Machine Learning Based Heart Disease Prediction Model with GUI Interface

Harveen Kaur[1], Balraj Preet Kaur[2], Kumari Ragini[3], Archana Gupta[4]

[1 ,3, 4]Student, [2]Research Scholar
[1,2,3,4] Department of Computer Science and Engineering, DAV University,Jalandhar,India

[1]harveenk409@gmail.com,[2]balrajpreetkaur2@gmail.com,[3]itsrcragini2004@gmail.com,[4]archnagupta8656@gmail.com

ABSTRACT

One of the top causes of death around the globe is a heart attack. According to current statistics, one person dies from heart disease every minute, making it one of the primary problems in everyday modern life. The ability to predict the onset of illness at an early stage is extremely difficult nowadays. When used in the healthcare industry, machine learning has the potential to accurately and quickly diagnose diseases. The circumstances under which heart disease may arise are estimated in this study. Medical parameters are characteristics of the datasets utilized. The datasets are analyzed using the Random Forest Algorithm, a machine learning algorithm, in Python. This method makes use of historical patient data from the past to forecast future ones at an early stage, saving lives. In this study, a trustworthy system for predicting heart disease is put into place utilizing a powerful machine learning algorithm called the Random Forest method. This reads a CSV file containing patient record data. After gaining access to the dataset, the procedure is carried out, and a useful heart attack level is generated. The suggested system's benefits include High success rates are attained, along with excellent performance and accuracy rates, flexibility, and adaptability.

Keywords: Machine learning, Artificial Intelligence, Heart disease.

INTRODUCTION

The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained using statistical techniques to produce classifications or predictions and to find important insights in data mining projects. The decisions made as a result of these insights influence key growth indicators in applications and enterprises, ideally. Data scientists will be more in demand as big data continues to develop and flourish. In the actual world, we are surrounded by people who are able to learn from their experiences thanks to their capacity for learning, and we also have computers or other robots that carry out our orders. But can a machine learn from past facts or experiences the same way a human does? So now the role of machine learning is revealed. Without being explicitly programmed, computers may learn from data and even get better on their own, thanks to machine learning algorithms[9][10].

Without being explicitly coded, software systems can predict events more accurately thanks to a class of algorithms known as machine learning (ML). Building algorithms that can take input data and apply statistical analysis to predict an output while updating outputs as new data

becomes available is the fundamental idea behind machine learning. The machine learning lesson explains both fundamental and sophisticated machine learning principles. Both students and working professionals can benefit from our machine learning tutorial [11].

A developing technology called machine learning makes it possible for computers to learn automatically from historical data. Machine learning uses a variety of techniques to create mathematical models and make predictions based on previous information or data. Nowadays, it is utilized for many different things, including recommender systems, email filtering, Facebook auto-tagging, image identification, and speech recognition [12].

The most prevalent kind of machine learning algorithm is supervised learning. It trains an algorithm with a known set of input data (called features) and known outputs using a known dataset (referred to as the training dataset). Labeled input data that corresponds to desired outputs or response values is included in the training dataset. The supervised learning algorithm uses it to make predictions of the response values for a fresh dataset after building a model from it by identifying connections between the features and output data[13].

Machine learning under supervision discovers patterns and connections between input and output data. The usage of labeled data defines it. A dataset with many examples of features and goals is referred to as "labeled data." Algorithms for supervised learning employ a dataset to learn the link between features and targets. Training or fitting are terms used to describe this process. In supervised learning, models are trained using a labeled dataset, where the model learns about each type of data. Once the training process is complete, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output [14].

CLASSIFICATION

In supervised machine learning, classification is a type of outcome or event prediction where algorithms learn from the data. For instance: bank might have a customer database with information on loans, investments, and credit histories and wonder if any of their clients will default. We'll have features and goals in the history data. Features are a customer's characteristics, such as their credit history, loans, investments, etc. If a specific client has defaulted in the past, the target will indicate this (often expressed by 1 or 0; true or false; yes or no). Classification techniques are used to predict the outcome when there are only two possible values for a discrete outcome, such as true or false, default or no default, yes or no [15].

1. Logistic Regression
2. Decision Tree
3. K-nearest Neighbor
4. Random Forest
5. Neural Networks

Unsupervised learning commonly referred to as "unsupervised machine learning," analyses and groups unlabeled datasets using machine learning algorithms. These algorithms identify hidden patterns or data clusters without the assistance of a human. It is the best option for exploratory data analysis, cross-selling tactics, consumer segmentation, and picture identification because of its capacity to find similarities and differences in information [16]. Clustering, association, and dimensionality reductions are the three basic tasks that unsupervised learning models are used for. Compared to supervised learning, unsupervised learning algorithms enable users to carry out more complication processing tasks. Yet, compared to other natural learning processes, unsupervised learning might be more unpredictable. Clustering, anomaly detection, neural networks, and other techniques for unsupervised learning are examples. Finding valuable insights from the data is made easier with the aid of unsupervised learning. Unsupervised learning is considerably more like how humans learn to think through their own experiences, which brings it closer to actual artificial intelligence. Unsupervised learning is more significant since it operates on unlabeled and uncategorized data. Unsupervised learning is necessary to handle situations where the input and output are not always the same in the actual world[17][18].

LITERATURE OVERVIEW

Asif Niwaz et al [1] proposed a model to make a decision-support system by using clinical reports and laboratory test results. In this model, the dataset is split into 5 folds, and it is ensured that it is equally divided into each fold. Furthermore, standard random forest was unable to handle the imbalanced classification problem and became biased in favour of the majority class. To make it work, the strategy is incorporated in the construction stage of the model of a random forest classifier.

Fatma et al [2] used classification algorithms to predict heart failure disease. In this model, success rates have been measured using various classification algorithms. This category includes Gaussian naive bayes, support vector machines, linear discriminant analysis, decision trees, and the random forest classifier algorithm. All these algorithms are compared with their success rates and complexity matrix estimation data.

Fahd Saleh Alotaibi et.al [3] made a model to improve the prediction of heart failure. Basically, it emphasis on the accuracy measured in previous work and the proposed model by using decision trees, logistic regression, random forests, Naive Bayes, and system vector machines. All models are run in parallel to compare previous work to the proposed model's work. To make this possible, data expansion and 10-fold cross-validation are effective tools.

Ramesh TR et.al [4] proposed a framework suitable for evaluating genuine models, which necessitates using a dependable, robust, and trustworthy framework, including such machine learning methodologies. This framework works immensely well for static and dynamic datasets. To make this model work effectively, 303 patients' records are used to predict better results.

ApurvGarg et.al[5] made a model whether a person will suffer from cardiovascular disease or not including machine learning concepts like supervised learning algorithms.K-nearest neighbor(K-NN) and random forest algorithm both are used and results has been calculated foreach.Both of these algorithms worked well on the dataset.

The publisher of this proposed model, Mahesh Parmar et.al [6], used deep learning's concept of a neural network to predict heart disease. Because it involves the implementation of K-NN, system vector machines, and hyper-parameter optimization, the accuracy of hyper-parameter optimization is found to be high in comparison to others, this paper aims to achieve better accuracy to make it more sufficient and possible. Hence, the chances of developing heart diseases can be predicted by this model.

Abdallah A dellatie et.al [7] proposed a model for improving detection of heart disease and worked on improving weighted random forests. A comparison has been done on the results between IWRF and SMOTE-RF on the HD clinical record database, including optimization and excluding it too. In this paper, performance evolution is compared with previous studies for HD detection using a statlog dataset. Therefore, it played a significant role in improving accuracy compared with existing models.

M.A. Jabbar et.al [8] proposed a model in which random forest was used as a classifier, chi-square, and genetic algorithm to measure heart disease risk. It led to improved accuracy in classification, and this model can be useful for health care professionals. To make it possible, various parameters are compared for the given dataset. Therefore, an efficient approach has been made for heart disease prediction risks.

PROPOSED MODEL

The suggested model includes a dataset of patients with cardiac disease, a machine learning algorithm, and assessment criteria. The dataset is divided into training data (70%) and testing data (30%). During the classification procedure, the random forest method is applied to the dataset. The suggested algorithm uses patient data on the key health indicators to predict heart disease.

Obtaining the dataset, including the characteristics of a person with a cardiac condition and a person who is not, as well as the conclusion as to whether the person has the condition or not, is the first stage in the setup. This experiment's data set was obtained from the Kaggle website (https://www.kaggle.com/ronitf/heart-disease-uci). Python was employed as the experiment's programming language.The figure 1 show the flowchart of proposed model.In this study, the information was used to employ various machine learning methods for determining whether or not an individual has cardiac disease. The dataset will then be cleaned up of null values, missing values will be handled, the dataset will be visualised, and various machine learning methods'

precision will be evaluated. Below is a definition of the machine learning methods used.

