Prediction of Heart Diseases Using Machine Learning Techniques: Application of Artificial Intelligence

Balamurugan M, Blessed Prince P

Research Scholar
Department of computer science and engineering
Presidency university, Bangalore
Associate professor
Department of computer science and engineering
Presidency university, Bangalore
balamurugan.m@presidencyuniversity.in,blessprince25@gmail.com

Abstract

Heart-related diseases, also known as cardiovascular diseases (CVDs), have emerged as the most life-threatening disease, not just in India but all over the world. The heart-related diseases account for the majority of deaths that have occurred around the globe over the last few decades. Therefore, there is a need for a system that is dependable, accurate, and practical in order to detect such disorders in a timely manner so that appropriate treatment may be administered. The algorithms and methods of machine learning have been applied to a variety of medical datasets in order to automate the research of extensive and complicated data sets. We developed a method for determining whether or not a patient is likely to be diagnosed with a heart disease by utilising the patient's medical history in our model. This allowed us to construct a heart disease prediction system. A dataset that was gathered from the archives of every international university was used in conjunction with a number of supervised machine learning methods. These methods were used to make predictions regarding cardiac disease.The highest performance was obtained using KNN (accuracy = 93.3%, precision = 100%, sensitivity = 80%) , when the value of k=1.

Keywords

Cardiovascular Diseases; Logistic Regression; Support Vector Machines; Machine learning, heart disease, , prediction model, classification algorithms.

1 Introduction

The human heart is an essential component of the human body. It distributes blood to all of the different parts of our body. In the event that it does not operate properly, the brain and a number of other organs will cease operating, and the individual will pass away within a few minutes. It is estimated that approximately 17.9 million people pass away each year as a result of cardiovascular disorders; of these fatalities, 80 per cent may be attributable to coronary artery disease and cerebral stroke [1]. These statistics are from the World Health Organization. It is by far the most common cause of mortality in people. With the use of a person's medical history, our effort may assist in the prediction of individuals who are likely to be diagnosed with a cardiac ailment [2]. Chest pain and high blood pressure are two examples of these symptoms. Testing and training the data is the foundation of machine learning. The system learns from experience by taking in data, and after it has been trained, it makes predictions based on the test dataset. Heart disorders have emerged as one of the most prevalent causes of mortality all over the globe.Each year, cardiovascular ailments take the lives of 17.7 million people. Even in India, heart-related disorders have now surpassed other causes of death to become the leading cause of mortality [3]. In India, cardiac diseases were responsible for the deaths of 1.7 million people in the year 2016. Diseases that affect the heart not only drive up the cost of medical treatment but also lower an individual's overall productivity. According to projections provided by the World Health Organization (WHO), India may have suffered economic losses of up to $237 billion between the years 2005 and 2015 as a direct result of heart-related or cardiovascular disorders [4]. Hence, the prediction of heart-related disorders in a way that is both possible and accurate is highly significant. It is widely agreed that neural networks are the most effective tool for the prediction of diseasees such as cardiovascular disease and brain disease. The suggested technique that we utilise contains a total of thirteen characteristics that may predict heart disease. The findings demonstrate a higher degree of performance in comparison to the approaches that are currently used in works such as [5]. In recent years, the Carotid Artery Stenting (CAS), which stands for carotid artery stenting, has emerged as a popular therapeutic option in the world of medicine. Patients in their latter years who are suffering from heart disease are more likely to have major adverse cardiovascular events (MACE) when they have CAS. Their judgement becomes highly crucial moving forward. In order to obtain findings, we make use of an Artificial Neural Network (ANN), which has shown to be effective in the prediction of cardiovascular disease [6]. Approaches based on neural networks are presented, and these methods integrate not only the posterior probability but also the projected values from numerous techniques that came before them. This model reaches an accuracy level of up to 89.01 percent, which is an impressive result when compared against other studies. As was discussed before in [7], the Cleveland heart dataset is put to use in all of the experiments in conjunction with a neural network in order to increase the diagnostic accuracy of heart disease.

