I.J. Information Technology and Computer Science, **2023, 5, 45-54**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2023.05.05

A Comparative Analysis of Algorithms for Heart Disease Prediction Using Data Mining

Snigdho Dip Howlader*
Department of Computer Science, American International University - Bangladesh, Dhaka, Bangladesh
E-mail: snigdho.howlader@gmail.com
ORCID iD: https://orcid.org/0009-0005-2949-3042
*Corresponding Author

Tushar Biswas
Department of Computer Science, American International University - Bangladesh, Dhaka, Bangladesh
E-mail: tusharbiswas1014@gmail.com
ORCID iD: https://orcid.org/0009-0005-5805-9835

Aishwarjyo Roy
Department of Computer Science, American International University - Bangladesh, Dhaka, Bangladesh
E-mail: aishwarjyo@gmail.com
ORCID iD: https://orcid.org/0009-0007-3973-2523

Golam Mortuja
Department of Computer Science, American International University - Bangladesh, Dhaka, Bangladesh
E-mail: saiham394@gmail.com
ORCID iD: https://orcid.org/0009-0004-9973-4190

Dip Nandi
Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh
E-mail: dip.nandi@aiub.edu
ORCID iD: https://orcid.org/0000-0002-9019-9740

Received: 23 March 2023; Revised: 15 June 2023; Accepted: 20 July 2023; Published: 08 October 2023

Abstract: Heart disease is very common in today's day and age, with death rates climbing up the numbers every year. Prediction of heart disease cases is a topic that has been around in the world of data and medical science for many years. The study conducted in this paper makes comparison of the different algorithms that have been used in pattern analysis and prediction of heart diseases. Among the algorithms that have been used in the past included a combination of machine learning and data mining concepts that essentially are derived from statistical analysis and relevant approaches. There are a lot of factors that can be considered when attempting to analytically predict instances of heart diseases, such as age, gender, resting blood pressure etc. Eight such factors have been taken into consideration for carrying out this qualitative comparison. As this study uses a particular data set for extracting results from, the output may vary when implemented over different data sets. The research includes comparisons of Naive Bayes, Decision Tree, Random Forest and Logistic Regression. After multiple implementations, the accuracy in training and testing are obtained and listed down. The observations from implementation of these algorithms over the same dataset indicates that Random Forest and Decision Tree have the highest accuracy in prediction of heart disease based on the dataset that we have provided. Similarly, Naive Bayes has the least accurate results for this scenario under the given contexts.

Index Terms: Heart Disease Predictions, Machine Learning, Data Mining.

1. Introduction

The human body is a very complicated system composed of many different sub structures. One of these more important sub structures is the cardiovascular system, at the center of which is the heart. [1] The importance of the heart

is simple, even though the procedure is very complicated. As our body requires oxygen, the heart makes sure that oxygen and nutrients are distributed all throughout the body in a uniform manner. Such importance of the heart is what makes it an alarming event whenever heart diseases are identified, diagnosed or experienced. [2] heart diseases come in many different forms; however, they share some characteristics that remain the same or at least similar throughout. Heart diseases can be ones that have the patient experience high blood pressure, coronary artery disease, strokes or cardiac arrests and so on. They are all commonly classified as cardiovascular diseases. An age old question amongst medical experts has been to identify factors that contribute to heart diseases. [3] Knowing the chances of any individual experiencing heart diseases or heart problems can help medical experts greatly to prevent them. By identifying, even an approximation of the chances of a patient encountering heart diseases, both the patient and the physicians can take required measurements, such as the type of diet they should be maintaining. There are no precise or proper solutions for this right now, and research is being conducted regularly on how it may be possible to make the prediction results more accurate. There are various limitations to why properly useful results are unable to be obtained from studies in this field. An important limitation is that when it comes to prediction of heart diseases, the impact of environment, nutrition, fitness and many other factors make prediction more difficult. The same study done on two different data sets will yield very different results because of things like age and diet. The goal of this study is to take four algorithms that have been widely used for identifying heart disease prediction in the past, and draw them up for comparison against one another. The difference in their results will be able to suggest which algorithm or algorithms are better suited in this scenario, and therefore why they should be used more in the future.

2. Related Works

2.1. Overview of Cardiovascular Diseases

According to the World Health Organization (WHO), there has been a rise in cardiovascular diseases all throughout the world. In 2012 alone, there were an estimated 17.5 million people who died from cardiovascular diseases. According to another report by the WHO, 17.9 M people pass away every year from cardiovascular diseases. This is an astounding 31% of all deaths over the world, and 85% of the CVDs are either due to stroke or heart attacks. There is a lack of preventive measures being taken by people, which lead to increasing rates in heart diseases. [4] Although medical data is collected in large amounts on a regular basis, not all of it proves to be too useful most of the time. The effectiveness is not always constant when looking for important relationships and patterns across data. Knowledge discovery has been used across different domains such as business for making the utmost usage of data. The same can be said about the usage of data in the medical industry. [5] Below are some of the various heart diseases that are most commonly experienced by people:

- Coronary Heart Disease: One of the more common forms of heart diseases, coronary heart disease is where the patient experiences an event when blood vessels become clogged or blocked. This reduces blood flow and slows down the supply of oxygen to the heart.
- Congestive Heart Failure: An illness that progressively weakens the heart's muscles and their ability to pump blood. The final stage of this illness leads to the patient not getting enough blood pumped by the heart. [6]
- Angina: Angina, or Angina Pectoris is a medical condition that leads to chest pain and discomfort because of the heart not receiving enough blood to pump. [7]
- Arrhythmia: Inability of the heart to maintain a stable pulse. Irregular heart beats can be very problematic, cause extreme discomfort and also lead to extreme amounts of pain due to the irregular intervals.

2.2. Discovering Patterns across Patients' Data

There are various types of algorithms that have been used for finding patterns across medical data, including ones that deal with heart diseases. [8] This study will try to draw a comparison between some of the more widely practiced algorithms that are used for finding patterns across heart disease data. A variety of different algorithms have been carried across datasets of heart disease patients over the years. Tweaking each algorithm slightly has resulted in different outputs of accuracy. This is applicable to patient data sets that are outside of the domain of heart disease as well, such as skin disease by Anurag Kumar Verma et. al. [9] A large number of studies and experiments have been carried out using various techniques, such as and not limited to: Naive Bayes, K-NN Clustering Support Vector Machine, Neural Networks.

These, and a lot of other techniques will be discussed in detail and their existing results so far. The main goal for each of the studies have always been the same, however - To identify a method that can predict the chances of an individual to experience heart diseases. [10] This ends up making use of the huge bulk of data that is gathered annually by hospitals and results in outputs that help patients and the overall medical industry. In the long run, there are possibilities that these bulks of data may contain information that can save many lives. [11].

2.3. Problem Statement

While there are plenty of algorithms and techniques to mine data, it is a very vast field of research and thus, it makes it a case to case situation. The same technique will fail to provide the most accurate results in every context, as different datasets have vastly different factors. For this study, the aim is to make a comparison between three of the most commonly

used data mining techniques in the heart disease sector, i.e. Naive Bayes technique, the Decision Tree Technique, the Random Forest technique and the Logistic Regression Approach. The goal will be to give an explanation for each of the techniques, how they are carried out and how they differ from each other in terms of memory, complexity and accuracy. Data mining problems like these often require the data to be in a specific format, and that pre-processing will also be required. Clean data is not always easy to find, and it will ease the process to implement the data mining techniques for the best results.

The objective of this study is to run tests across the three algorithms/techniques that are mentioned, and to find out which produces the most accurate results. It is to be noted that the results may vary if the dataset is changed, as different datasets will always contain certain inconsistencies that result in the learning step of a machine learning implementation differently. This will hopefully give a better insight on which of these three aforementioned algorithms prove to be superior, and thus which is more likely to be used when carrying out a prediction study such as this.

2.4. Background

For a long time, the healthcare industry has been at it to find patterns for all sorts of diseases and health disorders. Scanning and mining large data sets in order to find valuable information in them is performed to make decisions earlier and possibly make predictions for disease contraction before the patient actually experiences or contracts it. [12, 13] Data Mining has become one of the top fields of research in computer science, and the main reason for this is perhaps due to businesses and industries benefiting much more from utilizing data than without it. Even though the medical industry produces such a bulk amount of data, most of it ends up being useless. Therefore, it has proved to be very beneficial over the recent years when big data from the medical industry have been attempted to be analyzed, mined and utilized. [14, 15] There have been various studies over the decades. This has gone as advanced as trying to predict what kinds of inherited illnesses are more likely to occur than others, as proposed by Lahiru Iddamalgoda et al. They discovered that most inherited diseases are most likely to be caused by single nucleotide polymorphism variants and these SNPs are occurring more prominently. The K-Nearest Neighbor algorithm proved to be the most efficient algorithm for studying inherited diseases. [16, 17] As factors constantly keep changing from disease to disease, it is very important to note that each type of disease works with different factors and will yield best results under very different algorithms. Before putting the data into use, it is always important for risk factors to be identified as the parameters that the data will be classified or mined under. Factors associated with heart disease include age, blood pressure, sugar levels in the blood, obesity etc. [18] For multiple studies [4] over the last decade or two, the risk factors that have been identified have varied but have mostly stuck to containing some common fields. These are:

- Age
- Sex
- Type of chest pain
- Resting blood pressure
- Cholesterol
- Fasting blood sugar
- Maximum heart rate
- Exercise induced angina

2.5. Data Mining in Heart Disease Prediction

Amongst the various algorithms and techniques that have been used to explore diseases containing heart disease factors, some of the most common are. Some of the more commonly used techniques include Classification, Clustering, Association and Prediction. [19, 20] Amongst these, classification and clustering techniques have proven to be more prominent in the domain of heart disease prediction. [21] Chaitrali S. Dangare et al. tested out Naive Bayes, Decision Trees and Neural Networks to carry out on a dataset of 573 records in total in WEKA. The results ended up showing that Neural Networks proved to be more accurate when compared to Naive Bayes and Decision Trees. It was suggested that their results may vary upon using a different dataset, or switching out variables. A sophisticated system was proposed by V. Krishnaiah et al. that would be able to remove uncertainty in heart disease datasets using a Fuzzy K-NN approach. The experiment resulted in removing redundant data to make it more useful and for better results for future use. It was found in another study that amongst J48 Decision Tree, K-Nearest Neighbor, Naive Bayes and Sequential Minimal Optimization, the classifier with the highest accuracy value was the J48 Decision Tree. According to this study, J48 Decision Tree obtained results good enough to carry out in a clinical environment, but results may vary with change of data. A hybrid approach was proposed by Saman Iftikhar et al. which made utilization of the SVM classifier, Genetic Algorithm and Particle Swarm Optimization. It was noticed that GA and PSO algorithms were able to select discriminative features to make sure that the SVM classification worked to its fullest. A study by Divya Tomar and Sonali Agarwal. was able to conclude stating that no individual data mining technique could give consistent results for all types of medical or healthcare data. Performance of such techniques solely depends on the cleanliness and types of datasets that are being provided. A clean, and more favorable data will always provide more accurate results than one that does not. Therefore, drawing comparisons across different techniques is only able to assure what kind of dataset a set of techniques will perform best or worst on. [22]

A prototype was suggested for a heart disease prediction system by et al. that made use of three different data mining

techniques. They were: Naive Bayes, Neural Networks and Decision Trees. In this case, it was identified that the Naive Bayes algorithm proved to result with better accuracies than the other two. It was suggested that Time Series, Clustering and Association Rules can also be incorporated into the system to make it more useful and let it result more dynamically. In a different study, the Naive Bayes proved to be superior once again when classifying heart disease instances with a higher accuracy when compared to PSO. Namely, the Naive Bayes combined with the Genetic Algorithm proved to be very accurate when compared to Naive Bayes combined with the PSO. However, it was suggested that the PSO was better at noting the different aspects in feature selection. It was evident in most cases that the Naive Bayes and Neural Network techniques were superior compared to a lot of other techniques when conducting implementation in this domain. However, the Neural Network can be more ineffective in terms of memory utilization and is slower, and this is significant when considering the amount of data that will be analyzed or learned from in this domain of research. [23] In another study, it was identified that the Naive Bayes algorithm proved to contain significantly higher accuracy when put up against the Decision Tree when applied on the Cleveland and Statlog datasets. There were certain advantages and disadvantages when applying any of the algorithms on these datasets, but it was clear that for effective results,a multitude of algorithms would be required over the long run. [24] et al. were able to conclude stating that over time, other medical sectors, heart diseases and CHD included, would require Machine Learning and Artificial Intelligence to make it possible for predicting illnesses or their outcomes, or at least get to a state of prevention. However, this would get more complicated with time as different diseases contained different datasets, with vastly different factors involved that would result in the data mining and machine learning techniques behaving very differently and resulting nonuniformly.

In another study by Devansh Shah et al, a comparison was made between Naive Bayes, Decision Tree, Random Forest and K-Nearest Neighbor. The dataset had to go through pre-processing before these algorithms were implemented, and the results showed that the K-Nearest Neighbor algorithm showed the most accurate results, where 14 attributes were taken into account. S. Ramasamy et al. proposed to create a detection system using association rule mining, along with keyword based clustering algorithms. In that study, the keyword based clustering algorithm was able to come up with the more accurate results. [25] et al examined a group of different data mining techniques, such as Decision Tree and SVM for performing classification - which resulted in them suggesting a group of factors regarding prediction of heart diseases. The factors are: age, sex, chest pain, blood pressure, personnel history, previous history, cholesterol, blood sugar when fasting, resting ECG and maximum heart rate. [26] A study on Apriori carried out further was able to show that Apriori is a fairly effective algorithm when it comes to heart disease datasets, which was unlikely as this algorithm is more famous for finding prediction through itemsets in analysis of market and stock data. It is to be noted that this study was carried out by WEKA and MATLAB simultaneously. [27] In a different study, it was concluded that despite assessing the risks involved in contraction of cardiovascular diseases, the risk reduction is not always adequately reflected. The most definite statement that was made was the chances of cardiovascular diseases were incremented as age increased. [28] One of the bigger challenges of studies in this field was identifying where to collect data from. A study suggested that it was crucial to identify if the Hospitals were maintaining data properly before approaching them for their data. It was also important to understand the overall quality of a hospital before their data is used for a study. [29] A separate study tried to look into the reliability of a dataset in a study. The results of an algorithm on a dataset may be accurate with the real life scenario, but this result may not be similar when using a separate dataset. Multiple factors, such as null values and overfitting contribute to results being flawed in such circumstances. [30] A separate study done on comparing KNN, SVM, C5.0, Logistic Regression and Neural Networks showed that the Decision Tree was able to achieve the best results. Decision Trees were also easier to implement when compared to the other algorithms for this study. [31] A set of ensemble algorithms were used in another study such as bagging, boosting and stacking. The accuracy is more improved when bagging is used. The weaker classifiers were ensemble with majority voting, which had very poor accuracy. [32]

2.6. Discrepancies and Challenges

Haleh Ayatollahi et al. conducted a study to compare the results between ANN and Standard Vector Machine. They concluded the study with results that suggested that SVM had a better fitness of the data with a reduced level of error. [33] It should also be noted that modern datasets are more likely to result in better outcomes with machine learning and data mining algorithms compared to that of the past, due to heart diseases becoming more prominent. Another study conducted to make a comparison between the performance of different algorithms resulted in the Random Forest algorithm being the most efficient when compared to others, namely Logistic Regression and Decision Tree. The study mentioned a possible better scope for the future with a more stable, appropriate dataset for more consistent results. [34]

Besides discrepancies in data, there are other concerns when it comes to prediction of heart diseases using machine learning. One of the biggest concerns is the generalization and struggle of interpretation. Data mining and machine learning models find it difficult to interpret data from different populations. This often leads to overfitting of data and therefore an overall poor performance across the research conducted. Additionally, ethics are also a concern whenever data in medicine and healthcare is concerned. The data in this regard is always concerned with the patient's privacy, consent and liability. Often, researchers find it hard to get their hands on data for machine learning and data mining purposes for prediction of heart diseases, or the medical sector in general.

3. Methodologies

As mentioned above, the objective for this research is to figure out which algorithm proves to be the most accurate and efficient amongst four of the most prominently used algorithms in heart disease prediction research. The techniques that will be used are taken on a basis of popularity and effectiveness. They are the Naive Bayes technique, the Decision Tree Technique, Random Forest Technique and the Logistic Regression Approach. The following study will be carried out using the Cleveland (which was obtained from Kaggle) dataset for heart diseases, drawing a comparison between three different data mining techniques. The dataset contains a total of 14 different factors that will be taken into account, and will result in boolean values to say an instance will experience heart disease, or not. The implementation will be carried out in Python. The same Cleveland dataset will be used with the same factors for both approaches. The Python code will be implemented in Google Colab. The study conducted in this paper makes comparison of the different algorithms that have been used in pattern analysis and prediction of heart diseases.

The Naive Bayes

The Naive Bayes classifier is a popular probabilistic classification technique. It focuses on predicting class memberships and their probabilities. It assumes that the presence of any one particular feature in a class is not related to any other feature at all, which is taken straight from the Bayesian Theorem. [33] The biggest reason why the Naive Bayes method is so popular in Machine Learning and Data Mining is because it is able to be applied across large datasets quite easily and simply and is widely known for outperforming other techniques that are considered highly sophisticated.

The pseudocode for the Naive Bayes algorithm is provided below:

Step 1: Read and store the training dataset
Step 2: Generate the Mean and Standard Deviation for the predictor variables
Step 3: Continue
    Calculate probability of the next instance for each class
    Stop when all predictor variables have been iterated
Step 4: Generate the likelihood for each of the classes
Step 5: Find the higher likelihood

The following equation sums up Bayes' Theorem.

$$P(C_i \mid X) = \frac{P(C_i \mid X)P(C_i)}{P(X)} \tag{1}$$

Where

- P(Ci|X) is the probability of an event occurring after taking new information into consideration, which is also known as the posterior probability.
- P(Ci) is the probability of an event occurring before taking new information into consideration, also known as prior probability of the class.
- P(X|Ci) is the likelihood
- P(X) is the prior probability
- P(X) is the only constant for all classes

Another way of seeing this can be: Posteriori = Likelihood * Prior / Evidence

Naive Bayes is used widely in the field of machine learning and data mining for its simple, yet easy to understand algorithm. For the training phase of any data mining implementation, Naive Bayes is the best and fastest probabilistic classifier. [34]

Decision Tree

One of the more powerful techniques in data mining, Decision Trees are very popular when it comes to classification and prediction. Decision Trees are able to create rules, just like inferred by humans and thus they can be used in a database. There are different versions of the Decision Tree approach, such as the C4.5. A Decision Tree works very similar to a flow chart, and it is composed of leaf nodes and parent nodes. All of them contain a root node, which is the source of all other nodes. This technique may not always be able to efficiently handle data that contains repetition or replication of data. [35]

The aim of a Decision Tree is always to be able to generate a training model that can predict the class or value of the target variable. It can do this by learning the decision rules that are inferred from the prior data, as mentioned before. In order to predict a class, the traversal must be initiated from the root of the tree. For it to work, first of all, the best attribute is discovered from the dataset and set at the root of the tree. The training set is then split into subsets, which must be done in a way that each subset contains data with the same value for an attribute. The two steps are carried on until leaf nodes

can be discovered for all branches in the tree. That's the entire traversal of a generic decision tree.

Random Forest

Random Forest is a classification technique commonly used in machine learning, which has a tendency to work on non-linear datasets. A random forest is named so because it is built from hundreds of trees, all containing randomized inputs inside them. [36] Random Forests usually work with the Ensemble technique. Random forest can be used for both classification and regression problems, and is essentially a large collection of Decision Trees. [37]

A Random Forest classifier is able to handle missing values from a dataset, does not overfit a model (even with a lot of trees in the forest) and can model the forest classifier for categorical values. A random forest pseudocode would essentially be this:

Step 1: Select "x" features from "y" features at random where x is less than y
Step 2: Calculate the node "q" by using the best split point
Step 3: Split the node "q" into child nodes using best split point
Step 4: Continue the steps 1 to 3 until number "r" number of nodes are achieved
Step 5: Produce the forest by repeating steps 1 to 4 "n" number of trees

This will essentially create the random forest, which will contain the values for predictions as per requirement. [38]

Logistic Regression

Logistic regression is a process that is able to model the probability of a particular discrete outcome, based on an input that is provided to it. It is very useful when it comes to machine learning and data mining. [39] One of the particular advantages of using a Logistic Regression model is that it is able to produce simple probabilistic formulae for classification.

Logistic Regression is useful for the situations where the requirement is to predict the absence or presence of certain characteristics, based on the values provided by a set of predictor variables. While it is somewhat similar to linear regression, but suited towards models where dependent variables are dichotomous. [40]

4. Analysis and Discussion

In this paper, four different classifier algorithms are studied and experiments are conducted to find out the best algorithm among them. The dataset contains 1025 records. The total records are divided between two sub-datasets. One is for training which has 820 records and another is for testing which has 205 records. This research was performed using Python 3 and Google colab. The Dataset has been preprocessed to check if it has any missing values. After preprocessing the data set Naive Bayes, Decision Tree, Random Forest, Logistic Regression were applied. For better understanding, confusion matrices for each algorithm were used. A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. In this research the confusion matrix has two classes, so this is a 2X2 confusion matrix.

Here:

Class 0 = Does not have heart disease
Class 1 = Has heart disease

Table 1. Confusion Matrix for the cases of having heart disease and not having heart disease

	0	1
0	True Positive	False Negative
1	False Positive	True Negative

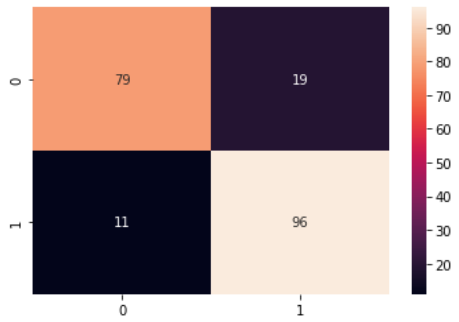

Fig.1. Confusion matrix for naive bayes

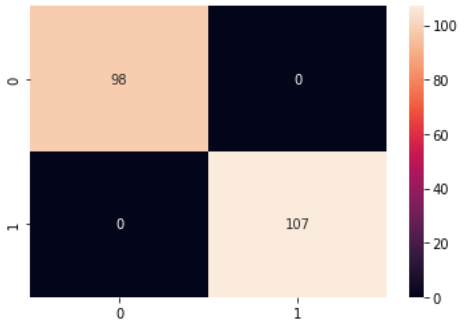

Fig.2. Confusion matrix for decision tree

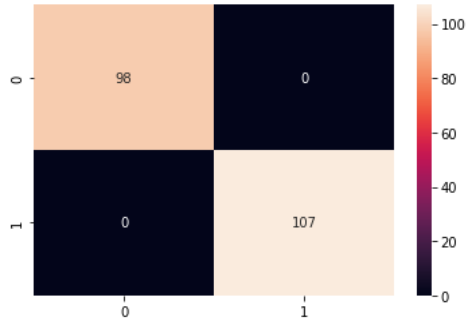

Fig.3. Confusion matrix for random forest

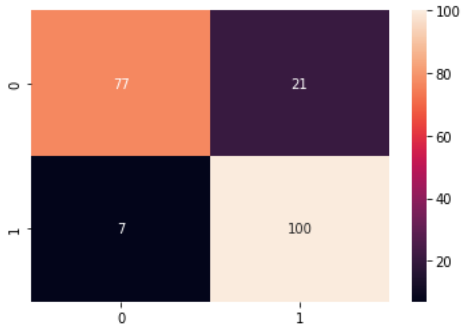

Fig.4. Confusion matrix for logistic regression

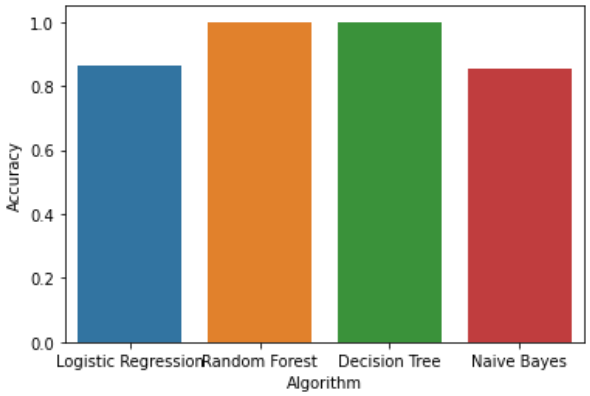

Fig.5. Comparative result of different algorithms

Figure 1, 2, 3 and 4 display the confusion matrices for the four algorithms that are being compared in this study. They give a visual representation of the performance of the models involved. It also helps understand and identify which classes have been correctly or incorrectly classified. Figure 5 shows the different accuracies that were obtained from the four algorithms, and clearly displays Random Forest and Decision Tree to be the best of the four in terms of accuracy. Logistic Regression is significantly less accurate, while Naive Bayes performs the least well for this case.

Table 2. Accuracy of different algorithms in training and testing phase

	Accuracy in training (%)	Accuracy in testing (%)
Logistic Regression	86.22	86.34
Random Forest	100.00	100.00
Decision Tree	100.00	100.00
Naive Bayes	82.07	85.37

5. Conclusions

The heart disease sector in the medical industry is one of the most active and complicated fields there is. The progress in data science and machine learning in medicine is significant, as it makes it easier for individuals to understand how to live their lives in order to avoid encountering certain illnesses. With such a vast number of techniques and algorithms already existing for predicting heart disease cases, it gets more and more difficult to properly keep track of the efficiency of each of these techniques. The study conducted here has therefore been done to present a definitive way of measuring the most efficient technique amongst four of the most commonly used techniques for heart disease prediction.

To ensure accuracy and reliability, the dataset that was used was obtained from Kaggle, the data of which are already preprocessed and cleaned for the most part. This same data set is used quite popularly in many other types of research that use instances of cardiovascular diseases. It has mostly been cleaned from anomalous values and outliers. Some of the data is still necessary for a realistic outcome of the research, as removal of all outliers is going to cause a risk of overfitting. The algorithms that were used in the comparison were all used in the past for researching on prediction of cardiovascular diseases, and therefore already are proven to be significantly good in this criteria. This study only solidifies their abilities to come up with accurate and efficient results, and to differentiate between them evidently.

In this paper, we worked with four different algorithms. From our observation we understand that Random Forest & Decision Tree has the highest accuracy and lowest number of false negative & false positive in confusion matrix. However, the accuracy of these two algorithms can vary in their depths, if depth is lower than the accuracy will decrease. Varying data might also lead to differences in results. We can further expand this research by working on optimization, adding new attributes, adding new data mining techniques.

References

[1] Chaitrali S. Dangare, Sulabha S. Apte, PhD. “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[2] V. Krishnaiah, G. Narsimha, and N. Subhash Chandra “Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach”, Springer International Publishing Switzerland 2015

[3] Devansh Shah, Samir Patel, Santosh Kumar Bharti, “Heart Disease Prediction using Machine Learning Techniques”, Springer Nature Singapore Pte Ltd 2020

[4] R Fadnavis, K Dhore, D Gupta, J Waghmare and D Kosankar, “Heart disease prediction using data mining”, International Conference on Research Frontiers in Sciences (ICRFS), 2021

[5] Boshra Bahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February - 2015

[6] S. Ramasamy, K. Nirmala, “Disease prediction in data mining using association rule mining and keyword based clustering algorithms”, International Journal of Computers and Application, 2017

[7] Pavithra M., Sindhana A.M, Subajanaki T. et al.“Effective Heart Disease Prediction Systems Using Data Mining Techniques”, Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021

[8] M. A. Nishara Banu, B. Gomathy, “Disease Predicting System Using Data Mining Techniques”, International Journal of Technical Research and Applications, 2013

[9] Anurag Kumar Verma, Saurabh Pal, Surjeet Kumar, “Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study”, Applied Biochemistry and Biotechnology, 2019

[10] Sellappan Palaniappan, Rafiah Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, IEEE, 2008

[11] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, “Heart Disease Diagnosis Using Data Mining Technique”, International Conference on Electronics, Communication and Aerospace Technology, 2017

[12] Sneha Grampurohit, Chetan Sagarnal, “Disease Prediction using Machine Learning Algorithms”, International Conference for Emerging Technology, 2020

[13] M. A. Nishara Banu, B. Gomathy, “Disease Forecasting System Using Data Mining Methods”, International Conference on Intelligent Computing Applications, 2014

[14] Hasib Qaiser, “Data Mining in Healthcare for Heart Diseases”, International Journal of Innovation and Applied Studies, 2015

[15] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”, IEEE Access, 2017

[16] Lahiru Iddamalgoda, Partha S. Das, Achala Aponso, Vijayaraghava S. Sundararajan, Prashanth Suravajhala and Jayaraman K. Valadi, “Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: Challenges and Implications”, Frontiers in Genetics, 2016

[17] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Hindawi, 2021
[18] M. A. Nishara Banu, B. Gomathy, "Disease Forecasting System Using Data Mining Methods", International Conference on Intelligent Computing Applications, 2014
[19] V. Krishnaiah, G. Narsimha, Ph.D., N. Subhash Chandra, Ph.D., "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications, 2016
[20] Sravani Nalluri, Vijaya Saraswathi Redrowthu, Somula Ramasubbareddy, Kharisma Govinda, E. Swetha, "Chronic Heart Disease Prediction Using Data Mining Techniques", Data Engineering and Communication Technology, 2020
[21] Umair Shafique, "Data Mining in Healthcare for Heart Diseases", International Journal of Innovation and Applied Studies, 2015
[22] Divya Tomar, Sonali Agarwal, "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology, 2013
[23] Ibomoiye Domor Mienyea, Yanxia Suna, Zenghui Wang, "An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk", Elsevier, 2020
[24] Md Mamun 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", Elsevier, 2021
[25] Neelam Singh, Santosh Kumar Singh Bhadauria, "Early Detection of Cancer Using Data Mining", International Journal of Applied Mathematical Sciences, 2016
[26] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012
[27] Mirpouya Mirmozaffari, Alireza Alinezhad, Azadeh Gilanpour, "Data Mining Apriori Algorithm for Heart Disease Prediction", Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE), 2017
[28] Michael J. Pencina, Ann Marie Navar, Daniel Wojdyla, MS Robert J. Sanchez, Irfan Khan, Joseph Elassal, Ralph B. D'Agostino Sr, Eric D. Peterson, Allan D. Sniderman, "Quantifying Importance of Major Risk Factors for Coronary Heart Disease", Circulation - The American Heart Association Inc., 2019
[29] Mary K. Obenshain, MAT, "Application of Data Mining Techniques to Healthcare Data", The Society of Healthcare Epidemiology of America, 2014
[30] Qi Rong Huang, Ph.D. Zhenxing Qin, Ph.D., Shichao Zhang, Ph.D., Chin Moi Chow, Ph.D., "Clinical Patterns of Obstructive Sleep Apnea and Its Comorbid Conditions: A Data Mining Approach", Journal of Clinical Sleep Medicine, 2008
[31] Moloud Abdar, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto, Goli Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases", International Journal of Electrical and Computer Engineering (IJECE), 2015
[32] C. Beulah Christalin Latah, S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Elsevier, 2019
[33] Haleh Ayatollahi, Leila Gholamhosseini, Masoud Salehi, "Predicting coronary artery disease: a comparison between two data mining algorithms", BMC Public Health, 2019
[34] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT), 2020
[35] Vikas Chaurasia, Saurabh, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2013
[36] Uma N. Dulhare, "Prediction System for Heart Disease using Naive Bayes and Particle Swarm", Biomedical Research, 2018
[37] Cincy Raju, Philipsy E, Siji Chacko, L Padma Suresh, Deepa Rajan S, "A Survey on Predicting Heart Disease using Data Mining Techniques", IEEE Conference on Emerging Devices and Smart Systems, 2018
[38] Yeshvendra K. Singh, Nikhil Sinha, Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", Springer Nature Singapore Pte Ltd., 2017
[39] Vikas Chaurasia, Saurabh Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2013
[40] Mohammed Khalilia, Sounak Chakraborty, Mihail Popescu, "Predicting disease risks from highly imbalanced data using random forest", BMC Medical Informatics & Decision Making, 2011

Authors' Profiles

Snigdho Dip Howlader completed his Bachelor's of Science in Computer Science and Software Engineering from American International University Bangladesh in 2021. His research interests are Software Engineering and Machine Learning.

Tushar Biswas completed his Bachelor's of Science in Computer Science and Engineering from American International University Bangladesh in 2021. His research interests are Machine Learning and Artificial Intelligence.

Aishwarjyo Roy completed her Bachelor's of Science in Computer Science and Engineering from American International University Bangladesh in 2021.Her interests are Software Quality Testing and Assurance.

MD. Golam Mortuja completed his Bachelor's of Science in Computer Science and Engineering from American International University Bangladesh in 2021. His research interest is Software Engineer.

Prof. Dr. Dip Nandi is the Director of Faculty of Science & Technology (FST), at American International University- Bangladesh (AIUB). His research interests include the concept of Software Engineering Model & Process, Data Science, E-Learning, Machine Learning etc.

How to cite this paper: Snigdho Dip Howlader, Tushar Biswas, Aishwarjyo Roy, Golam Mortuja, Dip Nandi, "A Comparative Analysis of Algorithms for Heart Disease Prediction Using Data Mining", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.5, pp.45-54, 2023. DOI:10.5815/ijitcs.2023.05.05