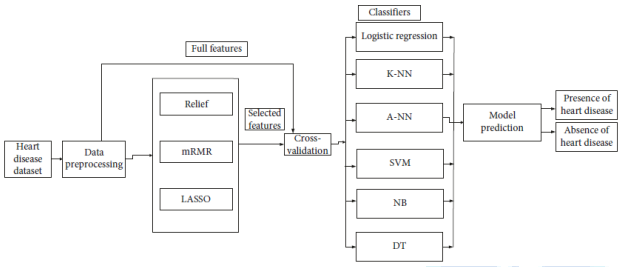


Fig. 1: A hybrid intelligent system framework predicting heart disease.

A few risk factors are under your control. In addition to the aforementioned variables, lifestyle choices such eating patterns, inactivity, and obesity are thought to be significant risk factors [5, 8, 15]. Heart disorders come in many forms, including myocarditis, cardio-myopathy, congenital heart disease, angina pectoris, congestive heart failure, and arrhythmias. Manually calculating the likelihood of developing heart disease based on risk factors is challenging [1]. Machine learning methods, however, are helpful for forecasting the result based on available data. Therefore, in order to estimate the risk of heart disease based on risk variables, this article uses a machine learning approach known as classification. It also employs an approach known as ensemble technique to attempt to increase the precision of heart disease risk prediction.

2. Review of Literature

1. S. Abdullah and R. R. Rajalaxmi, "A data mining model for coronary heart disease prediction using random forest classifier," published in 2012. The creation of a data mining model using the Random Forest classification method is the primary goal of the proposed study. The created model will include features like anticipating the happening of different events connected to every patient record, preventing risk factors and the cost metrics that go along with them, and increasing the prediction accuracy overall.

H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, "Using PSO algorithm for producing best rules in diagnosis of heart disease," 2017. The experimental results show that the PSO algorithm achieved higher predictive accuracy and much smaller rule list than C4.5. In this paper we proposed PSO algorithm for production of best rules in prediction of heart disease. The experiments show that the rules discovered for the dataset by PSO are generally with higher accuracy, generalization and comprehensibility. Based on the average accuracy, the accuracy of the PSO method is 87% and the accuracy of C4.5 is 63%. The task of classification becomes very difficult when the number of possible various combinations of parameters is so high. The self-adaptability of evolutionary algorithms depended on population is very useful in rule extraction and selection for data mining.

2. N. Al-milli, The article "Back propagation neural network for heart disease prediction" 2013,In this study, we describe a back propagation neural network based model-based method to detection of cardiac disease. This research study describes the development of a neural network-based heart disease prediction system. Thirteen medical variables were employed by the proposed algorithm to predict heart disease. The work's trials have demonstrated the suggested algorithm's strong performance in comparison to comparable state-of-the-art methods.
3. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of a heart disease prediction system based on neural networks, 2018. Using data mining and artificial neural network (ANN) approaches, we have presented the Heart Disease Prediction System (HDPS) in this research study. The system is developed using a multilayer perceptron neural network and back propagation method, derived from the ANN. Because the MLPNN model functions realistically well even without retraining, it demonstrates superior results and aids domain specialists and even those with a connection to the area in planning for a better diagnosis and providing the patient with early diagnosis findings.
4. P. K. Anooj, "Weighted fuzzy rules for risk level prediction of heart disease in a clinical decision support system," 2012. This is a laborious procedure that mostly relies on the subjective judgements of medical specialists. Machine learning algorithms have been created to automatically learn from examples or raw data in order to address this issue. In this case, a weighted fuzzy rule-based clinical decision support system (CDSS) that automatically learns from the patient's clinical data is provided for the detection of heart disease.
5. L. Baccour, "Some UCI data sets using modified fused TOPSIS-VIKOR for classification (ATOVIC)" (2016) An essential function of expert and intelligent systems is the classification process. The creation of new categorization algorithms with higher accuracy or true positive rates may have an impact on some real-world issues, such as the ability to anticipate medical diagnoses. It is anticipated that multi-criteria decision making (MCDM) techniques would look for the optimal option based on a number of factors. Every criterion has a value in relation to every option.

3. Software Requirements

Methods for determining which of them is the best and obtaining the outcome in SVM'sfavour. The UCI machine learning dataset, which consists of 303 samples with 14 input features, was used to train the various machine learning and data mining algorithms that Kumar et al.[5] worked on. The results showed that svm is the best algorithm among them; other algorithms that were used here included naivybayes, knn, and decision trees. Gavhane and others[1] have been working on the multilayer perceptron model for human heart disease prediction and algorithm accuracy utilising CAD technology. The more people who use the prediction system to forecast their ailments, the more people will become aware of these conditions, which will lower the death rate for cardiac patients. A couple of algorithms for illness prediction have been developed by several academics. In their study, Krishnan et al.[2] shown that decision trees are more accurate than naïve bayes classification algorithms. Machine learning algorithms are used to predict many kinds of illnesses, and several researchers—including Kohali et al.—have worked in this area. [7] Research was done on the prediction of heart disease using logistic regression, diabetes using support vector machines, and breast cancer using Adaboot classifiers. The results showed that the logistic regression had an accuracy of 87.1%, the support vector machine had an accuracy of 85.71%, and the Adaboot classifier had an accuracy of up to 98.57%, all of which are good from the standpoint of prediction. A review report on the prediction of cardiac illnesses has demonstrated that, in contrast to hybridization, which performs well and provides improved prediction accuracy, the older machine learning algorithms do not perform well in terms of accuracy.(8).

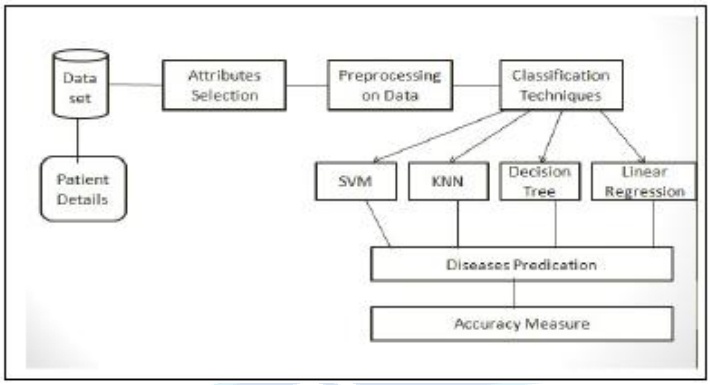


Fig.2. Architecture of Prediction System

C. Preprocessing of data

Preprocessing is necessary for the machine learning algorithms to provide a distinguished outcome. For instance, the Random Forest technique does not allow datasets with null values; thus, we must handle null values from the original raw data. We must use the following code to convert some categorised values into fake values, which are represented as "0" and "1," for our project:

D. Data Balancing

Since the data balancing graph shows that the two target classes are equal, data balancing is necessary for accurate results. The target classes are shown in Fig. 4, where "0" denotes a patient with heart disease and "1" denotes a patient without heart disease.

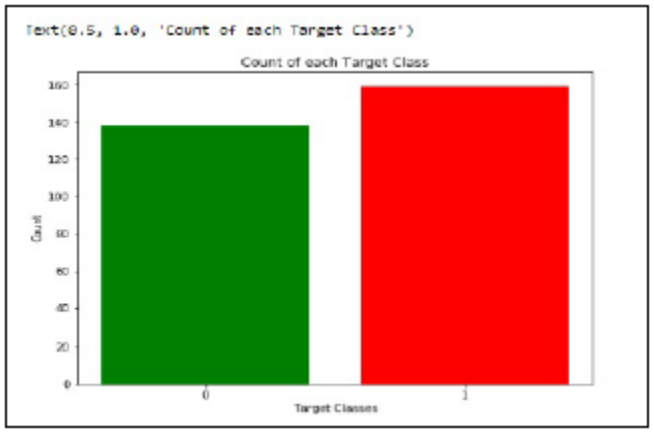
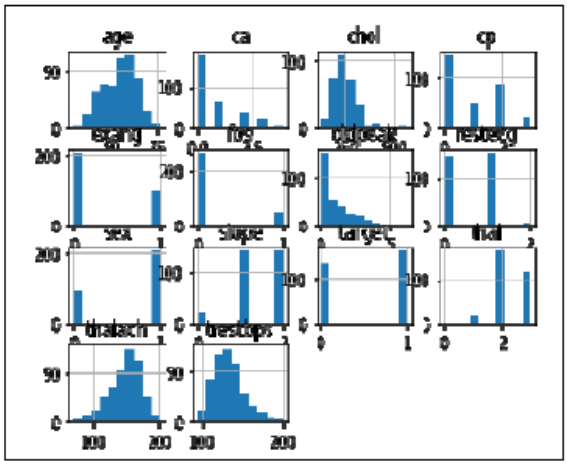