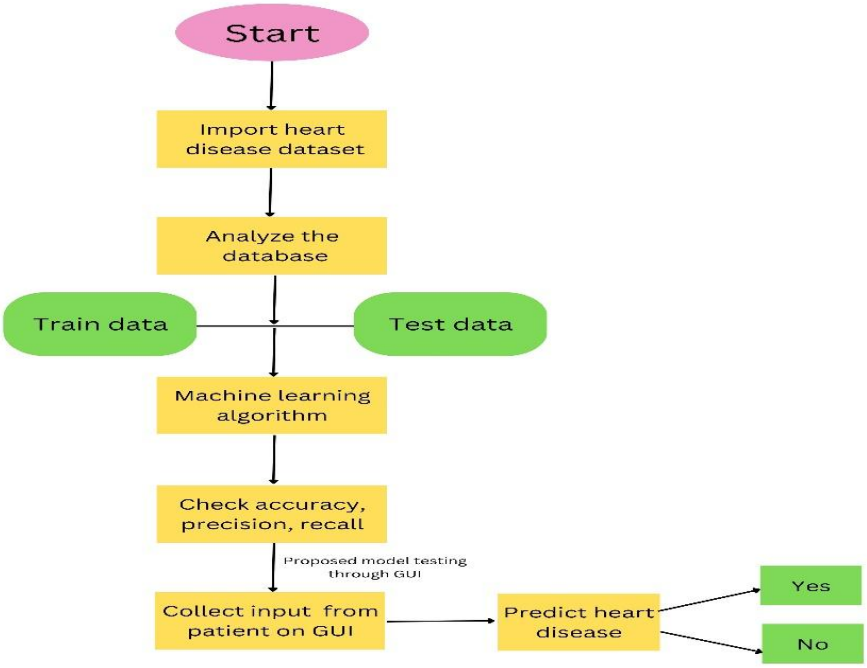

Figure-1 Flowchart of proposed model

DATASET DESCRIPTION

The dataset was collected from Kaggle(https://www.kaggle.com/ronitf/heart-disease-uci). It contains 14 attributes and 304 patients' data to make its prediction better. The target variable is whether a patient will survive or not. It has been reported by the donors that all legal factors were considered when collecting the data. These attributes are listed in short form and described in Table no.1.

Features	Description	Statistics
Age	Age of patient	Range:29-76
Sex	Gender of the patient	0 and 1 0->Female 1->Male

Chest pain type	Displays the type of chest-pain experienced by individual	Range:0-3 0->typical angina 1->atypical angina 2->non-anginal pain 3->asymptotic
Resting blood pressure	Displays the resting value of blood pressure	Range:94-172
Cholesterol	Displays the serum cholesterol	Range:126-417
Fasting blood sugar	Compares the Fasting blood sugar value of an individual	If fasting blood sugar level>120mg Then: if 1: true else: false
Resting ECG	Displays resting electrocardiographic results	Range:0-2 0->normal 1->having ST-T wave abnormality 2->asymptotic

Thalach	Person's maximum heart rate achieved	Range:71-202
Exang	Exercise induced angina	0&1(0->True: 1->False)
Old peak	ST depression induced by exercise	Range:0-6.2
Slope	Peak exercise ST segment	Range:0-2
CA	No.of major vessel colored y fluoroscopy	Range:0-3
Thal	Displays the thalassemia	Range:1-3 1->normal 2->fixed defect 3->reversible defect
Target	Whether a person will suffer from heart disease or not	0 false and 1 true

DATA PREPROCESSING

Correlation demonstrates the relationship between the traits and the objective variable or both. When one value grows, the value of the objective measure also rises, which is known as a

positive correlation (increase in one value, the value of the target variable decreased). From this heatmap, it is simple to categorise the features that are most important to the goal variable and to use the Seaborn library to plot the related features of the heatmap. We can conclude that chest discomfort plays the largest role in predicting the existence of heart disease when compared to the relationship between the other two factors. The figure 2 show the correlation map [19][20].

