iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48764

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Heart Disease Prediction System Using Random Forest Technique

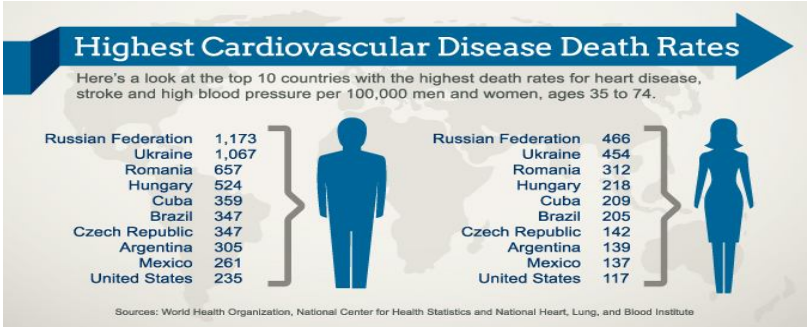
G. Harinadha Babu[1], Gunda Jayasree[2], Chattu Ashika[3], Vajja Ahalya[4], Katta Asha Niroopa[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Undergraduate Students, Department of Information Technology, KKR & KSR Institute of Technology and Sciences(A), Guntur, India.*

Abstract: ***Heart disease prediction and diagnosis have always been challenging tasks for medical experts. Hospitals and other medical facilities offer pricey procedures and treatments for cardiac diseases. As a result, being able to people all around the world can take the necessary actions to treat cardiac disease before it becomes severe if it is discovered in its early stages. The main causes of heart disease, a severe problem in recent years, are drinking alcohol, smoking cigarettes, and not exercising.***
A significant amount of data generated over time by the health care sector has allowed machine learning to offer efficient results in decision-making and prediction. We attempt to predict probable heart conditions in patients using machine learning approaches. In this project, we compare various classifiers, including decision trees, Naive Bayes, logistic regression, SVM, and random forests. We also suggest an ensemble classifier, which performs hybrid classification by combining the best features of both strong and weak classifiers and can handle large amounts of training and validation samples. We contrast already-in-use classifiers with others that have been proposed, such as Ada-boost and XGboost, which can offer higher accuracy. The main advantages of heart disease prediction using machine learning are that it manages the largest (enormous) quantity of data using the random forest algorithm and feature selection, as well as reducing the complexity of the doctors' time and being cost- and patient-friendly.
Keywords: Naive Bayes, Logistic Regression, Decision Tree, SVM, Random Forest

I. INTRODUCTION

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the main causes of illness and death among the world's population is heart disease. The prediction of cardiovascular illness is one of the most important subjects in the field of data analysis. The prevalence of cardiovascular disease has been rapidly increasing worldwide since a few years ago. Numerous research have been conducted in an effort to pinpoint the most crucial heart disease risk factors and precisely calculate the overall risk. Because it results in death without any overt symptoms, heart disease is sometimes known as the "silent killer." The ability to make decisions about lifestyle changes for high-risk individuals significantly depends on the early detection of cardiac disease, which reduces consequences.

The vast amount of data produced by the healthcare industry has made machine learning an effective tool for prediction and decision-making. By evaluating patient data that uses a machine-learning algorithm to categorise whether a patient has heart disease or not, this study hopes to predict future cases of heart disease. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. We may say that this technique can be very well adapted to accomplish the prediction of heart disease by gathering the data from many sources, classifying them under appropriate headings, and then analysing to get the needed data.



The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Additionally, the goal of this research is to determine the optimum classification method for detecting the likelihood of cardiac disease. Three classification algorithms, namely Nave Bayes, Decision Tree, and Random Forest, are employed at various levels of evaluations in a comparative study and analysis to support this work. Although these machine learning methods are widely utilised, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to evaluate the three algorithms. This will enable scientists and medical professionals to create a better.

II. LITERATURE REVIEW

An easy With the advancement of medical science and machine learning, various experiments and researches have been conducted in recent years, resulting in the publication of significant papers.

1) Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms .They used Cleveland dataset and pre-processing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set. A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.
2) AWAIS NIMAT et al. suggested a very effective expert system based on two support vector machines (SVM). Each of these two SVMs serves a specific function. The first one is used to eliminate extraneous characteristics, and the second one is used to make predictions. Additionally, they optimised the two approaches using the HGSA (hybrid grid search algorithm). Compared to the older traditional SVM models, they have achieved 3.3% greater accuracy with this model.
3) Ashir Javeed et al. built a model to address the issue of overfitting, which occurs when the suggested model performs and provides improved accuracy on testing data but provides worse accuracy results for training data while predicting the heart disease. They have created a model that will provide the best accuracy on both training and testing data in order to address this issue. The two methods used to anticipate the model are the RAS (random search algorithm) and the random forest algorithm. They obtained better outcomes using both training and testing data with the suggested model.
4) Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.
5) Deepika et al. presented predictive analytics using machine learning methods such naive Bayes, support vector machines, decision trees, and artificial neural networks to prevent and control chronic disease. To determine the accuracy, they used datasets from the UCI machine learning repository. The Support Vector Machine offers the best accuracy of 95.55% among them.
6) Lakshmana Rao et al. proposed "Machine Learning Techniques for Heart Disease Prediction" in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease.To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.
7) Anjan N. Repaka et al, proposed a model stated the performance of prediction for two classification models, which is analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.
8) Dr.kanak Saxena et al. created a data mining approach to accurately forecast cardiac disease. It basically assists medical professionals in making effective decisions based on the provided parameters. Age, sex, resting blood pressure, chest discomfort, serum cholesterol, fasting blood sugar, and other variables were employed by the author while using the Cleveland dataset from UCI. Additionally, they separated the datasets into two categories: training and testing. To determine accuracy, they employed a 10- fold approach
9) Senthil Kumar Mohan et al, proposed "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" in which their main objective is to improve 4 exactness in cardiovascular problems. The algorithms used are KNN, LR, SVM, NN to produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with linear model (HRFLM).

10) Aakash Chauhan et al, proposed “Heart Disease Prediction using Evolutionary Rule Learning”. Data is directly retrieved from electronic records that reduce the manual tasks. The number of services is decreased and shown major number of rules helps within the best prediction of heart disease. Frequent pattern growth association mining is performed on patient’s dataset to generate strong association.

III.PROBLEM IDENTIFICATION

All Heart disease is being stressed more and more as a silent killer that causes mortality without obvious symptoms. Growing concern about the illness and its effects is a result of the disease's nature. Therefore, efforts to foresee the possibility of this past lethal diseases are still present. To satisfy the demands of contemporary health, numerous technologies and methods are routinely tested. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. We can draw conclusions by compiling data from numerous sources, organising it under useful headings, and then analysing it to get the desired data. This method is very well suited for use in heart disease prediction. Early prognosis and its control can help to prevent & lower the death rates related to heart disease, as the adage goes, "Prevention is better than cure." Heart conditions comprise: diseases of the blood vessels, such as coronary artery disease

1) Abnormal heartbeats (arrhythmias)
2) Heart problems are a birth defect (congenital heart defects)
3) muscular disease of the heart
4) Heart valve failure

Finding heart disease is the most challenging part of treating it. Although there are methods for predicting heart disease, they are either too costly or ineffective to estimate the risk of heart disease in people. It has been demonstrated that early detection of heart conditions can lower mortality and overall effects. However, it is not always possible to properly monitor patients, and 24 hour access to a doctor is not feasible due to the additional intelligence, time, and skill required. In the current context, where there is a large amount of data, we can use a variety of machine learning techniques to analyse the data for hidden patterns. To diagnose illnesses, hidden patterns in medical data can be used.

A. Task Identification

A common cardiac disorder called coronary artery disease affects the main blood arteries that nourish the heart muscle. Coronary artery disease is typically brought on by cholesterol buildup (plaques) in the heart arteries. Atherosclerosis is the term for the accumulation of these plaques (ath-ur-o-skluh-ROE-sis). Reduced blood flow to the heart and other body organs is a result of atherosclerosis. It may result in a heart attack, angina, or a stroke.

Men and women may experience different symptoms of coronary artery disease. For instance, chest pain is more common among men. Along with chest tightness, women are more prone to experience additional symptoms like breathlessness, nausea, and excessive weariness.

Symptoms of the coronary artery disease can include:

1) Chest pain, tightness, pressure, and discomfort are some of the signs of coronary artery disease (angina).
2) Breathing difficulties
3) Neck, jaw, throat, upper abdominal, or back pain.
4) If the blood arteries in the legs or arms are restricted, you may experience pain, numbness, weakness, or coldness there.

IV.PROPOSED SYSTEM

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The relevant data is then pre-processed into the format needed. After that, the data is split into training and testing data. The formulas are used and the training data are used to train the model. By testing the system with test data, the correctness of the system is determined. The modules listed below are used to implement this system.

1) Collection of Dataset
2) Selection of attributes
3) Data Pre-Processing
4) Balancing of Data
5) Disease Prediction

A. Collection of Dataset

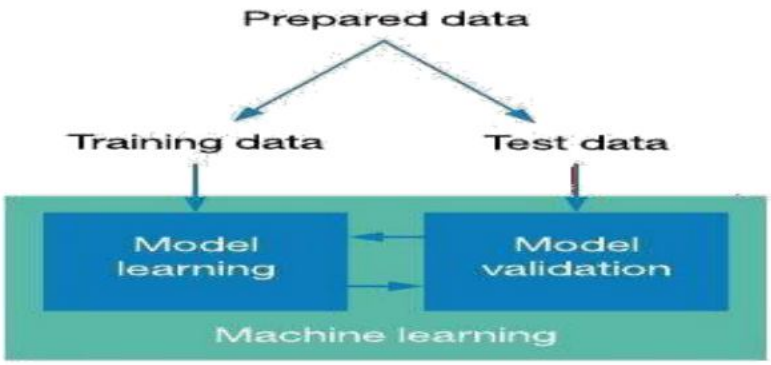


Fig1: Collection of Dataset

For the foundation of our heart disease prediction system, we first gather a dataset. We divided the dataset into training and testing data after it was collected. The learning of the prediction model is done using the training dataset, and the evaluation of the prediction model is done using the testing dataset. 30% of the data are utilised for testing in this project, while 70% are used for training. Heart Disease UCI is the dataset used for this study. It has 76 attributes, 14 of which are utilised by the system.

B. Selection of Attributes

Fig2: Selection of Attributes

C. Data Pre-Processing

The pre-processing of data is a critical stage in the development of a machine learning model.Data that isn't initially clean or in the model's required format can lead to inaccurate results. Pre-processing involves transforming data into the format we need. It is used to handle the dataset's noise, duplication, and missing values. Activities like importing datasets, partitioning datasets, attribute scaling, etc. are all part of data pre-processing. Pre-processing the data is necessary to increase the model's accuracy.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Fig3: Data Pre-Processing

D. Balancing of Data

There are two techniques to balance unbalanced datasets. They are both under- and over-sampling.

1) Under Sampling: Dataset balance in Under Sampling is achieved by reducing the size of the large class. When there is enough data, this process is taken into account.

2) Excessive sampling: Dataset balance in over sampling is achieved by enlarging the small samples. When there is not enough data, this process is taken into account.

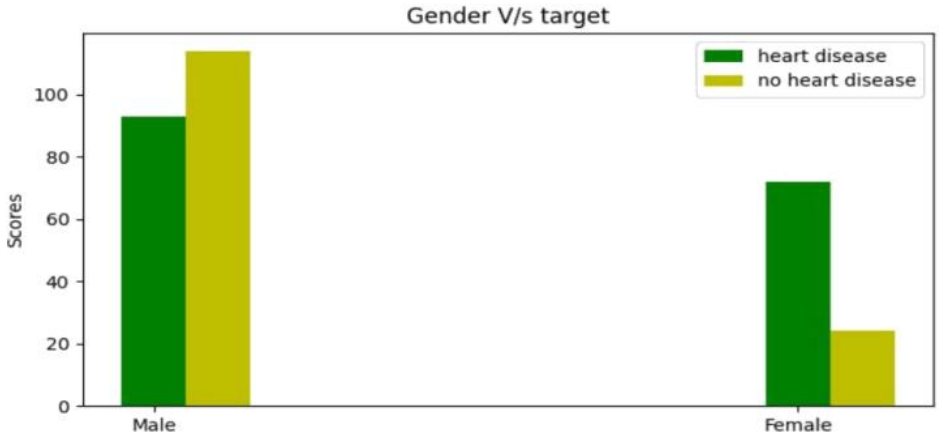


Fig4: Balancing of Data

E. Prediction of Disease

For classification, a variety of machine learning algorithms are employed, including SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Ada-boost, and XG-boost. Algorithms are compared, and the one that predicts heart disease with the best degree of accuracy is chosen.

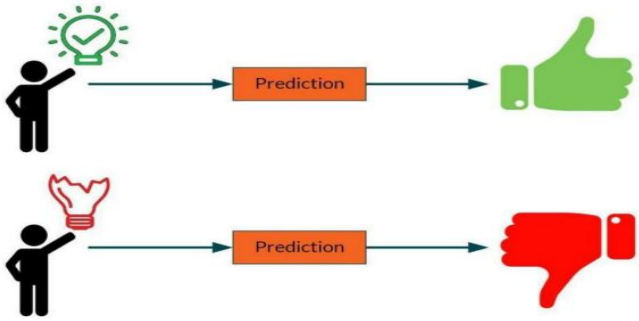


Fig5: Prediction of Disease

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

V. SYSTEM ARCHITECTURE

The system architecture provides a high-level understanding of how the system functions.
The following is a description of how this system functions:
Dataset collection is the act of gathering information containing patient specifics. selection of attributes process chooses the beneficial characteristics for heart disease prediction. The available data resources are located, then further chosen, cleansed, and transformed into the required form. To accurately forecast cardiac disease, various classification approaches will be applied to pre-processed data. The accuracy of several classifiers is compared using the accuracy measure.

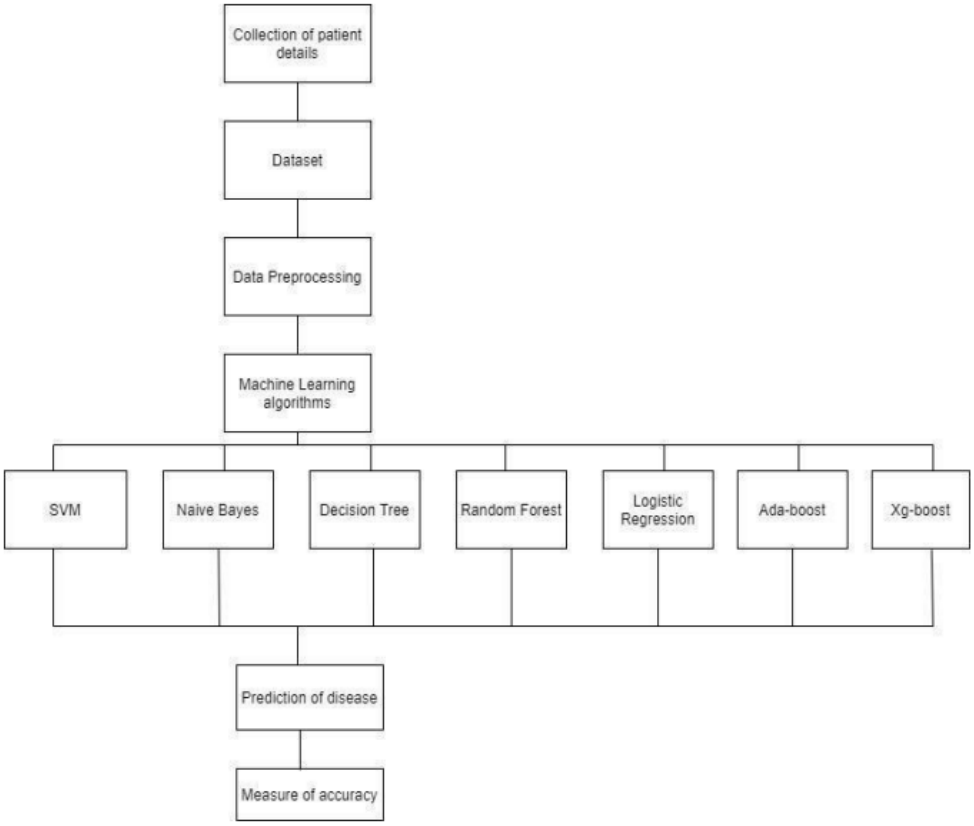


Fig6: System Architecture

A. Support Vector Machines

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyper-plane. The approach is referred described as a "support vector machine" because of these extreme circumstances.

