iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52789

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 V May 2023- Available at www.ijraset.com

Analysis of Cardiovascular Disease using Machine Learning Techniques

Aishwarya Dabir[1], Pratiksha Khedkar[2], Laxmi Panch[3], Tejal Thakare[4], Dr M A Pradhan[5]
[1, 2, 3, 4]*Student,* [5]*Associate Professor Department of Computer Engineering AISSMS College of Engineering, Kennedy Road, Pune-411001, India*

Abstract: ***Detecting cardiovascular disease early is crucial in healthcare, especially in the field of cardiology. With 12 million deaths worldwide annually, the disease is a significant health concern. Early detection can lead to lifestyle changes that reduce the risk of complications. Machine learning techniques can be used to extract useful data from large datasets and make accurate predictions, requiring less investment. Machine learning algorithms can also be used to address unbalanced datasets and feature selection, which can improve diagnostic accuracy. Using ensemble learning with a machine learning algorithm, feature selection, and biomedical test values can help classify cardiovascular disease. Multiple classifier models can also be applied to improve accuracy with an ensemble classifier.***

I. INTRODUCTION

According to the World Health Organization, cardiovascular disease causes 12 million deaths worldwide each year, with more than half of those deaths occurring in the United States and other countries. This group of diseases affects the heart or blood vessels and is responsible for 32.1% of all deaths globally, making it the leading cause of death except in Africa. Low- and middle-income countries have a particularly high mortality rate from cardiovascular disease, with over 80% of global deaths occurring in these regions. By 2030, it is estimated that over 23 million people will die from cardiovascular disease annually.

There are numerous risk factors for heart disease, including age, sex, smoking, lack of exercise, unhealthy diet, obesity, family history, high blood pressure, diabetes, high cholesterol, and psychosocial factors such as poverty and low education. Treatment typically begins with lifestyle changes and dietary interventions, with influenza vaccination potentially decreasing the risk of cardiovascular events in those with heart disease. Research into the causes, prevention, and treatment of cardiovascular disease is ongoing and remains an active field of biomedical research, with new studies published regularly.

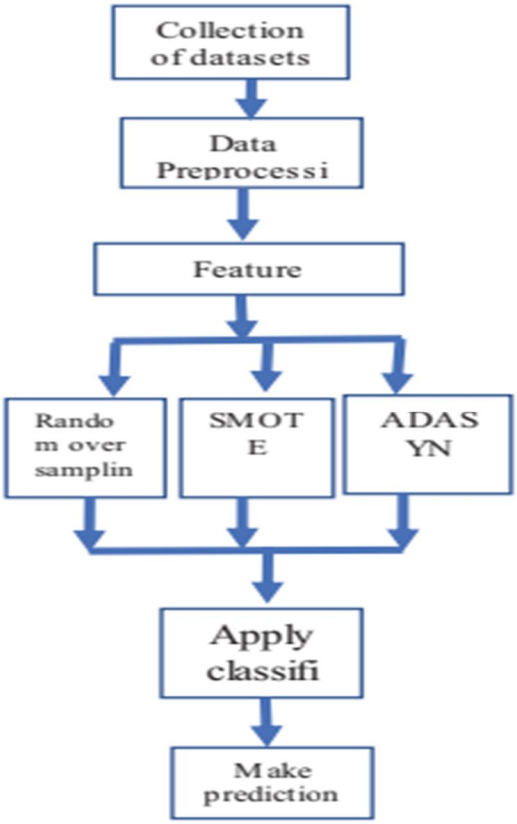
II. LITERATURE REVIEW

Aqsa Rahim; Yawar Rasheed; Farooque Azam; Muhammad Waseem Anwar; Muhammad Abdul Rahim, [1] The paper begins with an introduction to CVDs and their prevalence and impact on public health. The authors then provide a literature review of various machine learning techniques that have been used for CVD prediction, including logistic regression, decision trees, support vector machines, and artificial neural networks. The authors also discuss the importance of feature selection and data preprocessing techniques for improving the accuracy of CVD prediction models. The authors then present their proposed framework, which combines multiple machine learning algorithms and feature selection techniques to develop a more accurate and robust model for CVD prediction. The authors use a dataset of patient records and medical information, including demographic information, medical history, and laboratory test results, to train and test their model. The authors compare the performance of their proposed framework with those of other machine learning techniques using several evaluation metrics, including accuracy, sensitivity, specificity, and AUC. The results show that the proposed framework outperforms other machine learning techniques in terms of accuracy, sensitivity, and specificity.