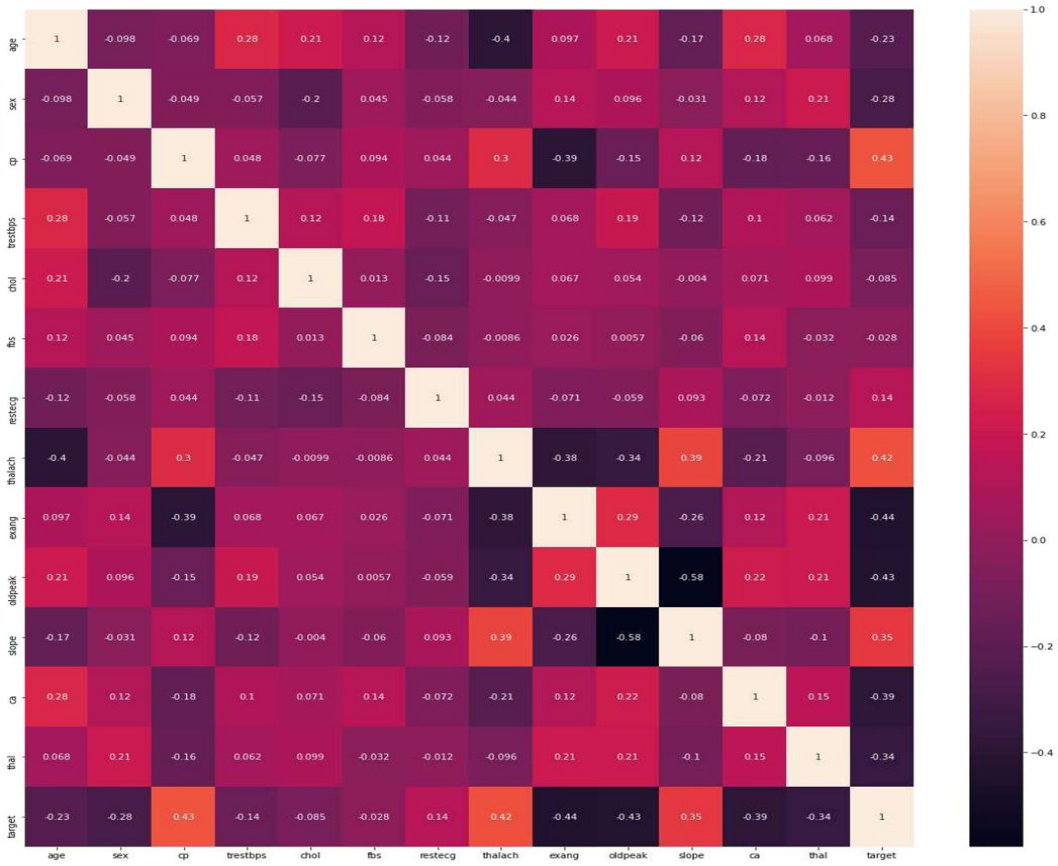

Figure-2 Correlation Map

MACHINE LEARNING ALGORITHM

In this research three machine learning algorithm used for analysis as explained below.

1. KNN is a non-parametric algorithm, meaning it does not make any assumptions about the distribution of the data. Instead, KNN classifies new data points based on the closest neighbors in the training set. The value of k, which represents the number of nearest neighbors to consider, is a hyperparameter that can be tuned to improve the accuracy of the model.Let's say we have a dataset with two classes, represented by different colored points on a graph. To classify a new data point (represented by the red point), we find the k-nearest neighbors in the training set (represented by the blue points). In this case, we set k=3, so the three closest neighbors are selected. The majority class among the neighbors (in this case, the blue class) is then assigned to the new data point. KNN can

also be used for regression problems, where the output is a continuous variable. In this case, the predicted value for a new data point is the average of the values of the k-nearest neighbors in the training set.One advantage of KNN is that it is easy to interpret and implement, making it a good starting point for beginners in machine learning. However, KNN can be computationally expensive for large datasets, and the choice of k can affect the accuracy of the model[21].

2. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier[22].

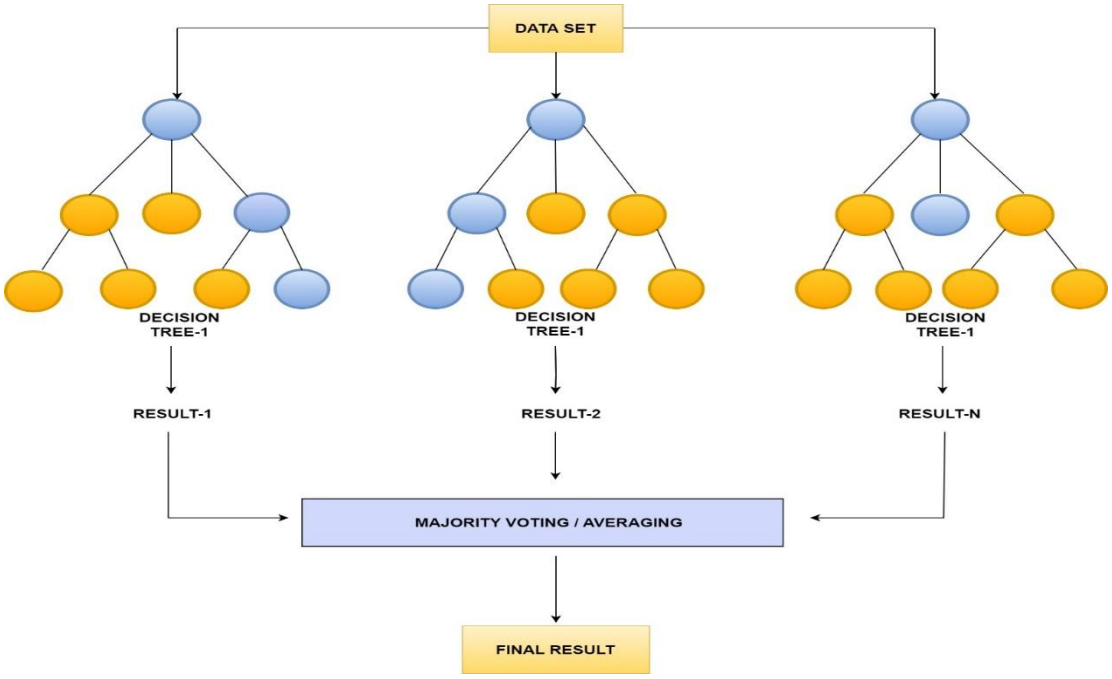

Figure-3

1. Random Forest is a classifier that uses more decision trees on various subsets of the supplied dataset and takes the average into account to increase the dataset's prediction accuracy. The RFA method uses the results from each decision tree to anticipate the

ultimate output, as illustrated in Fig. 3, rather than relying on a single decision tree. The number of trees determines how accurate the outcome will be; the better the accuracy rate, the more trees there are. Also, it prevents the issue of overfitting. The following step provides an explanation of how the algorithm functions[23][24].

Step 1: Selecting K data points from the chosen training set.

Step 2: Create as many decision trees as possible linked to the chosen data points.

Step 3: Decide how many decision trees, N, you want to construct.

Step 4: Steps 1 and 2 should be repeated.

Step 5: Assign new data points to the category that receives the majority of votes by looking up each decision tree's predictions for the new data point.

RESULTS AND DISCUSSION

The proposed model has the advantages of high performance and accuracy. It has a high success rate and is extremely adaptable. Our project's primary goal is to determine whether or not a person has heart disease and make recommendations about how to move forward. High accuracy rates may be attained with the Random Forest method.

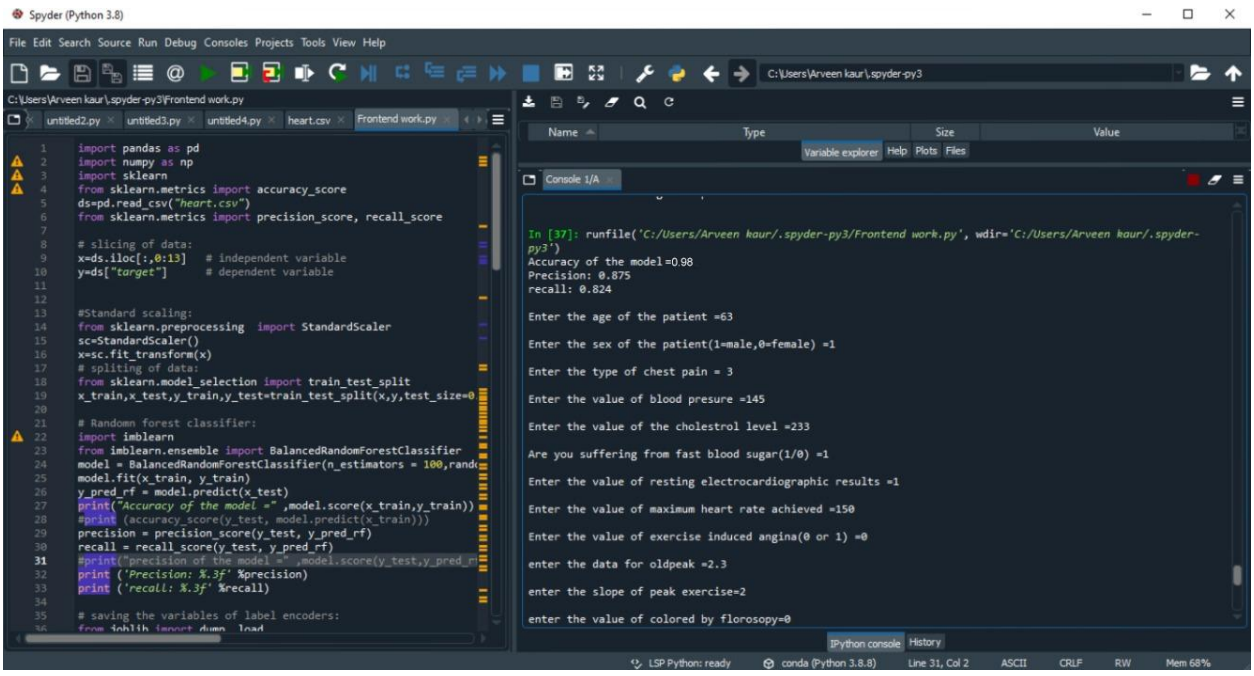

Figure-4

The above figure show the python interface of proposed model. The precision value of model is 87.5% and recall value is 82.4%.

