Journal of Advanced Research in Instrumentation and Control Engineering
Volume 10, Issue 1 - 2023, Pg. No. 1-5
Peer Reviewed Journal

Research Article

Analysis Study and Research on Heart Attack Prediction System

Harkirat Singh[1], Shivam Gupta[2], Vinay Chopra[3]

[1,2]Student, DAV Institute of Engineering and Technology, Jalandhar, Punjab, India.
[3]Assistant Professor, CSE DAV Institute of Engineering and Technology, Jalandhar, Punjab, India.

INFO

Corresponding Author:
Harkirat Singh, DAV Institute of Engineering and Technology, Jalandhar, Punjab, India.
E-mail Id:
harkiratsingh.0212@gmail.com
Orcid Id:
https://orcid.org/0009-0003-6784-0549

How to cite this article:
Singh H, Gupta S, Chopra V. Analysis Study and Research on Heart Attack Prediction System. J Adv Res Instru Control Engi. 2023; 10(1): 1-5.

Date of Submission: 2023-04-18
Date of Acceptance: 2023-05-05

ABSTRACT

A blood clot disrupts blood flow to the heart, causing a heart attack. This may cause chest tightness, fullness, squeezing, or discomfort. This discomfort may spread.

The study uses machine learning to predict heart attacks. A big dataset of heart attack survivors and non-survivors trained and tested the algorithm. The algorithm calculates a patient's risk of a heart attack based on age, blood pressure, cholesterol levels, and other factors. Logistic regression, random forests, and support vector machines are machine learning methods. Finally, the novel heart attack prediction method may increase early diagnosis and prevention.

Keywords: Logistic Regression, Random Forest, and SVM are Keywords

Introduction

The muscular heart pumps blood throughout the body. The cardiovascular system circulates blood via a network of arteries. It includes the heart. cardiac disease may result from cardiac blood outflow disturbance. Low-income and middle-aged nations have significant cardiovascular disease mortality. Analysing large amounts of EHR data may predict cardiac disease. Data may be extracted to treat patients. Machine learning is utilised for this.

Examples Include:

1. Circulation published a machine learning heart failure prediction model. 2. Using data from over 5,000 cases, the system accurately predicted heart failure in people.
2. A PLOS One study used machine learning to predict cardiovascular disease risk in over 150,000 individuals. This method may identify high-risk cardiovascular disease patients to guide preventative measures.
3. Scientific Reports uses machine learning to find new heart disease risk factors. Using 300,000 patient data, the researchers created a machine learning model. This model showed that air pollution, social isolation, and alcohol intake increased heart disease risk. Learning machines might increase heart disease understanding and help physicians design more personalised illness preventive measures.

Related Work

Babu and colleagues[1] used UCI data for their investigation. Several methods may help us comprehend heart illness. SMO (89% accuracy) and Bayes Nett (87% accuracy) are more accurate than KStar (73% accuracy) and Multilayer perceptron (70% accuracy). Bayes Nett is 87% accurate, whereas SMO is 89% accurate. These algorithms are too imprecise.

Cai et al.[2] suggest using ANN and SVM to predict stroke patients using Kaggle data. ANN and SVM had an accuracy of 81.82% and 80.38% for the training data set, and 85.9% and 84.26% for the test dataset. Dangare et al.[3] test four machine learning algorithms-Naive Bayes, SVM, Decision Tree, and ANN-using UCI repository data. ANN's 85.3% accuracy is best. Decision Tree earned 80%, while Nave Bayes and SVM received 78%. Umair Shafique and colleagues[4] analyze machine learning approaches using WEKA. Combining ANN and PCA improved performance. PCA boosts its accuracy from 94.5% to 97.7%. This causes a big gap.[4]

Random Forest, Decision tree SVM and SVM were used to predict Cardiovascular Disease, with Random Forest having the greatest accuracy of 85%. Decision tree SVM and SVM were also used. Muhammad Usama Riaz.[5]

Methodology

The first step in the processing of the system is the collection of data. For this step, we utilize the dataset from the UCI repository, which has been thoroughly checked by a large number of UCI researchers and authorities.