Komal Kumar, G. Sarika Sindhu, D. Krishna Prashanthi, A. Shaeen Sulthana, [2] The paper begins with an introduction to CVD and its impact on public health. The authors then present a literature review of various machine learning algorithms that have been used for CVD prediction, including decision trees, logistic regression, support vector machines, and neural networks. The authors then describe their proposed approach, which involves using a dataset of patient records and medical information, including demographic information, medical history, and laboratory test results, to train and test several machine learning classifiers. The classifiers used in the study include decision trees, logistic regression, and support vector machines. The authors evaluate the performance of the classifiers using several evaluation metrics, including accuracy, sensitivity, specificity, and F1-score. The results show that the decision tree classifier outperforms the other classifiers in terms of accuracy, sensitivity, and F1-score.

The authors also conduct a feature selection analysis to determine the most important features for CVD prediction. The results of the analysis show that the most important features for CVD prediction include age, blood pressure, cholesterol levels, and glucose levels.

A. Lakshmanarao, A. Srisaila, T.Srinivasa Ravi Kiran. [3] The authors then present their proposed model, which combines feature selection and ensemble learning techniques. The feature selection technique is used to identify the most important features that contribute to heart disease prediction. The authors compare several feature selection techniques, including correlation-based feature selection, principal component analysis, and mutual information-based feature selection, to identify the most effective technique. The results show that the proposed model outperforms other machine learning techniques in terms of accuracy, sensitivity, specificity, and AUC. The authors also provide a detailed analysis of the factors that contribute to the accuracy of their model, including the choice of feature selection technique and the number of base classifiers used in the ensemble learning technique. Overall, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques" provides a valuable contribution to the field of heart disease prediction using machine learning techniques. The authors' use of feature selection and ensemble learning techniques to improve the accuracy and robustness of their model is innovative and shows promise for improving heart disease prediction. The paper provides a useful reference for researchers and practitioners in the field of healthcare who are interested in using machine learning techniques for heart disease prediction.



Riya Elizabeth Roy, Praveen Kulkarni, Sandeep Kumar [4] The paper covers various aspects of the problem, including feature selection, preprocessing, and model selection. The authors begin by introducing the problem of heart disease prediction and its importance in healthcare. They then provide a brief overview of machine learning and its applications in healthcare. The paper then discusses various feature selection techniques that can be used to select the most relevant features for predicting heart disease. The authors then provide a detailed discussion of various preprocessing techniques, including data cleaning, normalization, and feature scaling. The paper also covers various model selection techniques, such as logistic regression, decision trees, support vector machines, and artificial neural networks. The authors provide a comprehensive comparison of these models, including their strengths and weaknesses. The paper also discusses various evaluation metrics that can be used to evaluate the performance of the models, such as accuracy, sensitivity, specificity, and AUC. The authors provide a comprehensive comparison of these metrics and explain their significance in evaluating the performance of the models.

Yukti Sharma, Rikku Veliyambara, Rajashree Shettar[5] The authors combine two machine learning techniques, decision tree and support vector machine (SVM), to develop a more accurate and efficient classifier. The paper begins with an introduction to the problem of heart disease identification and its importance in healthcare. The authors then provide a literature review of various machine learning techniques that have been used for predicting heart disease, including decision trees, SVM, artificial neural networks, and logistic regression. The authors then present the proposed hybrid classification model, which combines decision tree and SVM. They explain the rationale behind the choice of these two techniques and how they complement each other to improve the accuracy of the classifier. The authors also provide a detailed explanation of the feature selection and preprocessing techniques used in the model. The paper presents the results of experiments conducted to evaluate the performance of the proposed hybrid classification model. The authors compare the results of their model with those of other machine learning techniques, including decision tree, SVM, and logistic regression. The results show that the hybrid classification model outperforms the other models in terms of accuracy, sensitivity, specificity, and AUC.

Rubini Pe, C A Subasini, A Vanitha Katharine[6[The paper begins with an introduction to cardiovascular disease and its impact on public health. The authors then provide a literature review of various machine learning algorithms that have been used for cardiovascular disease prediction. They explain the advantages and disadvantages of each algorithm and the factors that must be considered when selecting an appropriate algorithm. The authors then present their study, which uses a dataset of patient records to develop and evaluate the performance of the machine learning algorithms. The paper provides a detailed explanation of the feature selection and preprocessing techniques used in the study. The authors then compare the results of the different machine learning algorithms based on several evaluation metrics, including accuracy, sensitivity, specificity, and AUC. The results show that the SVM algorithm outperforms the other algorithms in terms of accuracy, sensitivity, and AUC, indicating that it is the most effective algorithm for predicting cardiovascular disease. The authors also provide a detailed analysis of the factors that contribute to the accuracy of the SVM algorithm, including feature selection and parameter tuning.

III. DATA PREPROCESSING

Steps followed for data preprocessing are:

1) Checking missing values

2) Label Encoding:

a) The data consist of categorical values which had to be treated before applying algorithm. Label encoding is a technique used to convert categorical variables into numerical values in order to be used as input for machine learning models.

b) Sklearn. preprocessing. LabelEncoder() used for label encoding as it is a convenient, fast, and reliable implementation of label encoding. Avoiding bias and improving performance of model.

3) Normalization:

The process of bringing all of a dataset's columns into the same scale sklearn.preprocessing.normalize() is utilized accomplish standardization.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

IV. CLASSIFICATION ALGORITHMS

The following are four machine learning algorithms that can be used for a comparative study:

A. Decision Tree (J48)

Weka J48 algorithm for decision tree is used for analysis. To calculate entropy the measure of impurity Gini index is used. Decision tree used divide and conquer strategy for classification. It is a tree base classification algorithm

B. Support Vector Machine (SMO)

This algorithm is used for classification and regression analysis. It creates a hyperplane to segregate n-dimensional space into classes, allowing new data to be classified into the correct category in the future**.** A function based classification algorithm. As dataset is binary the hyperplane for SVM will be a line.

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

C. K Neighbors (IBK)

This algorithm assumes that new data is similar to existing data and can be classified into a similar category. It stores all available data and classifies new data points based on similarity. The number of neighbors used is 4 for given dataset. The mathematical formula used is:

$$d=\sqrt{((x2-x1)^2+(y2-y1)^2)}$$

D. Random Forest

Random Forest is an ensemble learning method that creates multiple decision trees and aggregates their predictions to make a final prediction. One useful aspect of Random Forest is that it can provide feature importance scores for each input feature, which indicates how much each feature contributes to the model's predictions..

E. Logistic Regression

Logistic regression is a statistical method used for binary classification tasks. The logistic regression model estimates the probability of the outcome based on the input features, using a logistic function (also known as a sigmoid function). The output of the logistic function is a value between 0 and 1, which represents the probability of the event occurring. The model then makes a prediction by applying a threshold to the estimated probability.

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Dataset: For our system we have used heart disease dataset, which is collected from UCI repository. All information about the dataset is given in following table.

Table1.Total Instances

Dataset	Instance	Attributes
Heart Disease Dataset	918	12

Table2. Performance Analysis of algorithms without pre-processing dataset

Algorithm	Accuracy%	True Positive Rate %	False Positive Rate %	Error%
Random Forest	85.86	92.39	8.75	14.13
Decision Tree	82.06	86.95	14.45	17.95
Logistic Regression	79.89	85.86	16.04	20.10
KNeighbors	72.28	72.82	27.47	27.17
SVM	49.45	98.913	100	50.43

1) Accuracy: -correctly classified instances / total number of instances.
2) True Positive Rate: - TP/(TP+FN)
3) False Positive Rate: -FP/(FP+TN)
4) Error rate: - 1-Accuracy

Table3. Performance Analysis of algorithms by pre-processing dataset

Algorithm	Accuracy%	True Positive Rate %	False Positive Rate %	Error%
Random Forest	88.04	86.40	16.09	11.94