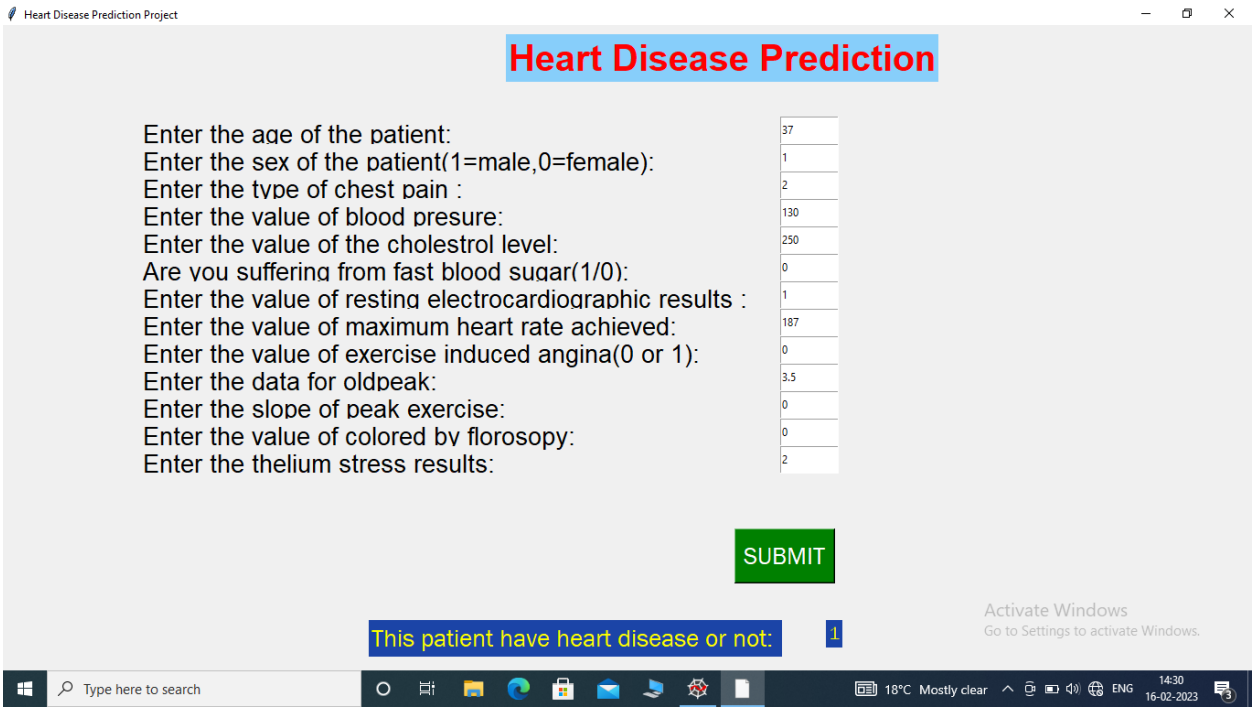

Figure-5

Figure 5 and 6 illustrates how the model functions as a graphical user interface (GUI) for a system that predicts cardiac illness. The application data collection comprises of a training/testing data set and user input. In this effort, the HTML environment is used to build a GUI. Figure 4 illustrates that a patient's test reports are given to GUI interface and result is 1 means patient may suffer from heart failure disease.

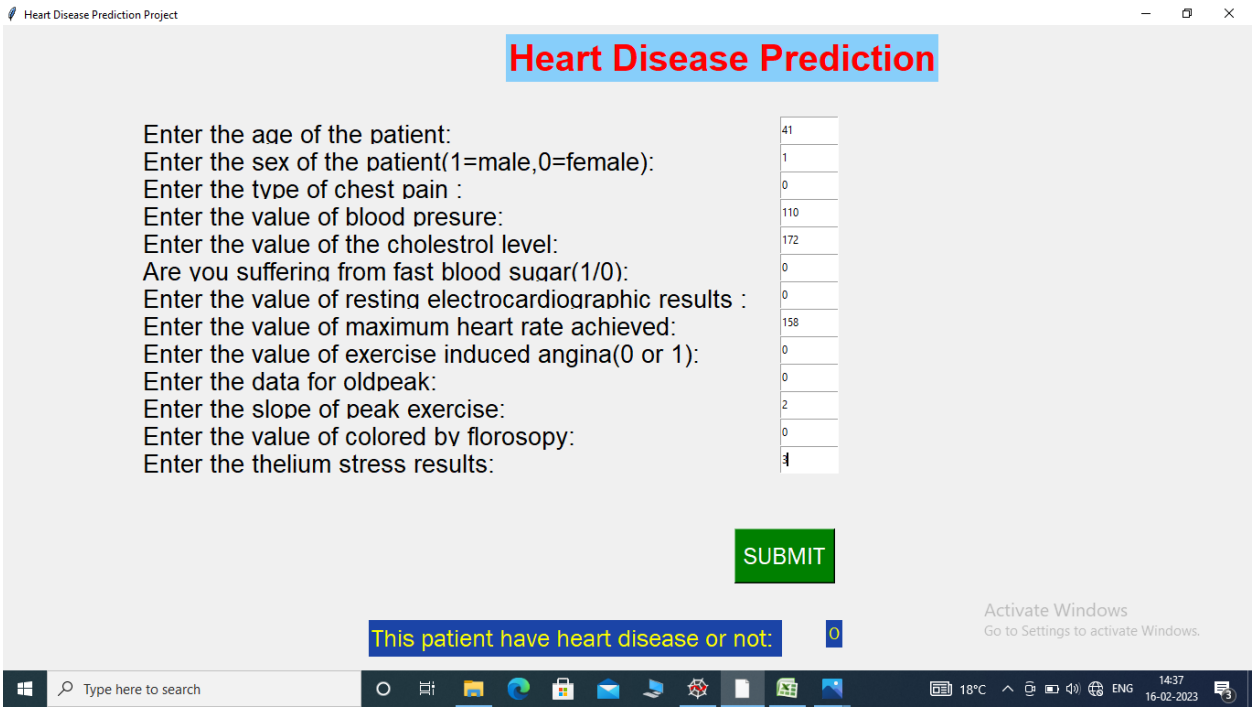

Figure-6

This screenshot illustrates that a patient's test reports are given to GUI interface and result is 0 means patient may not suffer from heart failure disease.