2 Related Work

Artificial neural networks, often known as ANN, are used in a wide variety of applications, including edge detection. A non-linear filter that makes use of neural networks is applied to an image in order to determine where the edges of the picture are located. In their research [8], Terry and Vu presenteda multi-layer feed-forward neural network to identify the edges of a laser radar image. This network is utilised to do this identification. For the purpose of carrying out this detection, this network is used. Synthetic edge patterns are employed throughout the training process of the networks. The network is able to identify a variety of edge types, including horizontal, vertical, diagonal, and more besides. Both Li and Wang contributed to the development of a novel neural network detector that operates on an image's 8-bit subimage. A Naive Bayes classifier technique was presented by Miranda et al. for the purpose of predicting cardiovascular disorders. When it comes to determining heart disease, the scientists have taken into consideration a few key risk factors. The author conducts research on a variety of machine learning (ML) methods that are applicable to the categorization of cardiac disease. Comparing the efficacy of the classification algorithms known as K-Means, K-Nearest Neighbors, and Decision Trees was the subject of a research that was carried out via the use of research. The results of this research indicate that the Decision Tree provides the best level of accuracy, and moreover, it was deduced that it may be made more efficient by combining a number of various methodologies and fine-tuning its parameters. Kohli[9] first proposed the use of neural networks to assist in the diagnosis and prediction of heart disease and blood pressure in addition to a variety of other characteristics.

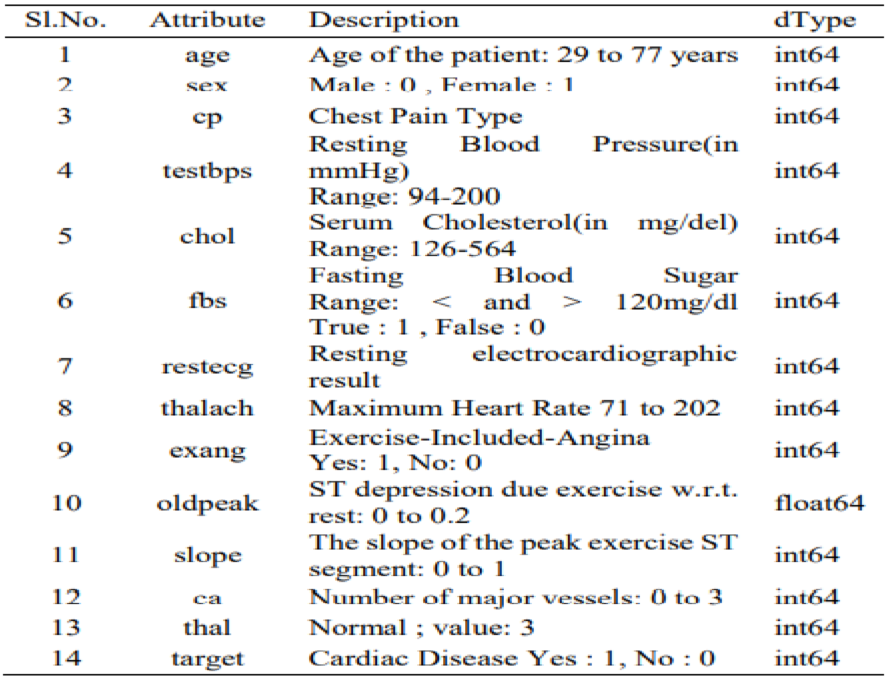
It was decided to construct a deep neural network that would include the specified characteristics associated with the condition. This network had the capability of producing an output, which was subsequently processed by the output perceptron. In other words, if we want to use the model for Test Dataset, we need to make sure that we have an accurate result. [10]These techniques are referred to collectively as "supervised classifiers".Yazdani[11] came up with the idea of a cost-sensitive ensemble technique to increase the effectiveness of diagnosis, which in turn minimises the cost of misclassification. There are a variety of different classifiers that have been employed, including random forest, logistic regression, SVM Model, extreme learning machine, and K-Nearest Neighbor. Ttest was used to examine the performance of the ensemble. The strategy that was presented did, in fact, obtain the best performance after undergoing 10 rounds of cross-validation. MdMamun Aliand colleagues [12] investigated how accurately heart disease might be predicted with the use of a group of different classifiers and majority vote. The accuracy was enhanced by 6.92 percent via the use of bagging, as opposed to 5.94 percent through the use of boosting, 7.26 percent through the use of majority voting when weak classifiers were ensembled, and 6.93 percent through the use of stacking. Sibo Prasad Patro[13] assessed the accuracies of the different machine learning algorithms that can categorise whether or not a person has heart disease using the dataset collected from the UCI repository. These strategies can determine whether or not a person has heart disease.When the algorithms were evaluated using this dataset[14],[15], it was determined that the K-Nearest Neighbor was the most effective of the bunch.

3 Methodology

In order to determine the causes of heart disease on the UCI, researchers employ a computational method in conjunction with the three association rules of mining, which are known as apriori, predictive, and Tertius as shown in figure 1

Cardiac diseases are any diseases that pertain to or relate to the heart, which is the organ in the human body that is responsible for pumping blood throughout the body. In the realm of medicine, it is well knowledge that the term "such disease" refers to a spectrum of disorders that may have an effect on the heart. Cleveland dataset. Based on the evidence that has been collected, one may draw the conclusion that women have a lower risk of developing heart disease in comparison to men. Accurate diagnosis is the first and most important step in treating cardiac problems. The conventional methods, on the other hand, are not sufficient for reliable prognosis and forecasting, as shown in Table 1.

Table 1. Dataset attribute



Sl.No.	Attribute	Description	dType
1	age	Age of the patient: 29 to 77 years	int64
2	sex	Male : 0 , Female : 1	int64
3	cp	Chest Pain Type	int64
4	testbps	Resting Blood Pressure(in mmHg) Range: 94-200	int64
5	chol	Serum Cholesterol(in mg/del) Range: 126-564	int64
6	fbs	Fasting Blood Sugar Range: < and > 120mg/dl True : 1 , False : 0	int64
7	restecg	Resting electrocardiographic result	int64
8	thalach	Maximum Heart Rate 71 to 202	int64
9	exang	Exercise-Included-Angina Yes: 1, No: 0	int64
10	oldpeak	ST depression due exercise w.r.t. rest: 0 to 0.2	float64
11	slope	The slope of the peak exercise ST segment: 0 to 1	int64
12	ca	Number of major vessels: 0 to 3	int64
13	thal	Normal ; value: 3	int64
14	target	Cardiac Disease Yes : 1, No : 0	int64

3.1 Data Source

An organised dataset of people was chosen for inclusion in the research with consideration given to their past experiences with heart diseases as well as any other preexisting medical disorders [14]. Diseases of the heart include a wide range of disorders that might have an adverse effect on the organ. According to the World Health Organization (WHO), cardiovascular diseasees are responsible for the highest number of fatalities among persons in the middle years of their lives. We use a data set that contains the medical histories of 304 individual patients, each of whom falls into one of many age categories. This dataset provides us with the much-needed information, which consists of the medical attributes of the patient. These attributes include things like the patient's age, resting blood pressure, and fasting sugar level, among other things. These characteristics assist us to assess whether or not the patient has been diagnosed with any kind of cardiac disease.This dataset consists of 304 individuals and comprises 13 medical parameters that help us determine if a patient is at risk of having a cardiac disease or not. Additionally, it categorise patients into those who are at risk of acquiring a heart disease and those who are not in danger of developingheart disease in the future.

3.2 Data Collection

The first step in getting started with the research activity is to collect the data that is available in the dataset held in the UCI repository [13]. This dataset is verified by a large number of scholars as well as the UCI authorities.

3.3 Data Pre-Processing

After the acquisition of a variety of records, the data on heart disease are then pre-processed. The patient records in this dataset total 303, although there are significant gaps in 6 of them due to missing data. The aforementioned 6 records have been eliminated from the dataset, and the pre-processing will be performed on the remaining 297 patient records. The multiclass variable and the binary classification are new concepts that have been applied to the characteristics of the dataset that has been presented.

3.4 Feature Extraction

As a result of this, an additional collection of features is produced from the data that were initially collected. The features are changed in some way throughout the process of feature extraction. As a result of this process, the change is often irreversible since some, maybe many, valuable pieces of information are lost in the process. Feature extraction is achieved with the use of the Principal Component Analysis (PCA) method. Principal Component Analysis is a well-known linear transformation method that may be discovered in a variety of different software applications. It does this by searching the feature space for the directions that have the greatest amount of variance and by searching for directions that are mutually orthogonal. The most accurate reconstruction may be achieved with the use of a global algorithm.

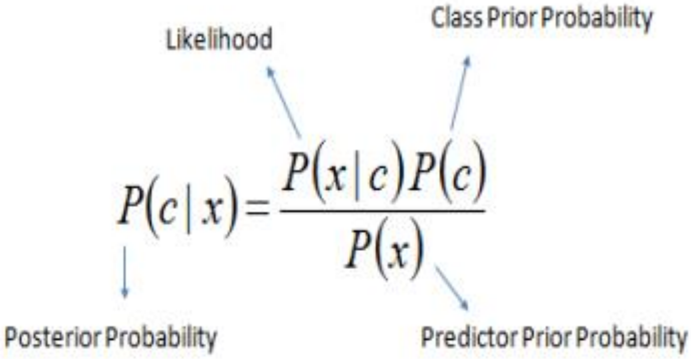
3.5 Reduction of Dimensionality

The process of dimension reduction is picking a mathematical representation in such a way that it is possible to connect the majority of the variation contained within the data that is provided, but not all of it. This allows one to include just the information that is most relevant. The information that is taken into consideration for a task or a problem may comprise a very large number of features or dimensions; yet, not all of these qualities or characteristics may have the same degree of effect on the result of the work or the issue. A high number of characteristics, also known as features, may influence the difficulty of the calculation and may even cause overfitting, which in turn produces inaccurate results. Therefore, dimension reduction is a very crucial phase that should be addressed whenever one is constructing a model. In most cases, dimension reduction may be accomplished via the use of two distinct approaches: feature extraction and feature selection.

4.Machine Learning Algorithms and Techniques

4.1 Naive Bayes

The Bayes Theorem provides the foundation for an approach to categorization known as Naive Bayes, which is both straightforward and useful. It presupposes that the predictors are independent of one another, which means that the characteristics or traits under consideration should not be associated with one another and should not in any manner be connected to one another. In the work [15], Naive Bayes obtains an accuracy of 84.1584 percent using the 10 most important features, which are determined via SVMRFE (Recursive Feature Elimination) and gain ratio methods. In other words, these techniques help Naive Bayes accomplish its results. On the other hand, according to the research presented in the research [17], the accuracy of Naive Bayes increases to 83.49 percent when all 13 features of the Cleveland dataset [16] are used. Both [15] and [17] may be accessed over the internet.The Naive Bayes classifier evaluates to92.0792% of correctly classified instance .



$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

4.2K nearest neighbours (KNN)

Logistic Regression and Random Forest Classifiers are two diagnostic tools that may assist doctors and other medical professionals in more correctly diagnosing heart disease. This paperwork includes looking through current data on cardiovascular disease, in addition to publications and papers that have been published and published data. The methodology provides a structure for the model that has been presented [18]. The technique is a procedure that contains stages that turn provided data into recognised data patterns for the understanding of the users. These patterns may be accessed by the users. The suggested approach consists of many stages, the first of which is the gathering of the data, followed by the extraction of important values in the second stage, and finally the preprocessing stage, in which we investigate the data. Based on the process that are used, the preparation of data may include the elimination of missing values, the cleaning of data, or the normalising of data [19].The dataset analysed using the KNN algorithm produces a precision of 100%

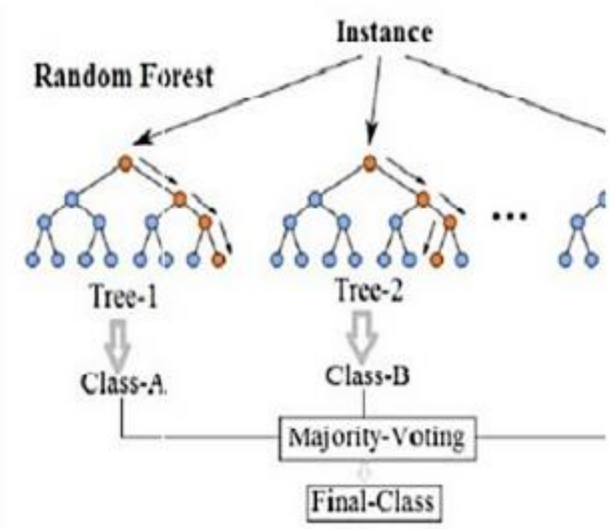


Fig. 2.Random Forest Simplified

4.3 Random Forest

The Random Forest method is used for machine learning that is supervised. classification problems but performs better than average in classification tasks in general. As its name suggests, the Random Forest methodology takes into consideration a variety of decision trees before arriving at a conclusion in order to provide an output. The reasoning for this approach is based on the hypothesis that if more trees were planted in the forest, the results will ultimately point in the right direction. When it comes to classification, the method utilises a voting process to choose the class; on the other hand, when it comes to regression, it takes the mean of all of the outputs from the decision trees. As can be seen in the figure below, it performs well with huge datasets that have a high dimensionality.

4.4Artificial Neural Network

They are used to either represent or imitate the distribution, functions, or mappings among variables, and they are modules of a dynamic system that are coupled with a learning rule or an algorithm for learning. This association allows them to learn. These are also referred to as "mappers." These modules act as neuron simulators in the nervous system; hence, artificial neural networks (ANN) collectively refer to the neuron simulators and the synapses between the many modules that make up the network. When building a neural network, many neurons are stacked on top of one another in layers to form the network's final output. The input layer is the very first one, and the output layer is the very final one. The layers that lie in between are referred to as hidden layers. An activation function is present in each neuron. Sigmoid, ReLU, tanh, and other similar functions are examples of well-known Activation functions. The weights and biases of each node make up the network's parameters, and they are set up in such a way that the actual result and the anticipated outcome are identical. The backend configuration of the network may be seen in figure 3.The estimation done throughANN produces an accuracy of 79.9043%

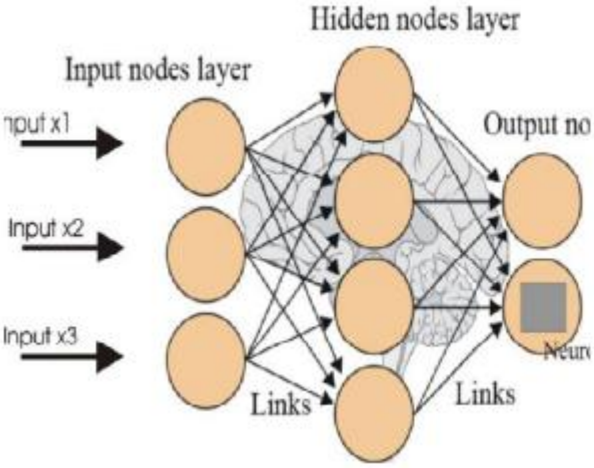


Fig. 3 ANN Layers

5 Result & Discussion

This research's objective is to evaluate the efficacy of a number of different classification algorithms and, as a result, to identify the classification method that yields the most reliable results for determining whether a patient will or will not develop heart disease. On the dataset shown in figure 4, this research was carried out by applying the Naive Bayes, K-Nearest Neighbor, and Random Forest analysis methods. When it comes to predicting whether or not a patient has heart disease, KNN performs much better than Random Forest Classifier. This demonstrates that KNN and Logistic Regression are superior approaches to use when diagnosing a cardiac condition. The next graph presents a plot of the number of patients that have been segregated and predicted by the classifier based upon the age group, resting blood pressure, sexual orientation, and chest pain which are all clearly depicted in figures 5 and 6, respectively.

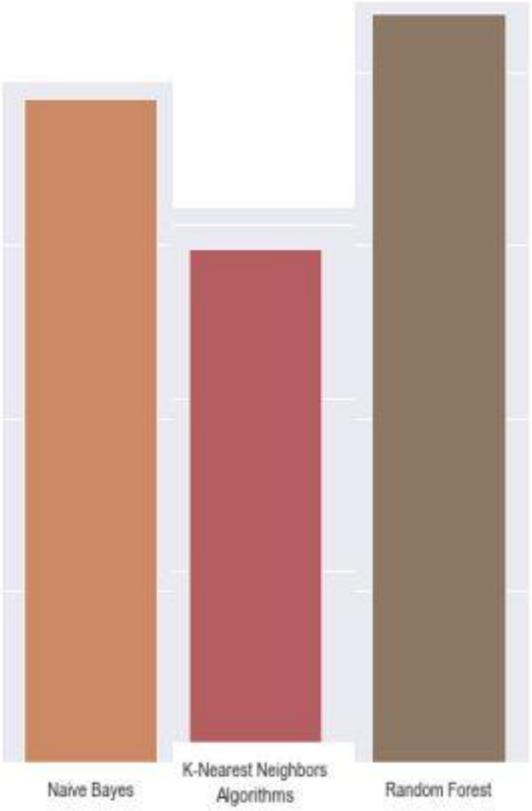


Figure. 4 Performances of Naive Bayes, KNN & Random Forest.

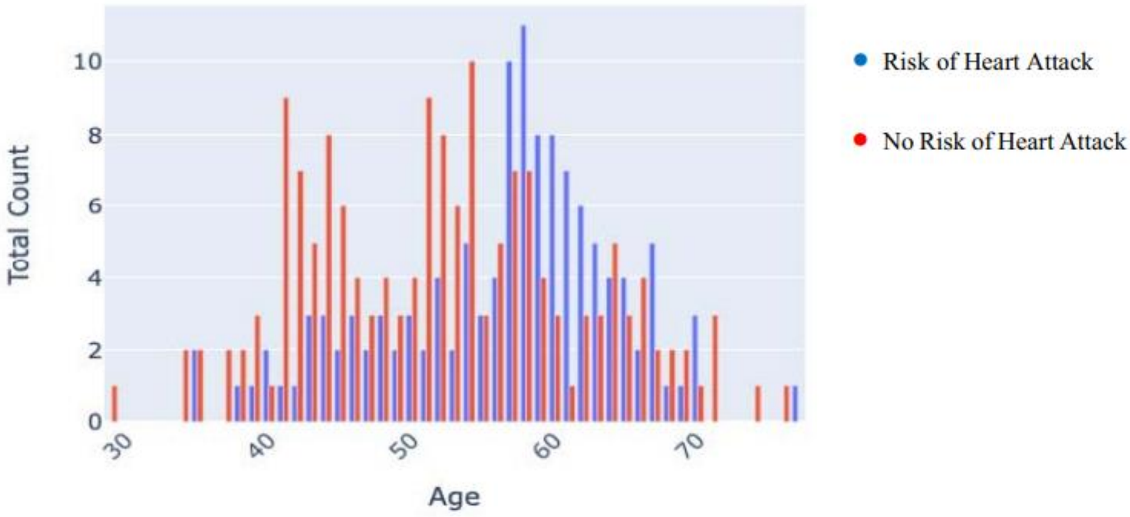


Figure 5. Heart Attack risk on the basis of their age.