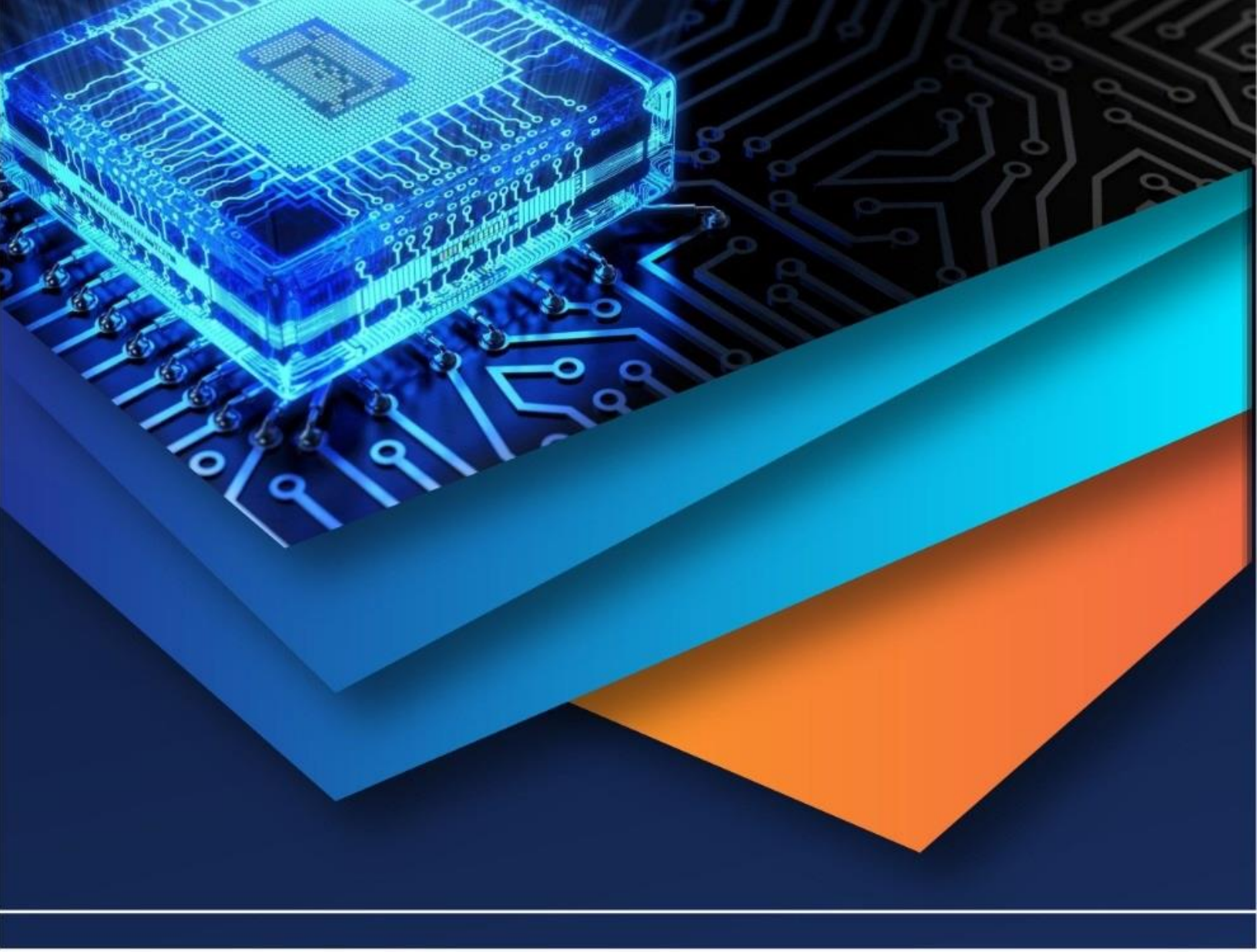
Decision Tree	80.43	69.09	28.97	19.56
Logistic Regression	86.95	83.94	18.68	13.04
KNeighbors	81.53	72.81	27.81	18.47
SVM	85.86	83.49	19.10	14.13

V. CONCLUSION

By analysis of the above algorithms performed on the dataset, we observed that the accuracy is majorly increased due to Random Forest and Logistic Regression after preprocessing data using normalization, label encoding and k-fold validation.

REFERENCES

[1] ”An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases” by Aqsa Rahim; Yawar Rasheed; Farooque Azam; Muhammad Waseem Anwar; Muhammad Abdul Rahim
[2] “Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers” by N. Komal Kumar, G. Sarika Sindhu, D. Krishna Prashanthi, A. Shaeen Sulthana.
[3] Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques by A. Lakshmanarao, A. Srisaila, T.Srinivasa Ravi Kiran.
[4] Machine Learning Techniques in Predicting Heart Disease a Survey by Riya Elizabeth Roy, Praveen Kulkarni, Sandeep Kumar
[5] Hybrid Classifier for Identification of Heart Disease by Yukti Sharma, Rikku Veliyambara, Rajashree Shettar
[6] A Cardiovascular Disease Prediction using Machine Learning Algorithms by Rubini Pe, C A Subasini, A Vanitha Katharine



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY