International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

Machine Learning for Cardiovascular Disease Risk Assessment: A Systematic Review

Danish Quamar[1], Mohammad Islam[2]
[1]Department Of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India
danishkhan.rock@gmail.com
[2]Assistant Professor
Department Of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India
islam@manuu.edu.in

Abstract— Accurate diagnosis and early detection of heart disease can help save lives because it is the primary cause of mortality. If a forecast is inaccurate, patients could potentially suffer significant harm. Today, it is challenging to predict and identify heart disease. 24 hour monitoring is not practical due to the extensive equipment and time required. Heart disease treatments can be both expensive and challenging. In order to obtain the data from databases and use this information to successfully forecast cardiac illness, a variety of data mining techniques and machine learning algorithms are now accessible. We have used every technique to put the heart disease prognosis into practise. The algorithms used in SVM, NAIVE BAYER, REGRESSION, KNN, ADABOOST, DECISION TREE, and XG-BOOST And Voting Ensemble Method.

Keywords- Adaboost, XG-boost, Naive Bayes, SVM, Decision Tree, Random Forest, Logistic Regression, confusion matrix, and correlation matrix.

I. INTRODUCTION

Heart disease, also known as cardiovascular disease, is a term that refers to a range of conditions that affect the heart and blood vessels. These conditions can range from simple ones, such as arrhythmias or irregular heartbeats, to complex ones, such as heart failure or coronary artery disease. Heart disease is one of the leading causes of death worldwide and is a major public health concern. In this article, we will discuss the brief history and types of heart disease [1].

1.1. Brief history:

The history of heart disease can be traced back to ancient times. The ancient Egyptians believed that the heart was the center of the body's emotions and intelligence. They believed that the heart was the organ that weighed the soul after death. The ancient Greeks also believed that the heart was the center of the body's emotions and intelligence. They believed that the heart was the seat of the soul and the source of the body's energy [2]. In the 19th century, the French physician René Laennec invented the stethoscope, which allowed doctors to listen to the sounds of the heart and lungs. This innovation enabled doctors to diagnose heart disease for the first time. In the early 20th century, the German physician Werner Forssmann performed the first cardiac catheterization. This procedure allowed doctors to diagnose and treat heart disease by inserting a catheter into the heart [3].

1.2. Types of heart disease:

The most prevalent kind of cardiac illness is coronary artery disease (**CAD**). When the arteries that carry blood to the heart are constricted or obstructed, it happens. This may result in angina, also known as chest pain or discomfort, or a heart attack.

Heart failure: Heart failure happens when there is insufficient blood flow from the heart to the body. As a result, fluid may accumulate in the legs, lungs, and other areas of the body. Heart failure can be brought on by a number of conditions, such as CAD, hypertension, and diabetes.

Arrhythmias: Abnormal heart beats are known as arrhythmias. They may be uneven, too quick, or too slow. A number of conditions, such as CAD, excessive blood pressure, and heart failure, can lead to arrhythmias.

Valvular heart disease: When the heart's valves are damaged, valvular heart disease results. The resultant backflow of blood through the valve may result in cardiac failure. A number of things, including infections, congenital heart defects, and age, can lead to valvular heart disease.

Congenital heart disease: is a form of cardiovascular disease that manifests at birth. Many causes, including genetic factors, infections, and environmental factors, can contribute to its development. Simple flaws like a hole in the heart and complex problems like transposition of the major arteries can both be caused by congenital heart disease [4].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

1.2.1. HEALTH BENEFIT

There are several benefits of using machine learning algorithms for heart disease prediction. Some of them are:

Early Detection: Machine learning algorithms can analyze large amounts of patient data and detect patterns that might not be apparent to human physicians. This can help in the early detection of heart disease, which can improve treatment outcomes and save lives.

Personalized Treatment: Machine learning algorithms can take into account individual patient data such as age, sex, medical history, and lifestyle factors to predict the likelihood of developing heart disease. This can help in tailoring treatments to individual patients, leading to better outcomes.

Reduced Healthcare Costs: By detecting heart disease early, machine learning algorithms can help reduce healthcare costs by preventing hospitalizations, emergency room visits, and expensive treatments.

Improved Patient Outcomes: Machine learning algorithms can help in predicting the likelihood of complications in patients with heart disease. This can help physicians take proactive measures to prevent complications and improve patient outcomes.

More Efficient Healthcare Delivery: Machine learning algorithms can help in the prioritization of patients based on their risk of heart disease. This can help in more efficient healthcare delivery by ensuring that high-risk patients receive timely interventions and treatments.

In summary, machine learning algorithms can help in early detection, personalized treatment, cost reduction, improved patient outcomes, and more efficient healthcare delivery in the context of heart disease prediction [5].Heart disease is a major public health concern that affects millions of people worldwide. The history of heart disease can be traced back to ancient times, and innovations in medical technology have enabled doctors to diagnose and treat heart disease more effectively. There are several types of heart disease, including CAD, heart failure, arrhythmias, valvular heart disease, and congenital heart disease. Understanding the different types of heart disease and their causes is important for preventing and managing this condition [6].

Type of Heart Disease	Description
Coronary Artery Disease	Narrowing or blockage of the blood vessels that supply the heart muscle with oxygen and nutrients
Arrhythmia	An irregular heartbeat that may be excessively rapid, too slow, or unsteady
Heart Failure	The body requires more blood than the heart can pump.
Heart Valve Disease	Problems with the valves that regulate blood flow through the heart
Congenital Heart Disease	Abnormalities in the heart's structure or function that are present at birth
Cardiomyopathy	Diseases of the heart muscle, which can lead to heart failure
Pericardial Disease	Inflammation or other problems with the sac that surrounds the heart
Aortic Aneurysm	The largest artery in the body, the aorta, has a weak spot or a bulge in its wall.
Hypertensive Heart Disease	Heart disease caused by high blood pressure
Myocarditis	cardiac muscle inflammation that is typically brought on by a viral infection

TABLE 1 VARIOUS TYPES OF HEART DISEASE [7]

II. MACHINE LEARNING ALGORITHM

A. **(AdaBoost Adaptive Boosting)**:- A well-known machine learning algorithm called AdaBoost combines a number of weak classifiers to create a powerful classifier. It updates the weights of samples that were incorrectly classified in each iteration and trains a new classifier using the revised weights. A group of weak classifiers that have been weighted based on their individual performance make up the final model. A visualization of the decision boundary and the classifier weights for each iteration can be used to visualize AdaBoost.In conclusion, this algorithm adjusts the weights of the training examples depending on the classification mistakes made by the weak classifier currently in use, and gives examples that were misclassified a higher weight to reduce their influence on the outcome prediction. Each weak classifier's weight is defined by its classification error, and the final prediction is a weighted average of the predictions from the weak classifiers. [8].

In summary, AdaBoost adapts the weights of the training examples based on their classification errors by the current weak classifier, and assigns higher weights to misclassified examples to correct their influence on the final prediction [9].

B. SUPPORT VECTOR MACHINE

A common machine learning approach for classification and regression tasks is called Support Vector Machine (SVM). The algorithm functions by identifying the optimum decision boundary with the maximum margin of separation between several classes of data points. The margin is the separation between each class's closest data points and the boundary.

SVM attempts to discover a hyper plane that divides the data into distinct classes by representing data points as points in a

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

multidimensional space. The ideal hyperplane is one that maximizes the margin.

By utilizing kernel functions to shift the data into a higher-dimensional space where the data is more separable, SVM can handle both linear and non-linear classification tasks. Furthermore resistant to overfitting, SVM is capable of handling high-dimensional data. In order to determine the hyperplane that maximizes the margin while minimizing the classification error, the SVM algorithm can be thought of as an optimization problem.

SVM is a strong and adaptable method that may be used to solve a variety of real-world issues, including anomaly detection, text classification, and image classification.[10].

C. NAÏVE BAYES

Naive Bayes is a probabilistic algorithm used for classification tasks. It calculates the probability of a new data point belonging to a particular class based on its features and the probability of each class occurring in the training data. The algorithm assumes independence between the features, which allows it to compute the joint probability of all features given the class label as the product of their individual probabilities. The class with the highest probability is then assigned to the new data point [11].

The Naive Bayes algorithm works as follows:

- Collect and preprocess the data.
- Calculate the prior probabilities of each class label by dividing the number of data points with that label by the total number of data points.
- For each feature and each class label, calculate the probability of observing that feature given that class label, using a probability distribution (such as Gaussian or multinomial) that best fits the data.
- Given a new data point with observed features, calculate the probability of each class label using Bayes' theorem and the conditional probabilities calculated in step 3.
- Select the class label with the highest probability as the predicted class for the new data point.
- The Naive Bayes algorithm is popular due to its simplicity, efficiency, and effectiveness on a wide range of classification tasks. However, the assumption of feature independence may not hold true in all cases, and the algorithm may perform poorly when there are strong correlations between features [12].

D. GRADIENT BOOSTING CLASSIFIER

Gradient Boosting Classifier is a popular machine learning algorithm used for classification tasks. It combines multiple weak classifiers (often decision trees) and builds a strong classifier by iteratively minimizing the loss function using gradient descent. The algorithm starts with a single weak classifier and calculates the errors on the training set. It then builds a new weak classifier that tries to correct the errors made by the first classifier. This process is repeated multiple times, with each new classifier trying to correct the errors of the previous ones. The final classifier is a weighted sum of all the weak classifiers.

The gradient boosting classifier involves the following steps:

1. Initialize the ensemble with a simple model, such as a single decision tree.
2. Fit the model to the data and calculate the residuals, which are the differences between the predicted and actual class labels.
3. Train a new decision tree to predict the residuals.
4. Add the new tree to the ensemble, adjusting the weights of the previous trees to minimize the loss function.
5. Repeat steps 2-4 until the desired level of accuracy is achieved or a stopping criterion is met.

The formula for the gradient boosting classifier involves the calculation of the residual and the update of the weights. The residual is given by:

$$r_i = y_i - f(x_i) \quad (2)$$

where y_i is the actual class label of the ith data point, f(x_i) is the predicted class label based on the current ensemble of trees, and r_i is the residual.

The weights of the trees are updated using a learning rate parameter, eta, and the gradient of the loss function, L, with respect to the predicted class labels, f(x_i). The update formula is:

$$f_{m}(x_i) = f_{m-1}(x_i) + eta * h_{m}(x_i) \quad (3)$$

where f_m(x_i) is the predicted class label based on the mth ensemble of trees, h_m(x_i) is the new tree added to the ensemble, and eta is the learning rate. The update rule aims to minimize the loss function by adjusting the weights of the previous trees and adding a new tree that corrects the errors of the previous trees.

In summary, the gradient boosting classifier is an ensemble method that iteratively adds decision trees to a model, adjusting the weights of the previous trees to minimize the loss function. The formula involves the calculation of residuals and the update of the weights using the gradient of the loss function and a learning rate parameter[13].

E. RANDOM FOREST

Random forest is a supervised machine learning algorithm that creates multiple decision trees and combines their predictions to produce a final output. It improves accuracy by using random feature selection and bootstrap aggregating (bagging) techniques. Each decision tree is trained on a random subset of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

the training data and features, reducing overfitting and increasing the generalization of the model. Random forest is commonly used for classification and regression tasks and is widely applied in various fields such as finance, healthcare, and image recognition**.**

Here's how the Random Forest algorithm works:

- Randomly select 'n' observations from the dataset.
- Use these observations as the training set to build a decision tree.
- Choose a number of 'm' random features from the total 'p' features.
- Split the nodes using the best split based on the selected features.
- Repeat steps 1-4 for 'k' times to create 'k' decision trees.
- Aggregate the predictions from the 'k' decision trees by taking the majority vote (in the case of classification) or the average (in the case of regression).
- The Random Forest algorithm can be summarized by the following formula:
- $Y = f(X) + e$ (1)
- Where:
- Y is the dependent variable to be predicted
- X is the set of independent variables
- f(X) is the function that maps the independent variables to the dependent variable
- e is the error term
- The Random Forest algorithm improves the accuracy of the predicted dependent variable by reducing the error term 'e' through the aggregation of multiple decision trees. By using multiple decision trees, the Random Forest algorithm reduces the risk of overfitting and provides more robust predictions [14].

F. LOGISTIC REGRESSION

The decision tree algorithm is a machine learning technique that recursively splits the dataset into smaller subsets based on the most significant attributes, creating a tree-like structure that can be used to make predictions. Each internal node of the tree represents a decision based on an attribute, and each leaf node represents a final outcome. The algorithm uses a heuristic approach to determine the best attribute to split the data, with the goal of maximizing the information gain or minimizing the impurity in the resulting subsets. Once the tree is built, it can be used for classification or regression tasks.

A decision tree is a predictive model that maps decisions and their possible consequences using a tree-like structure. Each internal node of the tree represents a decision or a test on a feature, and each leaf node represents a prediction or an outcome.

Here is the general step for a decision tree:

Let X be a set of input features, and Y be the output variable we want to predict.

- Select the best feature f from X to split the data into two or more subsets S1, S2, ..., Sk.
- For each subset Si, repeat step 1 recursively until a stopping criterion is met, such as reaching a certain depth, or all the instances in the subset belong to the same class.
- Assign a class label y to each leaf node.
- To predict the class label y of a new instance x, traverse the decision tree by following the path that corresponds to the values of x for each feature test until reaching a leaf node, and output the class label y assigned to that node.

The selection of the best feature f is based on a splitting criterion that measures the impurity or the homogeneity of the subsets. The most common splitting criteria are entropy, information gain, and Gini index, which are defined as follows:

Entropy (E) = - sum (p * log2 (p)), where p is the proportion of instances in each class in a subset.

Information Gain (IG) = E (parent) - sum ((|S|/|parent|) * E(S)), where parent is the set of instances before the split, S is one of the subsets, and |S| and |parent| are the number of instances in S and parent, respectively.

Gini Index (G) = 1 - sum (p^2), where p is the proportion of instances in each class in a subset.

The splitting criterion that results in the highest information gain or the lowest entropy or Gini index is selected to split the data [16].

III. EVALAUTION METRICS

TABLE 2[17]

Metric	Formula	Description
Accuracy	(TP+TN)/(TP+TN+FP+FN)	The percentage of accurate predictions out of all predictions.
Precision	TP/(TP+FP)	Among all positive predictions, the percentage of accurate positive predictions.
Recall/Sensitivity/True Positive Rate	TP/(TP+FN)	The percentage of accurate positive predictions among all real positives.
Specificity/True Negative Rate	TN/(TN+FP)	True negative forecasts as a percentage of all actual negatives.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

F1 Score	2*(Precision*Recall)/(Pre cision +Recall)	The harmonic mean of precision and recall, providing a balanced measure of the two.
ROC AUC	Area under the Receiver Operating Characteristic curve	A measure of the trade-off between true positive rate and false positive rate, indicating the overall performance of a classifier.
Mean Squared Error (MSE)	$(1/n)*\Sigma(y-y')^2$	The average of the squared differences between predicted and actual values, measuring the overall performance of a regression model.
Root Mean Squared Error (RMSE)	$sqrt((1/n)*\Sigma(y-y')^2)$	The square root of MSE, providing a more interpretable measure of the average prediction error.
Mean Absolute Error (MAE)	$(1/n)*\Sigma$	y-y'
R-squared	$1 - (\Sigma(y-y')^2/\Sigma(y-avg(y))^2)$	The percentage of the dependent variable's volatility that can be explained by the independent variable or variables, with a range of 0 to 1.

In the formulas above,
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative
y = actual target value
y' = predicted target value
n = number of data points.

IV. LITERATURE REVIEW

Rahul Katarya et-al (2020) [18]. The author emphasises the value of an automated approach for early heart disease prediction. The significance of feature selection and prediction as essential elements of any automated system is emphasised in the paper. The authors provide a number of techniques, including the hybrid grid search algorithm and the random search algorithm that are helpful for feature selection. They come to the conclusion that using search algorithms for feature selection and machine learning approaches for prediction would improve outcomes in heart disease prediction. The report offers a thorough analysis of the state of automated methods for heart disease prediction today.

Sanjay Patidar et-al (2022) [19].The Random Forest Classifier and Logistic Regression Algorithm both have a high accuracy of 88.00%, but Random Forest Classifier outperforms Logistic Regression with a higher Area Under Curve (AUC) value of 0.99. On the other hand, KNN performed less effectively with an AUC value of 0.84. In conclusion, Random Forest Classifier appears to be the best performing algorithm in terms of accuracy and AUC.

Shaik Farzana et-al (2021) [20].This research study sounds like it has a solid approach to predicting heart disease using a variety of machine learning algorithms. The use of the UCI Machine Learning Repository dataset will provide a strong foundation for the analysis. The inclusion of multiple algorithms like Support Vector Machine, Random Forest, KNN, Gaussian Naïve Bayes, and Xg-Boost is also a positive, as it provides the ability to compare and improve accuracy across different models. It's great to see a focus on improving standardization and accuracy, and the goal to minimize costs and retrieval time for patients.

PRONAB GHOSH et-al (2021) [21]. The study on predicting heart disease using machine learning with the Relief feature selection algorithm is a novel and innovative approach. The study achieved a high accuracy of 99.05% with 10 features, demonstrating the effectiveness of the RFBM algorithm. The study also highlights the importance of feature selection in improving the accuracy of the model. The future goals of the research include generalizing the model to work with other feature selection algorithms and applying deep learning algorithms. Overall, this research is a step forward in improving early diagnosis of heart disease and reducing mortality rates.

Vijeta Sharma et-al (2021) [22]. The research conducted provides an analysis of different machine learning techniques used to predict the likelihood of an individual developing coronary illness based on their attributes and indications. The Cleveland dataset was used and the accuracy of four algorithms was evaluated with Random Forest giving the highest accuracy of 99% and Decision Tree giving the lowest accuracy of 85%. Increasing the number of training data could potentially result in a more accurate outcome, but this would result in slower processing times due to the increased complexity of the system. Overall, the research provides valuable insights into the use of machine learning in predicting heart disease and the results suggest that Random Forest is the most effective algorithm for this particular dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

Ms.G.Thilagavathi et-al (2021)[23] This proposed system is aiming to forecast heart disease using a voting ensemble model, which integrates decision tree, KNN and random forest algorithms. The system is expected to provide maximum accuracy with the help of UCI dataset and sensors such as pulse, ECG, glucose, and BP monitoring sensors. The use of an ensemble method is expected to result in better prediction performance. The goal of the system is to decrease the death rate through earlier detection, thus leading to a healthier society.
Kuldeep Vayadande et-al (2022)[24].In summary, the author suggest the study conducted on the Kaggle heart disease dataset showed promising results using both machine learning and deep learning algorithms. Among the machine learning algorithms, Logistic Regression, Random Forest, and XGBoost achieved the highest accuracy of 88.52%. Deep learning techniques such as Multi-layer Perceptron's and Artificial Neural Network also performed well with an accuracy of 86.89% and 85.25% respectively. However, the study concluded that deep learning techniques are not suitable for small datasets. This heart disease model provides a useful tool for tracking a patient's health risk based on their age and other attributes. Overall, the model has the potential to be a valuable resource for both doctors and patients in monitoring and understanding heart disease.
Narendra Mohan et-al (2021] [25] Author suggest that Logistic Regression (LR) algorithm was the best among the four machine learning models used to predict heart diseases with a accuracy of 90.2%. This shows that LR is a powerful model for this particular task and can be used as a baseline for comparison with future models. However, it's important to note that the accuracy alone is not always the best metric for evaluating a model's performance and other metrics such as precision, recall, F1-score, and AUC should also be considered to get a comprehensive understanding of the model's performance.
Chaimaa Boukhatem et-al (2022) [26] This work provides an insightful approach to predicting heart disease in patients using various machine learning algorithms. The data was well pre-processed and the results of each algorithm were clearly presented. The SVM algorithm with linear kernel showed the best results with high accuracy, precision, recall, and F1 Score. The paper provides a good foundation for further research in this field and has potential for wider application in other diseases. Further data analysis and testing with different algorithms could potentially improve the results even more.
Manjula P et-al (2022) [27] Author Suggest the Random Forest algorithm is a highly effective tool in collaborative learning for both regression and classification. By aggregating the outputs of multiple Decision trees, the algorithm provides a more accurate prediction of heart conditions and potential heart attacks. In today's world, the ability to quickly and accurately predict heart attacks is crucial, and the utilization of this algorithm in healthcare can potentially save lives by enabling early recognition of heart conditions. By inputting health records, patients or users can make informed decisions about seeking medical attention. The application of the Random Forest algorithm in the field of healthcare has the potential to greatly improve the early detection and management of heart conditions.
Archana Singh et-al (2020) [28] a study paper assert that a major issue for people's health is how well machine learning algorithms can anticipate heart problems. The calibre of the training and test datasets affects how well the algorithms work. The KNN method is the best one, according to an examination of algorithms using a confusion matrix. The adoption of cutting-edge machine learning techniques will continue to be prioritised in the future in order to increase the precision of heart disease forecasts and reduce the number of deaths brought on by these illnesses. These methods will assist spread awareness and enhance results for people with heart disease.
N. Komal Kumar et-al (2020) [29] Author Suggestions It is notable that a variety of machine learning classifiers, including Random Forest, Decision Tree, Logistic Regression, SVM, and KNN, are used to predict cardiovascular disease (CVD). The findings demonstrate that the suggested approach, which employed a random forest classifier, outperformed all other classifiers analysed, achieving the maximum accuracy of 85.71% with a ROC AUC score of 0.8675. The study highlights the efficiency of the random forest classifier and offers useful insights into the potential of machine learning in predicting CVD. The outcomes show machine learning's potential for use in healthcare applications and call for more study in this field.
JIAN PING LI et-al (2020) [30] The system for diagnosing heart illness using machine learning is presented in this article and uses classifiers including LR, K-NN, ANN, SVM, NB, and DT. To enhance performance and cut down on processing time, the system makes use of feature selection algorithms like Relief, MRMR, LASSO, LLBFS, and a cutting-edge algorithm FCMIM. The study demonstrates that SVM with FCMIM has an excellent accuracy of 92.37% while ANN with Relief is the best predicting system with high specificity. Exercise-induced angina and chest discomfort of the Thallium Scan kind are the key indicators for the diagnosis of heart disease. By applying feature selection techniques to enhance performance and cut down on calculation time, the work makes a significant addition to the field of heart disease detection.
RAHATARA FERDOUSI et-al (2021) [31] In this work, wearable technology in healthcare is used to introduce a novel method for early-stage risk prediction of non-communicable diseases (NCDs). The framework drastically decreases the pre-processing stage of machine learning by utilising a medical professional's verified training dataset. There are now new chances to forecast the risk of NCDs from low-level sensor data thanks to a novel dynamic test dataset creation method using

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

IoT sensor data that has been introduced. The classifier build time is greatly reduced when completely corrected training data are used, and ML classification algorithms can attain accuracy of 94% or higher. The performance of the suggested approach, using diabetes as an example of an NCD, outperforms that of the existing work. This research may be expanded upon to forecast other NCDs.

Abdul Saboor et-al (2022) [32] With an SVM classifier's accuracy of 96.72%, the suggested method for heart disease prediction utilising machine learning approaches yields encouraging results. The system's disadvantage is that it performs worse as dataset sizes grow. With larger datasets, the classifier's prediction accuracy increases, but at a certain size, it starts to suffer. The authors want to investigate if they can forecast children's heart problems more accurately in the future by using XGBoost. Future research on the prognosis of heart disease will use the results of this proposed method as a standard.

Baban.U. Rindhe et-al (2021) [33] This project explores the use of machine learning techniques in predicting heart diseases. The accuracy of cardiovascular risk prediction can be significantly improved using machine learning algorithms, which can benefit patients by allowing early detection and preventive treatment. The performance of various algorithms is compared, with each showing strong performance in some cases but poor performance in others. Overall, there is significant potential for the use of machine learning in predicting heart-related diseases. Support Vector Classifier: 84.0 % which is relatively high accuracy.

Kummita Sravan Kumar Reddy et-al (2022) [34] This research paper presented a comparative study of different machine learning algorithms for heart disease prediction, including Support Vector Machine (SVM), K-nearest Neighbour (KNN), Convolutional Neural Network (CNN), and Decision Tree. The results showed that SVM had the highest accuracy of 92% among the algorithms compared. In contrast, the accuracy for KNN was 67% and CNN was 58%. The study also highlighted the speed of prediction for each algorithm and concluded that SVM was the most accurate for predicting heart disease. The institution's commitment to high-quality evidence-based research is evident in this study.

ASHIR JAVEED et-al (2019) [35] The RSA-RF learning system for heart failure prediction that was proposed in this research successfully combines the advantages of the random search and random forest algorithms to improve prediction accuracy. According to the study, the RSA-RF system surpasses 11 current approaches for detecting heart failure and other well-known machine learning models, leading to an increase in prediction accuracy of 3.3%. By limiting the amount of features, the system also lessens the complexity in terms of time. This cutting-edge learning system has the potential to enhance the accuracy of heart failure identification and support doctors' decision-making.

Victor Chang et-al (2022) [36] In this study, a machine learning method for predicting the occurrence of heart failure in a medical database is presented. The accuracy of four linear models (SVM, KNN, GNB, and MNB) was 67.24% or higher. Cat Boost (87.93%) and other ensemble learning models outperformed LGBM, HGBC, and XGB. The application of this strategy can enhance disease prediction for any ailment, not just heart failure. The investigation may be improved in the future to forecast patient survival.

Mrs. Archana Kadam et-al (2022)[37] This paper studied various machine learning algorithms for predicting cardiovascular diseases and concluded that the Random Forest classifier performed best with an accuracy of 90.16%. Other algorithms such as logistic regression, naive Bayes, and decision trees had lower accuracy levels. The integration of the Random Forest classifier with a web application provides an easy to use medium for disease prediction and analysis. The research highlights the potential for further improvement in the accuracy and reliability of cardiovascular disease prediction using machine learning techniques.

V. FINDINGS

- **Lack of comparison with other existing methods**: The paper mentions that the majority voting ensemble method is able to predict heart disease, but it doesn't compare the results with other existing methods. A comparison would provide a better understanding of how this method stands in relation to others and its strengths and weaknesses.
- **Limited evaluation metric**: The accuracy of 90% is mentioned as the evaluation metric, but other metrics such as precision, recall, F1-score, and AUC (Area under the ROC Curve) should also be considered to get a more comprehensive view of the performance of the model.
- **Dataset information**: The information about the dataset used to train the model is not provided in detail. It is important to know the size of the dataset, the distribution of classes, and any potential biases in the data to understand the limitations and limitations of the model.
- **Model explanation**: The paper does not provide a detailed explanation of the machine learning models used in the majority voting ensemble method, which would help in understanding the reasoning behind the choice of these models and their contribution to the final result.

VI. RESEARCH METHODOLOGY

The proposed methodology for heart disease prediction using machine learning is as follows:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

Data Collection: Compile a dataset of patient records including information on things like age, gender, blood pressure, cholesterol, blood sugar, family history of heart disease, etc.

Data Preprocessing: Remove any invalid or missing values to clean up the data. To ensure that all features are on the same scale, normalise or scale the data.

Feature Selection: Choose the most crucial features that have the biggest influence on the prediction of heart disease using feature selection approaches. Techniques like correlation analysis, principal component analysis, or other feature selection algorithms can be used for this..

Split Data: Create training, validation, and testing sets from the data. The validation set will be used to adjust hyperparameters and improve the model, the training set to train the model, and the testing set to assess the model's effectiveness..

Model Selection: Choose a machine learning model that can be used to foretell cardiac disease. Support vector machines, decision trees, random forests, and logistic regression are some popular models for predicting cardiac disease.

- **Model Training**: On the dataset set, train the selected model. Use methods like grid search and cross-validation to improve hyperparameters and avoid overfitting.
- **Model Evaluation**: Use a variety of assessment metrics, such as accuracy, precision, recall, F1 score, and ROC-AUC, to assess the model's performance on the testing set..
- **Model Deployment:** Deploy the model for use in the actual world after it has been trained and evaluated. This can entail developing a website or mobile application or incorporating the concept into an already-existing healthcare system.
- **Continuous Improvement**: Monitor the performance of the model in real-world settings and continuously improve it by incorporating new data, updating hyperparameters, and refining the model architecture.

Overall, this methodology involves collecting and preprocessing data, selecting features, choosing a suitable machine learning model, training and evaluating the model, deploying it into production, and continuously improving it over time.

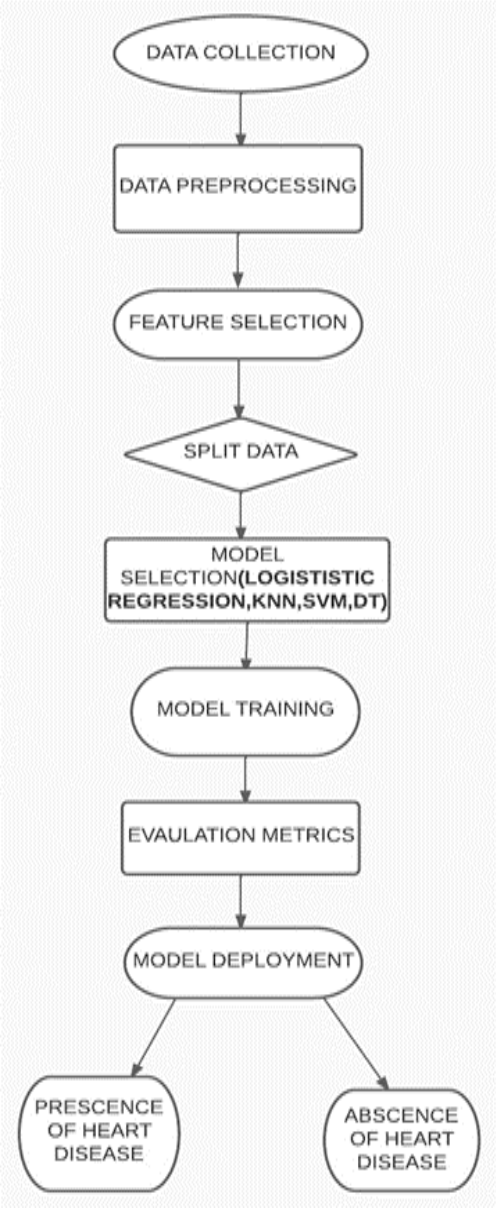

FIG 1:PROPOSED METHODOLOGY OF HEART DISEASE

DATA

The dataset is sourced from UCI and comprises five datasets with a total of 918 observations. The number of unique observations in the dataset is 918.Overall, this dataset seems useful for developing a model to identify individuals susceptible to heart disease, and can potentially aid in the development of better preventative measures and treatments for CVDs.

The data consists of 918 observations with 11 attributes.

This model for identifying individuals at risk of heart disease was created using data from a trustworthy source, specifically the UCI repository and kaggle. The data was drawn from five different datasets, all of which were carefully selected for their reliability and accuracy. By analyzing this data, the model is able to accurately identify those who may be susceptible to heart

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

disease, allowing for earlier intervention and potentially life-saving treatment [38].

- The Cleveland dataset comprises 303 observations.
- With 294 observations, the Hungarian dataset is a close second.
- The Stalag Data Set includes 270 observations.
- Switzerland's dataset has 123 observations.
- The Long Beach VA dataset consists of 200 observations.

TABLE 3 DATA ATTRIBUTE

Attribute	Description	Values
Age	In years	Continuous
Gender	Patients Gender	M, F
Chest Pain	Type of chest pain	TA, ATA, NAP, ASY
Resting BP	In mmHg	Continuous
Cholesterol	In mg/dL	Continuous
Fasting BS	1 or 0	1: >120mg/dL, 0: <=120mg/dL
Resting ECG	Findings	Normal, ST, LVH
Max HR	In bpm	60<Max HR<202
Exercise Angina	Stimulated angina	Yes, No
Old Speak ST Curve	Depression	Continuous,numeric value

DATA VISUALIZATION

Data visualization in machine learning involves the use of charts, graphs, and other visual representations of data to help understand and communicate complex patterns and relationships in the data. The main goal of data visualization is to make it easier for humans to interpret large amounts of data quickly and accurately.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.000000	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.000000	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.000000	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.500000	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.000000	Up	0
5	39	M	NAP	120	339	0	Normal	170	N	0.000000	Up	0
6	45	F	ATA	130	237	0	Normal	170	N	0.000000	Up	0
7	54	M	ATA	110	208	0	Normal	142	N	0.000000	Up	0
8	37	M	ASY	140	207	0	Normal	130	Y	1.500000	Flat	1
9	48	F	ATA	120	284	0	Normal	120	N	0.000000	Up	0

Fig 2. Dataset attribute visualization

Some common types of data visualizations used in machine learning include scatter plots, bar charts, line charts, heat maps, and histograms. There are also specialized visualization tools and libraries available in popular programming languages like Python and R that are specifically designed for machine learning tasks, such as matplotlib, seaborn, and plotly.

Checking for outliers in a dataset can give you an idea of the presence of extreme values that are significantly different from the rest of the data, which may impact the analysis and interpretation of the results. Identifying and addressing outliers can improve the accuracy and reliability of statistical models and conclusions drawn from the data [39].

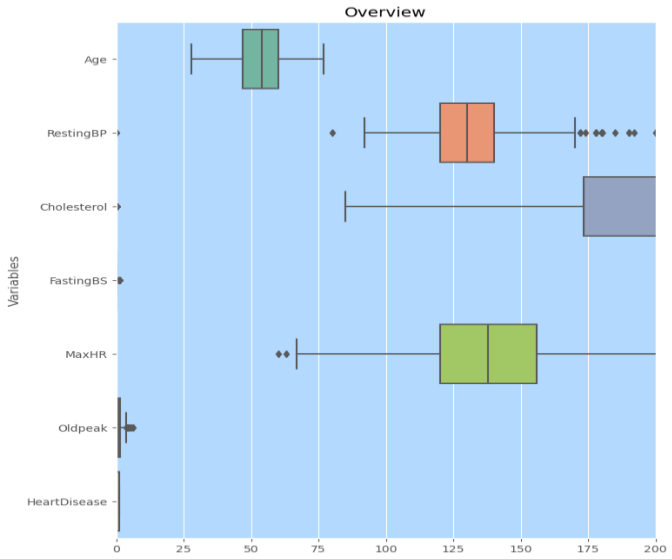

Fig 3 Outliers in data

A scatter plot matrix is a set of scatter plots organized in a grid, where each variable in a dataset is plotted against every other variable. This type of visualization can provide insights into the relationships between variables, including possible correlations and trends. The diagonal of the plot matrix typically shows the distribution of each variable, while the off-diagonal plots show the relationships between the variables [40].

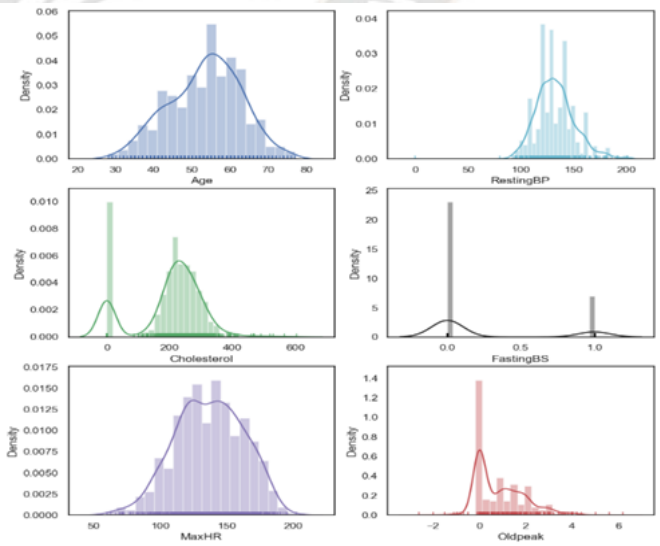

Fig 4. Scatterplot Matrix of the Dataset

VII CONCLUSION AND FUTURE SCOPE

In this study, we go over the significance of anticipating and spotting heart illness early on in order to avoid serious effects. Early disease detection can lessen the need for expensive

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

treatments and procedures. Several supervised machine learning techniques, including artificial neural networks, decision trees, random forests, support vector machines, naive Bayes, and the k-nearest neighbor algorithm, have been used to predict heart disease. Machine learning has shown promising results in making predictions and decisions in the healthcare industry. In summary, early identification and precise diagnosis of cardiac disease have improved with the application of machine learning and deep learning approaches. The suggested CNN model outperformed earlier research in terms of high accuracy, precision, recall, and F1 score for cardiac anomaly prediction. Also, the suggested CNN model performed better when combined with conventional machine learning algorithms including SVM, K-NN, DT, RF, and NB, as well as the voting ensemble approach. By enhancing early detection and treatment, the use of these cutting-edge methods may help to lower the death rate from cardiovascular disease.

CONFLICT OF INTREST

The author declare that they have no conflicts of interest with regards to conducting this research. There is no involvement of any financial institution or organization in investing any funds in this study till date. Hence, this research is completely free from any sort of conflicts.

ACKNOWLEDGEMENT

The authors express their gratitude to the management and academic staff of the Department of Computer Science and Information Technology at Maulana Azad National Urdu University in Telangana for their unwavering support and the seamless research environment they provided. We appreciate the facilities and resources that were made available to us at every stage of this research project.

REFERENCES

[1] Dr.S.Priyadarsini 1Karpagam.S, 2Kaleeswari.M, 3Kavitha.K, "HEART DISEASE PREDICTION USING MACHINE-," vol. 5, no. 8, pp. 334–337, 2020.

[2] A. Abdellatif and H. Abdellatef, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods," IEEE Access, vol. 10, no. August, pp. 79974–79985, 2022, doi: 10.1109/ACCESS.2022.3191669.

[3] A. Abdellatif, H. Abdellatef, J. Kanesan, C. O. Chow, J. H. Chuah, and H. M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods," IEEE Access, vol. 10, no. July, pp. 79974–79985, 2022, doi: 10.1109/ACCESS.2022.3191669.

[4] M. Abubaker and B. Babayigit, "Detection of Cardiovascular Diseases in ECG Images Using Machine Learning and Deep Learning Methods," IEEE Trans. Artif. Intell, vol. x, no. x, pp. 1–1, 2022, doi: 10.1109/tai.2022.3159505.

[5] A. K. Rajendran and S. C. Sethuraman, "A Survey on Yogic Posture Recognition," IEEE Access, vol. 11, no. December 2022, pp. 11183–11223, 2023, doi: 10.1109/ACCESS.2023.3240769.

[6] G. N. Ahmad and S. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection," vol. 10, 2022, doi: 10.1109/ACCESS.2022.3153047.

[7] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection," IEEE Access, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.

[8] S. Ahmed et al., "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models," J. Sensors, vol. 2022, 2022, doi: 10.1155/2022/3730303.

[9] M. Alkhodari, D. K. Islayem, F. A. Alskafi, and A. H. Khandoker, "Predicting hypertensive patients with higher risk of developing vascular events using heart rate variability and machine learning," IEEE Access, vol. 8, pp. 192727–192739, 2020, doi: 10.1109/ACCESS.2020.3033004.

[10] S. I. Ansarullah, S. Mohsin Saif, S. Abdul Basit Andrabi, S. H. Kumhar, M. M. Kirmani, and D. P. Kumar, "An Intelligent and Reliable Hyperparameter Optimization Machine Learning Model for Early Heart Disease Assessment Using Imperative Risk Attributes," J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/9882288.

[11] Ramya R. S., Parveen, M. S. ., Hiremath, S. ., Pugalia, I. ., S. H. Manjula, & Venugopal K. R. (2023). A Survey on Automatic Text Summarization and its Techniques . International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 63–71. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2478.

[12] S. Patidar, A. Jain, and A. Gupta, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Predictions," no. Iciccs, pp. 1340–1344, 2022.

[13] R. Atallah, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd Int. Conf. new Trends Comput. Sci., pp. 1–6, 2019.

[14] Anthony Thompson, Anthony Walker, Luis Pérez , Luis Gonzalez, Andrés González. Machine Learning-based Recommender Systems for Educational Resources. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/181.

[15] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "26. SABA BASHIR et-al (2021)," IEEE Access, vol. 9, pp. 130805–130822, 2021, doi: 10.1109/ACCESS.2021.3110604.

[16] S. Farzana, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques," 2020.

[17] Harsh, S. ., Singh , D., & Pathak , S. (2021). Efficient and Cost-effective Drone – NDVI system for Precision Farming. International Journal of New Practices in Management and Engineering, 10(04), 14–19. https://doi.org/10.17762/ijnpme.v10i04.126.

[18] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7112
***Article Received**: 19 March 2023 **Revised**: 22 May 2023 **Accepted**: 30 May 2023*

Support System," IEEE Access, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[19] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[20] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 302–305, 2020, doi: 10.1109/ICESC48915.2020.9155586.

[21] Thakre, B., Thakre, R., Timande, S., & Sarangpure, V. (2021). An Efficient Data Mining Based Automated Learning Model to Predict Heart Diseases. Machine Learning Applications in Engineering Education and Management, 1(2), 27–33. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/17.

[22] R. Katarya, "Predicting Heart Disease at Early Stages using Machine Learning : A Survey," no. Icesc, pp. 302–305, 2020.

[23] S. Patidar, D. Kumar, and D. Rukwal, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction," Adv. Transdiscipl. Eng., vol. 27, no. Iciccs, pp. 64–69, 2022, doi: 10.3233/ATDE220723.

[24] S. Farzana and D. Veeraiah, "Dynamic heart disease prediction using multi-machine learning techniques," Proc. 2020 Int. Conf. Comput. Commun. Secur. ICCCS 2020, 2020, doi: 10.1109/ICCCS49678.2020.9277165.

[25] P. Ghosh et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[26] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020, pp. 177–181, 2020, doi: 10.1109/ICACCCN51052.2020.9362842.

[27] G. Thilagavathi, S. Priyanka, V. Roopa, and J. S. Shri, "Heart Disease Prediction using Machine Learning Algorithms," Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022, pp. 494–501, 2022, doi: 10.1109/ICAAIC53929.2022.9793107.

[28] K. Vayadande, "Heart Disease Prediction using Machine Learning and Deep Learning Algorithms," 2022.

[29] N. Mohan, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," pp. 2021–2023, 2021.

[30] Chaimaa Boukhatem, "Heart disease prediction using machine learning," Handb. Res. Dis. Predict. Through Data Anal. Mach. Learn., pp. 373–381, 2020, doi: 10.4018/978-1-7998-2742-9.ch018.

[31] P. Manjula, U. R. Aravind, M. V Darshan, M. H. Halaswamy, and E. Hemanth, "Heart Attack Prediction Using Machine Learning Algorithms," vol. 10, no. 11, pp. 324–327, 2022.

[32] A. Singh, "Heart Disease Prediction Using Machine Learning Algorithms," pp. 452–457, 2020.

[33] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers," no. Ml, pp. 15–21, 2020.

[34] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, no. Ml, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[35] R. Ferdousi, M. A. Hossain, and A. El Saddik, "Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS," IEEE Access, vol. 9, pp. 96823–96837, 2021, doi: 10.1109/ACCESS.2021.3094063.

[36] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/1410169.

[37] Elena Petrova, Predictive Analytics for Customer Churn in Telecommunications , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[38] B. Shiva Shanta Mani and V. M. Manikandan, "Heart disease prediction using machine learning," Handb. Res. Dis. Predict. Through Data Anal. Mach. Learn., no. May 2021, pp. 373–381, 2020, doi: 10.4018/978-1-7998-2742-9.ch018.

[39] K. S. K. Reddy and K. V. Kanimozhi, "Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM," 2022 Int. Conf. Bus. Anal. Technol. Secur. ICBATS 2022, 2022, doi: 10.1109/ICBATS54253.2022.9758996.

[40] Ghazaly, N. M. . (2020). Secure Internet of Things Environment Based Blockchain Analysis. Research Journal of Computer Systems and Engineering, 1(2), 26:30. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/8.

[41] 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.

[42] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthc. Anal., vol. 2, no. September 2021, p. 100016, 2022, doi: 10.1016/j.health.2022.100016.

[43] S. P. Mrs. Archana Kadam, "A Cardiovascular Disease Prediction System Using Machine Learning," vol. 13, no. 9, pp. 7216–7225, 2023, doi: 10.47750/pnr.2022.13.S09.849.

[44] N. Louridi, S. Douzi, and B. El Ouahidi, "Machine learning-based identification of patients with a cardiovascular defect," J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00524-9.

[45] C. A. ul Hassan et al., "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," Sensors, vol. 22, no. 19, 2022, doi: 10.3390/s22197227.

[46] G. Choudhary and S. Narayan Singh, "Prediction of heart disease using machine learning algorithms," Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020, vol. 1, no. 3, pp. 197–202, 2020, doi: 10.1109/ICSTCEE49637.2020.9276802.