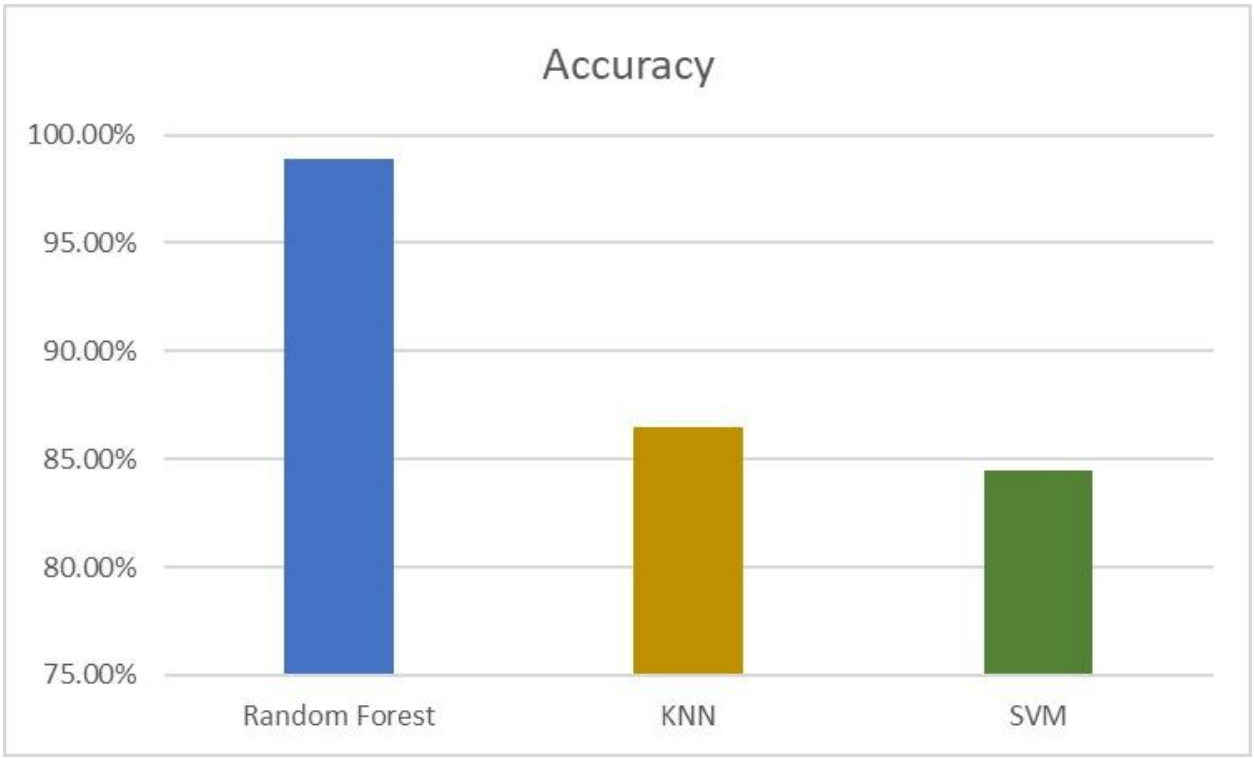

Figure-7 Result of Random forest

This graph depicts the accuracy,precision and recallvalue of the proposed model.The success of RF outperformed the other three classifiers in terms of training accuracy and assessment accuracy, as shown in Figures 8. RF's testing accuracy was 98.9%, and training accuracy was 100%. The second significant predictor was KNN, with training and testing accuracy of 90.5% and 86.5% respectively. Out of three classifications, the SVM displayed the worst results, with training accuracy and testing accuracy of 89% and 84.7%, respectively.

Table 2. Comparison Table of literature survey

Paper name	Accuracy	Dataset
Asif et al[1]	76.25%	Medical record of heart failure patient
Kalpyetmezligi et al[2]	90.76%	Heart failure prediction dataset
Fahd saleh et al[3]	89.14%	Heart disease dataset
Ramesh et al[4]	97%	UCI Repository
ApurvGarg et al[5]	81.967%	Ronitf/Heart disease dataset
Proposed model	98.9%	Heart disease detection

Table no.2 elucidates the comparison of different papers among the proposed paper on the basis of accuracies with heart disease dataset.

CONCLUSION

An effective ensemble learning methodology for regression and classification procedures is the Random Forest algorithm. The algorithm builds N decision trees and produces a class that represents the average result of each decision tree. Hence, early forecast accuracy is successfully

attained. Analyzing healthcare data, specifically data connected to the heart, will aid in the early diagnosis of heart disease or other aberrant cardiac conditions, preventing long-term fatalities. Predicting heart disease is a difficult task in today's world. By entering the report values, the patient or user can utilize this programmer to anticipate disease even if they are not in close proximity to a doctor and can choose whether to move forward or not. Future updates to this model might include new features like an early notification alerting all of the user's family members if they have heart disease. Also, the nearby hospital should receive the information. Another requirement is that the closest doctor should be reachable via online consultation.

REFERENCES

[1]Asif Newaz, Nadim Ahmed, FarhanShahriyarHaq,"Survival prediction of heart failure patients using machine learning techniques",Informatics in Medicine Unlocked,Volume 26,2021,100772,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2021.100772.

[2]]COŞKUN, Cevdet, and Fatma KUNCA, "EVALUATION OF PERFORMANCE OF CLASSIFICATION ALGORITHMS IN PREDICTION OF HEART FAILURE DISEASE", *KahramanmaraşSütçü İmam ÜniversitesiMühendislikBilimleriDergisi* 25.4 (2022): 622-632.

[3]Alotaibi, Fahd Saleh, "Implementation of machine learning model to predict heart failure disease", *International Journal of Advanced Computer Science and Applications* 10.6 (2019).

[4]Ramesh, T. R., et al, "Predictive analysis of heart diseases with machine learning approaches", *Malaysian Journal of Computer Science* (2022): 132-148.

[5]Garg, Apurv, Bhartendu Sharma, and Rijwan Khan. "Heart disease prediction using machine learning techniques", *IOP Conference Series: Materials Science and Engineering*. Vol. 1022. No. 1. IOP Publishing, 2021.

[6]Sharma, Sumit, and Mahesh Parmar. "Heart diseases prediction using deep learning neural network model", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 9.3 (2020): 2244-2248.

[7]Abdellatif, Abdallah, et al, "Improving the heart disease detection and patients' survival using supervised infinite feature selection and improved weighted random forest", *IEEE Access* 10 (2022): 67363-67372.

[8]Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", *Journal of network and innovative computing* 4.2016 (2016): 175-184.

[9]Bonaccorso, Giuseppe," *Machine learning algorithms*", Packt Publishing Ltd, 2017.

[10]Ray, Susmita,"A quick review of machine learning algorithms", *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, 2019.

[11]Singh, Amanpreet, Narina Thakur, and AakankshaSharma,"A review of supervised machine learning algorithms", *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee, 2016.

[12]Mohammed, Mohssen, Muhammad Badruddin Khan, and EihabBashier Mohammed Bashier,"*Machine learning: algorithms and applications"*Crc Press, 2016.

[13]Fatima, Meherwar, and MarufPasha,"Survey of machine learning algorithms for disease diagnostic", *Journal of Intelligent Learning Systems and Applications* 9.01 (2017): 1.

[14]Osisanwo, F. Y., et al. ,"Supervised machine learning algorithms: classification an dcomparison.", *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017): 128-138.

[15]Guo, Gongde, et al. "KNN model-based approach in classification." *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer Berlin Heidelberg, 2003.

[16] Guo, Gongde, et al. "KNN model-based approach in classification." On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003.

[17] Zhang, Shichao, et al. "Learning k for knn classification." *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.3 (2017): 1-19.

[18] Zhang, H., Berg, A.C., Maire, M. and Malik, J., 2006, June. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 2126-2136). IEEE.

[19] Wang, Haifeng, and Dejin Hu. "Comparison of SVM and LS-SVM for regression." *2005 International conference on neural networks and brain*. Vol. 1. IEEE, 2005.

[20] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 32-36. IEEE, 2004.

[21] Vishwanathan SV, Murty MN. SSVM: a simple SVM algorithm. InProceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) 2002 May 12 (Vol. 3, pp. 2393-2398). IEEE.

[22] Cherkassky, Vladimir, and Yunqian Ma. "Practical selection of SVM parameters and noise estimation for SVM regression." *Neural networks* 17, no. 1 (2004): 113-126.

[23] Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR).[Internet]* 9 (2020): 381-386.

[23] Bi, Q., Goodman, K. E., Kaminsky, J., &Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology*, *188*(12), 2222-2239.

[24] Mair C, Kadoda G, Lefley M, Phalp K, Schofield C, Shepperd M, Webster S. An investigation of machine learning based prediction systems. Journal of systems and software. 2000 Jul 15;53(1):23-9.

[25] Mackenzie, Adrian. "The production of prediction: What does machine learning want?." *European Journal of Cultural Studies* 18.4-5 (2015): 429-445.