The following are crucial SVM concepts:

1) *Support Vectors:* Support vectors are the data points that are closest to the hyper-plane. These data points will be used to define the separating line.
2) *Hyper Plane:* The hyper plane is a decision plane or space that is divided between a group of objects with various classes, as seen in the above diagram.
3) *Margin:* Margin is the distance between two lines on the nearest data points for various classes. The perpendicular distance between the line and the support vectors can be used to compute it. A large margin is viewed favourably, whereas a small margin is unfavourably.

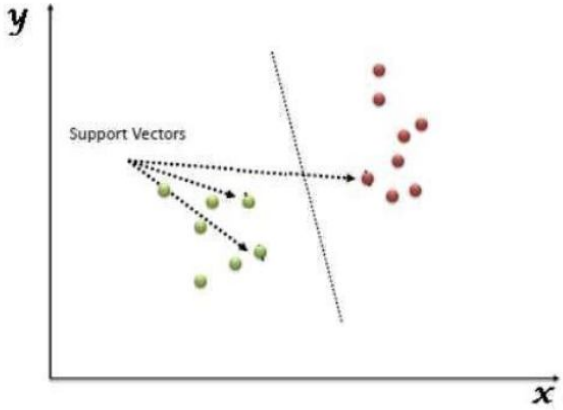


Fig7: Support Vector Machine

B. Decision Tree Algorithm

Although it may be used to solve classification and regression problems, Decision Tree is a Supervised learning technique that is more often favoured for classification problems. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. Because it is comparable to a decision tree, it has that name. It begins with the root node and expands on additional branches to create a structure like to a tree. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to construct a tree. A decision tree only poses a question and divides the tree into subtrees depending on the response (Yes/No). The supervised machine learning algorithm family includes the decision tree algorithm. It can be applied to regression problems as well as classification problems.

C. Random Forest Algorithm

A supervised learning algorithm is Random Forest. It is a development of machine learning classifiers that incorporates bagging to boost Decision Tree performance. Trees are dependent on a random vector, and it combines tree predictors which was randomly sampled. All trees are distributed in the same way. Instead of splitting nodes based on variables, Random Forests uses the best among a prediction subset that is randomly selected from the node itself. The worst case of learning with Random Forests has a temporal complexity of O(M(dnlogn)), where M is the number of growing trees, n is the number of occurrences, and d is the data dimension. Both classification and regression can be done with it. Additionally, it is the most user-friendly and adaptable algorithm. Trees make up a forest. The more trees there the greater strength a forest possesses. On randomly chosen data samples, Random Forests build Decision Trees, obtain predictions from each tree, and then vote on the best answer. Additionally, it offers a fairly accurate indicator of the feature's relevance.

Applications for Random Forests include feature selection, picture classification, and recommendation engines. It can be used to categorise dependable loan candidates, spot fraud, and forecast sickness. The Boruta algorithm, which chooses significant features in a dataset, is built around it. Popular machine learning algorithm Random Forest is a member of the approach for supervised learning. Both classification and regression can be done with issues with ML. Its foundation is the idea of ensemble learning, which is a method of utilising many classifiers in combination to solve a challenging problem and enhance effectiveness of the model. Like the name implies, "The classifier Random Forest includes a In order to increase the dataset's predicted accuracy, a number of decision trees are applied to different subsets of the given information and the average is taken "Rather of relying just on one decision tree, The random forest uses the majority votes of the participants to determine the prediction from each tree, these predictions forecasts the outcome.