- **Data Collection** - The initial step in the process of constructing a prediction system is collecting data and deciding which datasets will be used for training and testing the system. Within the scope of this study, we used 37% of the testing dataset and 73% of the training dataset to develop the system
- **Selection of characteristics** - The characteristics of the data included in the set are qualities of the data set that are used by both the system and the heart. The prediction algorithm takes into account a wide variety of factors, including a person's like heart rate, gender, age, and many more
- **Separation of Data:** The data are separated into a training set and a testing set. Only 25% of the data is utilized for actual testing, while the remaining 75% is put to use in actual training. We eliminated all the Nan values by using data normalization
- Data visualisation
- The pre-processing of data Pre-processing is necessary in order to get renowned results with machine learning algorithms. Because the Rand the om forest approach, for example, does not handle datasets with null values, we are required to manage null values from the raw data from which they were originally extracted
- For the purpose of our project, we are required to apply the following piece of code to convert certain categorised data to fictitious values in the form of "0" and "1"
- Data Balancing - In order to provide results that can be trusted, it is required to do data balancing since this proves that the target classes are comparable. The target classes are shown in Figure 3, where a "0" signifies those who have heart disease and a "1" indicates individuals who do not have heart disease

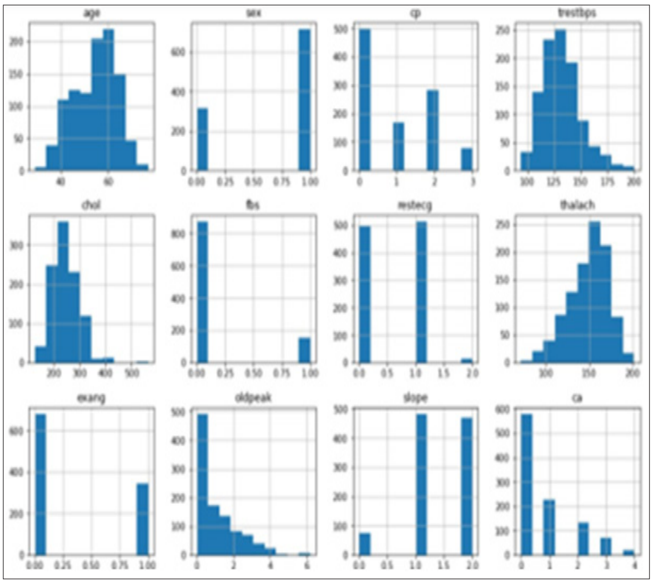
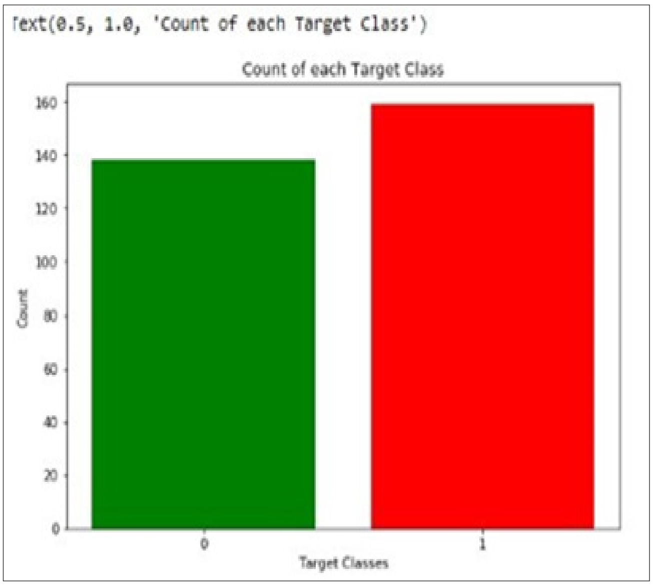


Figure 1.Bar Graph Visualization



```
Text(0.5, 1.0, 'Count of each Target Class')
```

Figure 2.Count of Each Target Class

Algorithms

An artificial neural network, often known as an ANN, is a kind of machine learning model that is inspired by the way the human brain works both structurally and functionally. It is made up of nodes (neurons) that are linked to one another and stacked in layers. It can be taught to recognise patterns in data thanks to its construction. To begin, we would need to do some preliminary processing on the data to convert it into a format that the ANN is able to make use of. Scaling the data, encoding categorical variables, and separating the data into training and testing groups might all be necessary steps in this process. Accuracy, Precision, Recall, F1-score, and a number of other metrics are examples of performance measurements.

Accuracy is calculated by dividing total points by total points plus total points plus total points plus total points plus total points.

The formula for recall, also known as sensitivity, is test performance divided by test performance plus false negatives, which is denoted by equation 2.

F1-Score=(2(Precision*Recall))⁄(Precision+Recall) (eq.3)

Precision is calculated by dividing the total pressure by the total pressure plus the final pressure (equation).

The number of true positives is denoted by TP, the number of true negatives by TN, the number of false negatives by FN, and the number of false positives by FP.

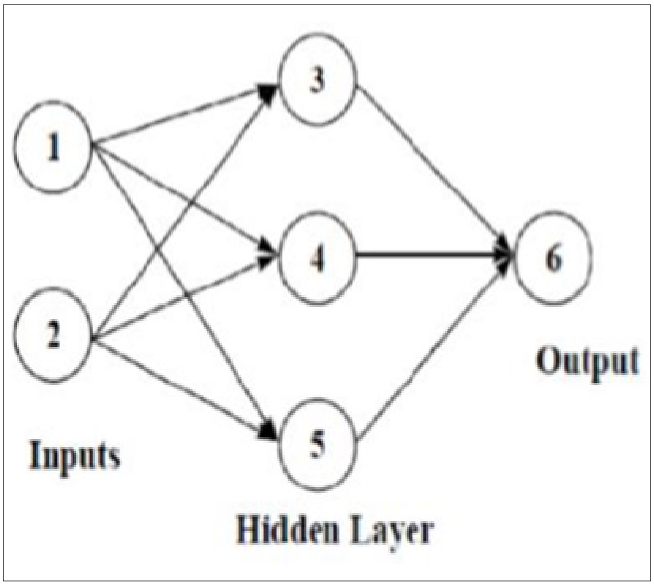


Figure 3.Samples of Artificial Neural Network

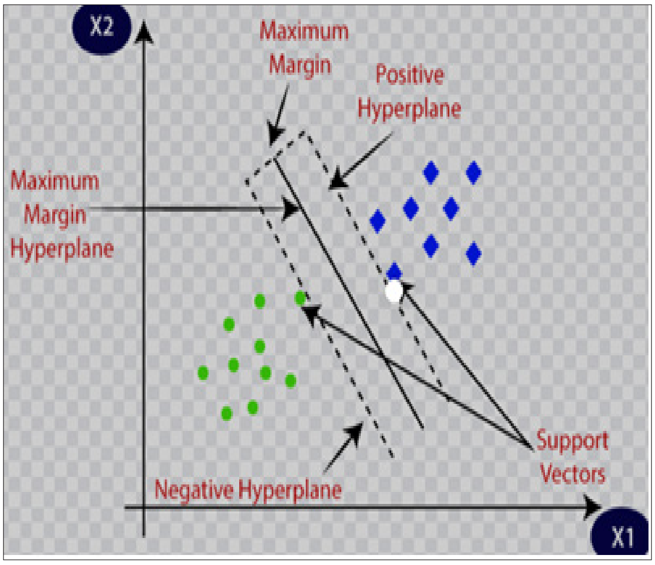


Figure 4.Hyperplane Structure through SVM

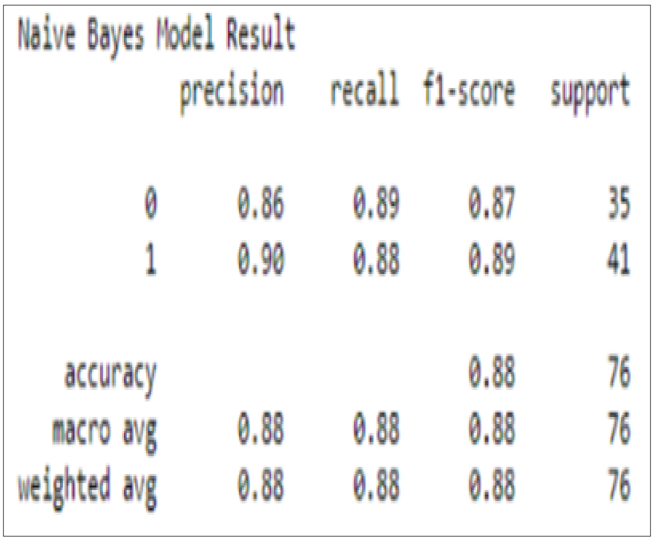
SVM:

The Support Vector Machine, or SVM, is a sort of algorithm that belongs to the field of machine learning. This technique may be used to problems involving classification or regression. On the basis of demographic and clinical information, SVM may be used in a heart attack prediction system to categorize people as being either high-risk or low-risk for developing heart disease.

Naïve Bayes:

The theorem of Bayes provides the foundation for the classification approach known as Naive Bayes. The naïve Bayesian classifier theorem states that the frequency of occurrences of some features of a class are unrelated to the presence or absence of other qualities. It is a reliable indicator of the presence of heart disease.

The Naive Bayes algorithm is used[4] in order to determine the posterior probability of each class based on the conditional likelihood of correctly categorising data sets. The equation may be found in the following. (C|X) =P X|C P(C) P(X), where X is the predicted instance and C is the class value, the aforementioned formula or equation helps in finding the class in which the feature is anticipated to categorise. X is the predicted instance, and C represents the class value.

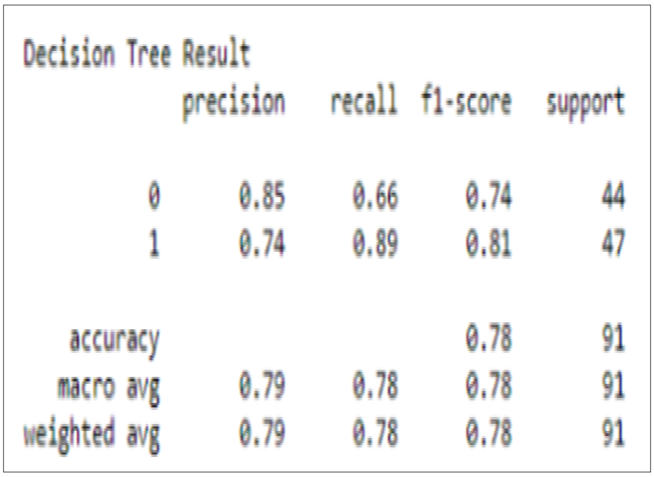


```
Naive Bayes Model Result
              precision    recall  f1-score   support

           0       0.86      0.89      0.87        35
           1       0.90      0.88      0.89        41

    accuracy                           0.88        76
   macro avg       0.88      0.88      0.88        76
weighted avg       0.88      0.88      0.88        76
```

This model gives an accuracy of 88 %.

Decision Tree:

A decision tree is a straightforward example of an algorithm for supervised learning that can classify data. They are responsible for working with numerical and categorical data. The decision tree is organised in the form of a tree with core nodes, branches, and leaf nodes; each branch of the tree reflects a different value from the data set being considered.



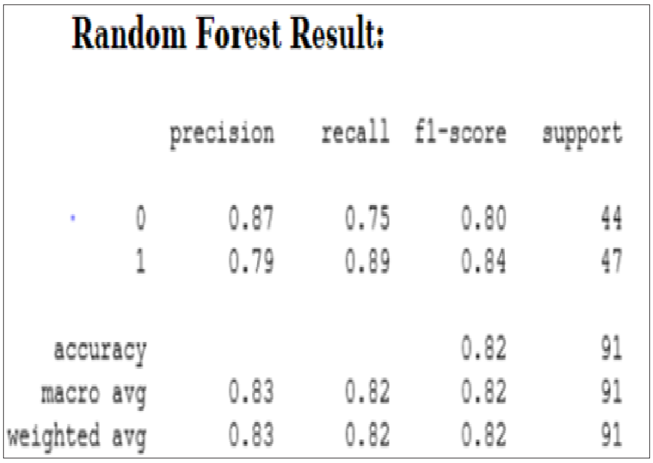
```
Decision Tree Result
              precision    recall  f1-score   support

           0       0.85      0.66      0.74        44
           1       0.74      0.89      0.81        47

    accuracy                           0.78        91
   macro avg       0.79      0.78      0.78        91
weighted avg       0.79      0.78      0.78        91
```

This model gives an accuracy of 78%.

Random Forest:

The method known as Random Forest is used for both classification and regression ensemble learning. At the time of training, it builds a number of different decision trees and then produces the classification that corresponds to the mode of the classes (regression) or the mean prediction (classification) of the individual trees.

It is possible to utilise the Random Forest algorithm in a heart attack prediction system in order to forecast the probability of a heart attack based on the presence of a number of risk variables.



```
Random Forest Result:

              precision    recall  f1-score   support

           0       0.87      0.75      0.80        44
           1       0.79      0.89      0.84        47

    accuracy                           0.82        91
   macro avg       0.83      0.82      0.82        91
weighted avg       0.83      0.82      0.82        91
```

This model is 82% Accurate.

Discussion

The use of data mining techniques to the treatment of heart disease was the primary focus of our investigation. Using five different data mining methods, we conducted various studies on our dataset on heart illness. By using a number of distinct classification algorithms, our team is working towards the goal of identifying the one that is most accurate at predicting the presence of heart disease. Which of these statements is the most accurate? We carried out five separate tests, all of which were geared at evaluating the efficacy of various machine learning algorithms, including SVM, neural networks, decision trees, naive bayes, and random forest.

A comparison of the Algorithms that were Actually Implemented.

The treatment of heart illness was the focus of our study, and we aimed to do this by using machine learning algorithms in the medical field. As a direct consequence of this, we carried out clinical studies on people suffering from heart disease using a number of different algorithms. Through trial and error, we will be able to ascertain which categorization approach is preferable for estimating the risk of developing heart disease.

The first part of this process involves the deployment of many machine learning algorithms. The second stage involves comparing the various machine learning algorithms that were used in these tests and selecting the machine learning algorithm that delivers the greatest level of accuracy. To provide a meaningful comparison between these trials, other performance metrics, such as accuracy, are used. ROC Curve, True Positive, False Positive, True Negative, and False Negative are some of the words that are used in this context. The compilation of algorithms is shown in the table below.

The values shown in table 2 indicate that the lowest possible accuracy on the available dataset is 78%, while the highest possible accuracy is 88%. The accuracy of the Naive Bayes model is the greatest, while the accuracy of the Decision Tree model is the lowest.

By keeping an eye on the several other performance metrics that are also used for outcomes. SVM has a TP rate of 40, ANN has a rate of 36, Naive Bayes also has a rate of 36, a Decision tree has a rate of 29, and a Random forest has a rate of 33. It can be seen from this that the SVM has the greatest true positive rate, whilst the Decision Tree has the lowest true positive rate. In a similar manner, the ANN has the lowest FP rate of 2, but the Decision tree has the greatest FP rate, which is 15.

The comparison that was just done demonstrates that Naive Bayes, Support Vector Machines, and Artificial Neural Networks are all excellent methods since they all have almost same accuracy and they have the highest TP rate. The FP rate for the SVM is 7, whereas the FP rate for the naive Bayes model is 4, and the FP rate for the ANN model is 2.

As is well knowledge, cardiovascular disease is a serious condition that is responsible for the deaths of millions of people. Because of this, we need to maintain a high TP rate while maintaining a lower FP rate. Patients who are afflicted with a disease are more likely to be cured when the condition is correctly diagnosed and when the diagnosis is made at an earlier stage. Therefore, it is to be anticipated that algorithms will perform admirably. When identifying people with heart disease, accuracy is also very important.

Concluding Remarks

The identification of heart disease was the primary focus of our research, which centred on the use of data mining techniques in medical settings. Diseases of the heart are significant conditions that may ultimately lead to death. The KNN, Neural Networks, Decision Trees, Naive Bayes, and Random Forests algorithms were used in the implementation of the data mining approaches. The performance of the system was evaluated using a number of different methods, as well as accuracy, TN, FP, FN, and TP rates.

In order to predict heart disease, we carried out five separate studies using the same data set. The outcomes of every algorithm that was built are presented in tabular format for ease of understanding and comparison. The findings of the experiment indicate that the Naive Bayes algorithm has the greatest accuracy, coming in at 88%, followed by the ANN algorithm and the SVM algorithm, both of which have an accuracy of 87%.

References

1. Babu S, Vivek E, Famina K, Fida K, et al. Heart disease diagnosis using data mining technique. In: 2017 International conference Using Machine Learning for Heart Disease Prediction 13 of *Electronics, Communication and Aerospace Technology (ICECA)*. 2017;1:750-753. IEEE.
2. Cai J, Luo J, Wang S, et al. Feature selection in machine learning. A new perspective. *Neurocomputing*. 2018;300:70-79.
3. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*. 2012;47(10):44-48.
4. Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser. “Data Mining in Healthcare for Heart 2015.
5. Muhammad Usama Riaz, Shahid Mehmood Awan, Abdul Ghaffar Khan, “Prediction of Heart Disease Using Artificial Neural Network”, 2018.
6. Komal Kumar, Napa G. Sarika Sindhu D, et al. “Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. 2020.
7. R El-Bialy MA. Salamay OH. Karam and ME Khalifa, “Feature analysis of coronary artery heart disease data sets,” Procedia Computer Science. 2015;65:459-468.
8. K Mankad PS. Sajja and R Akerkar “Evolving rules using genetic fuzzy approach - an educational case study,” *International Journal on Soft Computing (IJSC)*, 2011;2(1):35-46.
9. Heart disease dataset: https://archive.ics.uci.edu/ml/datasets/heart+Disease.
10. T M Franke T. Ho and CA. Christie, “The chi-square test: often used and more often misinterpreted,” American Journal of Evaluation. 2012;33(3):448-458.
11. Y Zhang, “Support vector machine classification algorithm and its application,” *Information Computing and Applications*. ICICA 2012, Springer, Berlin, Heidelberg. 2012.
12. G Ke Q. Meng T. Finley et al., “Light GBM: a highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*. 2017;30:3149-3157.