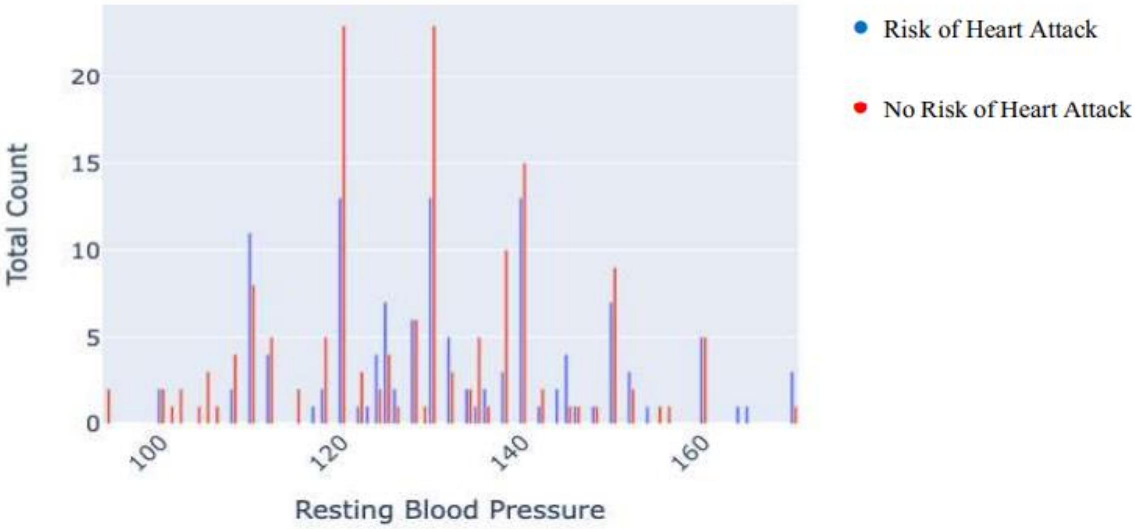
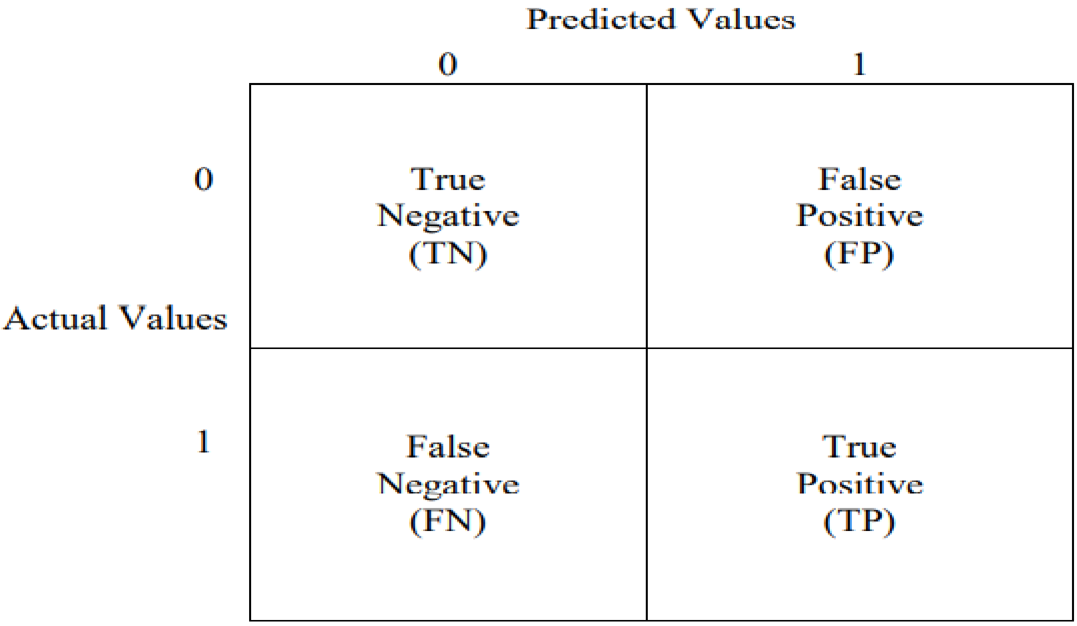


Figure 6. Heart Attack risk on the basis of their Resting Blood Pressure

A matrix of this kind gives information regarding the manner in which the classifier has performed in terms of comparing the instances that were properly predicted to the examples that were not correctly predicted. A confusion matrix is the name given to the tabular illustrationof the model estimates and the actual values of the dataset. For the purpose of monitoring how well their models are doing, classification-based challenges in machine learning make use of it. As can be seen in figure 7, it is made up of four distinct combinations of the expected and actual values. These combinations are denoted by the terms True Negative, True Positive, False Positive, and False Negative. These four components serve as the fundamental building blocks for the confusion matrix that is being designed.



		Predicted Values	
		0	1
Actual Values	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Figure 7 Confusion Matrix.

Conclusion

The overarching objective is to provide a variety of data mining strategies that are helpful in accurate prediction of cardiac disease. The purpose of this research is to develop a system that can provide accurate and reliable predictions using a reduced set of characteristics and evaluations. This research is able to forecast individuals who have cardiovascular disease by extracting, from a dataset that contains patient medical history, the patient medical history that leads to a deadly heart disease. This allows the project to make accurate predictions about people who have cardiovascular disease. This Heart Disease detection system provides aid to the patient based on his or her clinical information, which reveals whether or not the patient has previously been diagnosed with heart disease. Analysis of the implemented linear kernel in SVM along with other SVM Classifier Kernels can be performed in the future for the purpose of providing a distinct and understandable visualisation of the workings of the algorithm, which will allow for improved prediction. Both Random Forest and Ensemble models have been quite successful, and this may be attributed to the fact that they utilise numerous methods to combat the issue of overfitting. The Naive Bayes classifier-based models were quite efficient in terms of processing time, and they also performed very well. In the majority of these tests, SVM performed quite well. However, there is still a significant lot of work that has to be done in terms of research on how to cope with high-dimensional data and overfitting. This work needs to be done before it can be considered complete. Systems that are based on the methods and techniques of machine learning have proven to be highly effective in predicting heart-related disorders; however, there is still a great deal of room for improvement in this area. The continuation of this line of research might be carried out using a wide variety of different blends of machine learning approaches and improved prediction tools. In addition, the efficacy of heart disease prediction may be improved by developing new techniques for the selection of characteristics, which will provide a more comprehensive understanding of the factors that are relevant. Each of the algorithms that have been discussed up to this point has, in some circumstances, done very well but has, in other circumstances, fared quite badly.

Reference

[1] F. Dammak, L. Baccour, and A. M. Alimi, ‘‘The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,’’ in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015.

[2] Basheer, S., Alluhaidan, A.S. &Bivi, M.A. Real-time monitoring system for early prediction of heart disease using Internet of Things. Soft Comput 25, 12145–12158 (2021). https://doi.org/10.1007/s00500-021-05865-4

[3]Bhardwaj, S., Jain, S., Trivedi, N.K., Kumar, A., Tiwari, R.G. (2022). Intelligent Heart Disease Prediction System Using Data Mining Modeling Techniques. In: Kumar, R., Ahn, C.W., Sharma, T.K., Verma, O.P., Agarwal, A. (eds) Soft Computing: Theories and Applications. Lecture Notes in Networks and Systems, vol 425. Springer, Singapore. https://doi.org/10.1007/978-981-19-0707-4_79

[4]Balaha, H.M., Shaban, A.O., El-Gendy, E.M. et al. A multi-variate heart disease optimization and recognition framework. Neural Comput&Applic (2022). https://doi.org/10.1007/s00521-022-07241-1

[5]Jha, R.K., Henge, S.K., Sharma, A. (2022). Heart Disease Prediction and Hybrid GANN. In: Kahraman, C., Cebi, S., CevikOnar, S., Oztaysi, B., Tolga, A.C., Sari, I.U. (eds) Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation. INFUS 2021. Lecture Notes in Networks and Systems, vol 308. Springer, Cham.

[6]El-Shafiey, M.G., Hagag, A., El-Dahshan, ES.A. et al. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. Multimed Tools Appl 81, 18155–18179 (2022). https://doi.org/10.1007/s11042-022-12425-x

[7]Reddy GT, Reddy MPK, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G (2020) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evol Intel 13:185–196

[8]Chitra R, Seenivasagam V (2015) Heart disease prediction system using intelligent network. In: Power electronics and renewable energy systems. Springer, pp 1377–1384

[9]Kohli R, Garg A, Phutela S, Kumar Y, Jain S (2021) An improvised model for securing cloud-based E-Healthcare systems. In: IoT in Healthcare and ambient assisted living. Springer, pp 293–310

[10] Krishnaiah V, Narsimha G, Chandra NS (2015) Heart disease prediction system using data mining technique by fuzzy K-NN approach, vol 337. Springer, Cham

[11]Yazdani, A., Varathan, K.D., Chiam, Y.K. et al. A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform DecisMak 21, 194 (2021). https://doi.org/10.1186/s12911-021-01527-5

[12]MdMamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni,Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,Computers in Biology and Medicine, Volume 136,2021,104672,ISSN 0010-4825.

[13]Sibo Prasad Patro, GouriSankarNayak, NeelamadhabPadhy,Heart disease prediction by using novel optimization algorithm: A supervised learning prospective,Informatics in Medicine Unlocked,Volume 26,2021,100696,ISSN 2352-9148

[14]S. Verma and A. Gupta, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 249-253, doi: 10.1109/ICAIS50930.2021.9395963.

[15]T. P. Pushpavathi, S. Kumari and N. K. Kubra, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 915-923, doi: 10.1109/ICICT50816.2021.9358733.