Fig.3. Target class view

E. Histogram of attributes

The dataset's range of attributes and the code used to build it are displayed in the attributes histogram. Collection.hist()



4. Related Work

Analysis of heart-related matters is always necessary, whether for diagnosis, prognosis, or heart disease prevention.This study has contributions from a variety of domains, including data mining, machine learning, and artificial intelligence. Any algorithm's performance is dependent on the dataset's bias and variance[4]. According to study by Himanshuet al.[4] on machine learning for heart disease prediction, naïve bayes performs better with low variance and high biasness than with high variance and low biasness, or knn. The reason for the decline in knn performance is because it has an overfitting issue when there is low biasness and large variation. Utilising low variance and high biasness has several benefits as it requires less time for algorithm testing and training due to the smaller dataset size, but there are drawbacks as well. As the size of the dataset increases, asymptotic mistakes become more prevalent, and low variance based algorithms perform better in these kinds of situations. One nonparametric machine learning approach that is susceptible to overfitting is the decision tree. However, there are strategies available to address this issue, such as overfitting removal. Support vector machines provide a linear separable n-dimensional hyperplan for dataset classification using an algebraic and static background method. Because of the intricate architecture of the heart, care must be used when manipulating it to prevent mortality. A variety of techniques, including knn, decision trees, generic algorithms, and naïve bayes, are used to classify the severity of cardiac disorders [3]. Al Mohan et al. Describe the process of combining two distinct methodologies to create a hybrid strategy, which has a higher accuracy rate than any other approach at 88.4%. A few researchers have focused on using data mining techniques to forecast cardiac problems. After working on this,

Kaur et al.[6] have defined how the information and intriguing pattern are obtained from the big dataset. To determine which machine learning and data mining technique is the best, they compare the accuracy of several ways and arrive at a conclusion that favourssvm.

Kumar et al.[5] have worked on a variety of data mining and machine learning methods. These algorithms are analysed using the UCI machine learning dataset, which consists of 303 samples with 14 input features. The results show that svm is the best algorithm among them; other algorithms that have been considered include naivybayes, knn, and decision trees. Using CAD technology, Gavhaneet al.[1] have worked on the multilayer perceptron model for the prediction of human cardiac illnesses and the algorithm's accuracy. The more people who use the prediction system to forecast their ailments, the more people will become aware of these conditions, which will lower the death rate for cardiac patients. A couple of algorithms for illness prediction have been developed by several academics. In their study, Krishnan et al.[2] shown that decision trees are more accurate than naïve bayes classification algorithms. Machine learning algorithms are used to predict many kinds of illnesses, and several researchers—including Kohali et al.—have worked in this area.(7)Research was done on the prediction of heart disease using logistic regression, diabetes using support vector machines, and breast cancer using Adaboot classifiers. The results showed that the logistic regression had an accuracy of 87.1%, the support vector machine had an accuracy of 85.71%, and the Adaboot classifier had an accuracy of up to 98.57%, all of which are good from the standpoint of prediction. A survey report on the prediction of cardiac disorders has demonstrated that, in contrast to hybridization, which performs well and provides improved prediction accuracy, the older machine learning techniques do not perform well[8].

5. Coding and Algorithm (Add Codings)

The impact of CKD on enhancing the treated soil's water holding capacity and reducing water evaporation from the soil surface may be the cause of the increase in water usage efficiency brought about by treating the sandy soil. It gave the plants more access to water. More plant materials will be produced with the water that is available, leading to increased water usage efficiency. In some circumstances, the suggested approach can assist medical professionals in making a prompt diagnosis or testing novel ones in hospitals. This technique may be used by medical college students to study and assess what they have learned.

This research aims to explore the many data mining approaches that are currently available for predicting heart disease, compare them, and then aggregate the results from each methodology to obtain the best accurate result. The methods of prediction and categorization were the main focus. Hybridization or merging many algorithms into one potent algorithm can increase the algorithms' accuracy. Hospitals may utilise the new algorithm as an expert system to assist doctors swiftly detect heart problems and perhaps save lives. It can also be utilised in medical schools for educational purposes.

6. Conclusion and Future Scope

Since the heart is one of the body's most significant and crucial organs and heart disease prediction is a major worry for people, algorithm accuracy is one of the parameters used to analyse how well algorithms operate. The dataset that is utilised for testing and training determines how accurate the machine learning algorithms are. The KNN method is the best one, according to our examination of the algorithms using the dataset whose properties are displayed in TABLE.1 and the confusion matrix. In order to reduce the incidence of mortality cases by raising awareness of the diseases, a more advanced machine learning technique will be utilised in the future for the best analysis of cardiac diseases and for early disease prediction.

References

[1] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease usingMachine Learning Algorithms” ICIICT, 2019.

[2] Aditi Gavhane, GouthamiKokkula, IshaPanday, Prof. KailashDevadkar, “Prediction of Heart Disease using Machine Learning”,Proceedings of the 2nd International conference on Electronics,Communication and Aerospace Technology(ICECA), 2018.

[3] Senthilkumarmohan, chandrasegarthirumalai and GautamSrivastva,“Effective Heart Disease Prediction Using Hybrid Machine LearningTechniques” IEEE Access 2019.

[4] Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease usingMachine Learning Algorithms: A Survey” International Journal onRecent and Innovation Trends in Computing and CommunicationVolume: 5 Issue: 8 , IJRITCC August 2017.

[5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of HeartDiseases Using Data Mining and Machine Learning Algorithms andTools” International Journal of Scientific Research in Computer Science,Engineering and Information Technology ,IJSRCSEIT 2019.

[6] Amandeep Kaur and JyotiArora,“Heart Diseases Prediction using DataMining Techniques: A survey” International Journal of AdvancedResearch in Computer Science , IJARCS 2015-2019.

[7] Pahulpreet Singh Kohli and Shriya Arora, “Application of MachineLearning in Diseases Prediction”, 4th International Conference onComputing Communication And Automation(ICCCA), 2018.

[8] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of HeartDisease Using K- Nearest Neighbor and Genetic Algorithm,” ProcediaTechnol., vol. 10, pp. 85–94, 2013.

[9] S. Kumra, R. Saxena, and S. Mehta, “An Extensive Review on SwarmRobotics,” pp. 140–145, 2009.

[10] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “ A Heart DiseaseDiagnosis and Prediction Using Machine Learning and Data MiningTechniques: A Review” Advances in Computational Sciences andTechnology , 2017.

[11] Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction UsingMachine learning and Data Mining

Technique" Journals of ComputerScience &Electronics , 2016.
[12] ChavanPatil, A.B. and Sonawane, P."To Predict Heart Disease Risk andMedications Using Data Mining Techniques with an IoT BasedMonitoring System for Post-Operative Heart Disease Patients"International Journal on Emerging Trends in Technology, 2017.
[13] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms inDisease Prediction," vol. 38, no. 3, pp. 124–128, 2016.
[14] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of riskscore for heart disease using associative classification and hybrid featuresubset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634,2012.
[15] https://archive.ics.uci.edu/ml/datasets/Heart+Disease