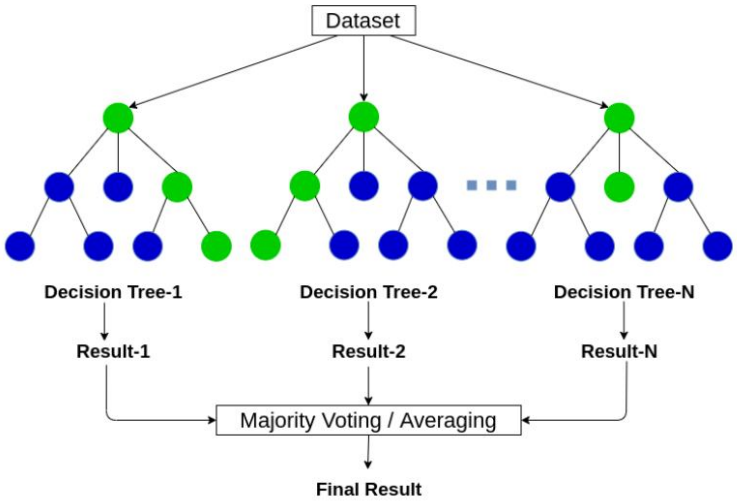


Fig8: Random Forest Algorithm

The algorithm has four steps to complete:

1) Choose arbitrary samples from a dataset.
2) Create a Decision Tree for each sample and determine the outcome of each one's prediction.
3) Cast a vote for each expected outcome.
4) Assign the title of "final prediction" to the outcome with the most votes.

D. Naïve Bayes

The Naive Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. One of the most straightforward and efficient classification algorithms is the Naive Bayes Classifier, which aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Spam filtration, Sentimental analysis, and article classification are some examples of Naive Bayes algorithms that are often used.

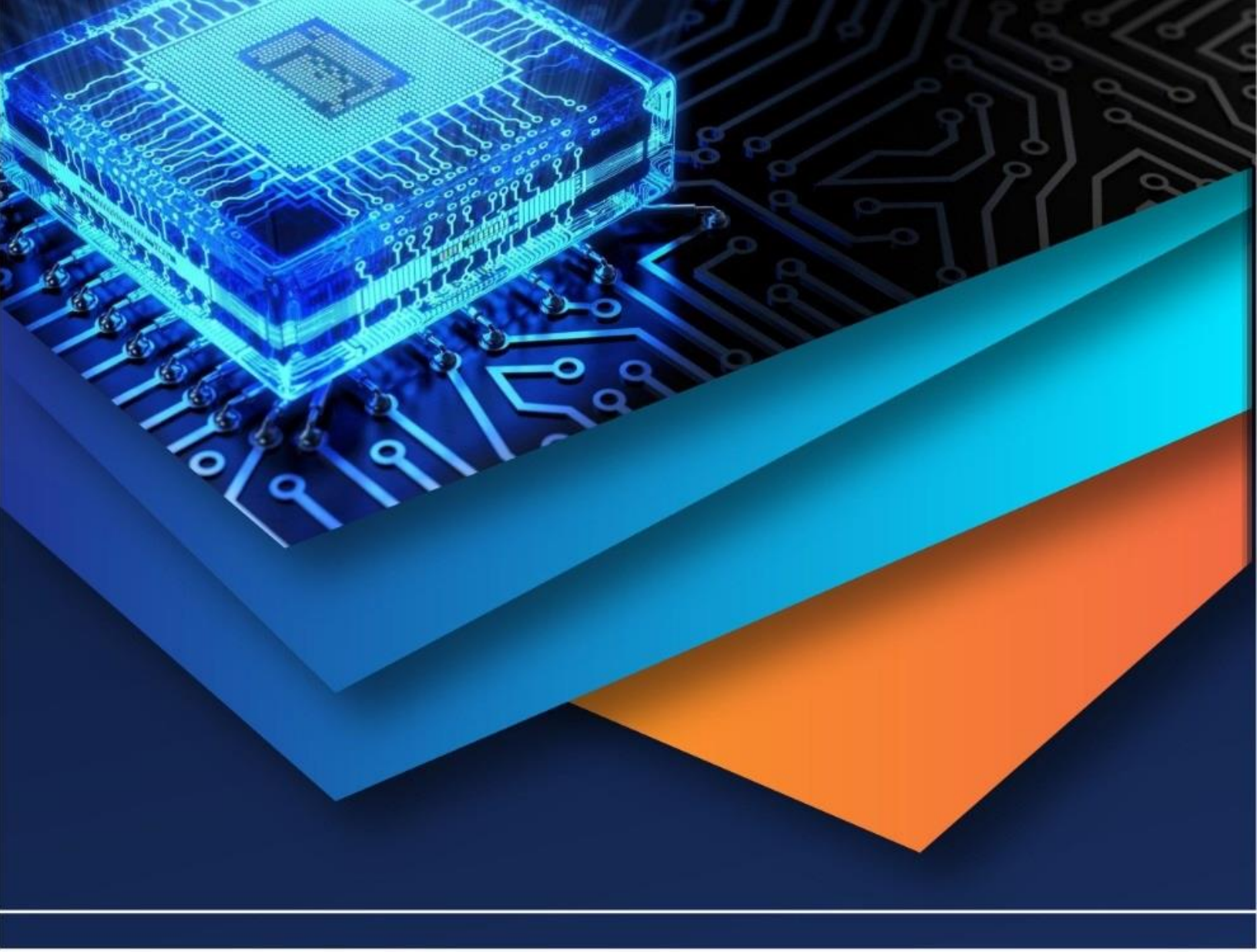
VI. CONCLUSIONS

The Application of promising technology, such as machine learning, to the first prediction of heart problems would have a significant social impact because heart diseases are a leading cause of death in India and around the world. The prognosis of cardiac disease at an early stage can help in making choices on lifestyle modifications for high-risk individuals can lower the problems and represent a significant advancement in medicine. Every year, there are more patients diagnosed with heart disease. This calls for an early diagnosis and course of action. The medical community as well as patients may benefit greatly from this use of appropriate technology support. SVM, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting are just a few of the seven machine learning algorithms employed in this study to evaluate performance. This project model is currently a good strategy that, if we include all the necessary algorithms and datasets as input, can produce good performance. In our upcoming work, we'll concentrate on how to prepare the dataset to fit the upcoming machine learning model. By making certain dynamic adjustments to our prediction system in accordance with the needs of the user, we will attempt to increase the speed and accuracy of our model.

REFERENCES

[1] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). “HDPS: Heart disease prediction system”. In 2011 Computing in Cardiology (pp. 557- 60). IEEE.
[2] Kuldeep Vayadande, Rohan Golawar, Sarwesh Khairnar, Arnav Dhiwar, Sarthak Wakchoure, Sumit Bhoite, Darpan Khadke, “Heart Disease Prediction using Machine learning and Deep learning algorithms”, International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), ©2022 IEEE

[3] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar "Prediction of Heart Disease Using Machine Learning", Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), DOI:10.1109/ICECA.2018.8474922, PP: 1275-1278 ©2018 IEEE
[4] Rahul Katarya, Polipireddy Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning", International Conference on Electronics and Sustainable Communication Systems (ICESC), DOI:10.1109/ICESC48915.2020.9155586, PP: 302- 305. ©2020 IEEE
[5] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, "Heart Disease Prediction using Machine Learning Techniques", 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), DOI:10.1109/ICACCCN51052.2020.9362842, PP: 177-181. ©2020 IEEE
[6] Alperen ERDOĞAN, Selda GÜNEY, "Heart Disease Prediction by Using Machine Learning Algorithms", 28th Signal Processing and Communications Applications Conference (SIU), DOI: 10.1109/SIU49456.2020.9302468, ©2020 IEEE
[7] Narendra Mohan, Vinod Jain, Gauranshi Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms", 5th International Conference on Information Systems and Computer Networks (ISCON), ©2021 IEEE.
[8] Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01
[9] J Prerana T H M, Shivaprakash N C et al "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS",Vol 3, PP: 90-99 ©IJSE, 2015.
[10] N. Oliver and F. F. Mangas, "HealthGear: a real-time wearable system for monitoring and analyzing physiological signals", IEEE International Workshop on Wearable and Implantable Body Sensor Networks, Cambridge, USA, (2016).
[11] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.
[12] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.
[13] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235– 180243, 2019, doi: 10.1109/ACCESS.2019.2952107.
[14] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016, no. January 2016, pp. 381–386, 2017, doi: 10.1109/ICATCCT.2016.7912028.
[15] N. H. Farhat, "Photonit neural networks and learning mathines the role of electrontrapping materials," IEEE Expert. Syst. their Appl., vol. 7, no. 5, pp. 63–72,1992,doi:10.1109/64.163674.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY