Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

A ROBUST PREPROCESSING AND FEATURE SELECTION FRAMEWORK IS PROPOSED TO ENHANCE HEART DISEASE PREDICTION ACCURACY

Aqib Mehmood[1], Hajar Bendaoud [2], Muhammad Ghaos Baksh UVES[2], Attiq Ullah[1], Mohsin Mahmood[3] , Mubashir Zainoor [1] , Salman Ali Khan[1]
[1]Iqra National University Peshawar
[2] Xidian University, china
[3]City University of Science & Information Technology Peshawar
Aqibmehmood@inu.edu.pk

DOI: https://doi.org/https://doi.org/10.5281/zenodo.20033315

Keywords:
Encompass Data Mining, Machine Learning, Logistic Regression, KNN, CNN, Random Forest, Prediction, UCI dataset

Article History
Received on 18 April , 2026
Accepted on 28April , 2026
Published on 05 May , 2026

Copyright @Author
Corresponding Author: *
Aqib Mehmood

Abstract
Heart disease is now the leading health issue in the world and requires proper measures to diagnose and prevent heart disease at an early stage. The paper introduces statistical and machine learning methods for forecasting heart disease by analyzing vital health indicators and lifestyle factors. To create a predictive framework, the University of California, Irvine (UCI) Heart Disease Dataset, comprising patient-specific characteristics, is used. The performance of three classification models, including the Logistic Regression, K-Nearest Neighbors (KNN), and the Random Forest, is compared in terms of their predictive performance. The research methodology can be divided into two stages: identification of the most important clinical characteristics that suggest cardiovascular risk; evaluation of the accuracy of the model on the data. The results indicate that machine learning and data mining tools can be used to diagnose and prevent cardiovascular diseases promptly.

1. Introduction

Medical dictation has been experiencing time, accuracy, and cost problems. The risks of human errors that accompany manual processes are very high, especially now that the world is faced with an epidemic of cardiovascular diseases. The condition of the situation is complicated by the complexity of the conditions, which means that the human judgment used in the determination of these conditions is not always accurate and is a threat to the health of patients. Data mining has turned out to be a potent tool for forecasting various outcomes in various fields. As the data mining techniques and the deep learning approaches are integrated, sophisticated models have been engineered to forecast some of the scenarios, such as the emergence of cardiovascular diseases. The three machine learning classification techniques—K-Nearest Neighbor, Logistic Regression, and Random Forest—have been employed by researchers working with the UCI Heart Disease dataset to create effective predictive models for identifying heart disease. These are easy but correct models that have been conditioned on 14 parameters using the UCI Dataset and can assist individuals in incorporating a healthy lifestyle and take some proactive measures that are necessary to minimize the risk of early heart disease development. Cardiovascular diseases, which impact about 1.5 million individuals annually and result in 31 percent of all deaths in the world, are a crisis health issue in society. Even more concerning is the fact that 82 percent of early deaths occur in the poor and middle-income countries, with cardiovascular diseases being the cause of 37 percent of them, contributed by late/bad predictions. Given the prevalence rate of such diseases as heart disease of the coronary, cardiomyopathy, hypertensive heart disease, and heart failure, early prediction and risk. Assessment in the process of minimizing the risk of mortality becomes extremely important. Predictive models can forecast cardiovascular disease by evaluating clinical markers and lifestyle variables that contribute to heart health risks. Patient-specific computational algorithms are deployed to calculate heart disease likelihood. This research applies four classification methods: K-Nearest Neighbor, Decision Trees, Logistic Regression, and Random Forest. The preprocessing workflow includes an initial Exploratory Data Analysis phase to examine feature attributes, followed by data normalization through standardization and outlier detection before model development. EDA techniques are first employed to assess data structure and quality, then standardization processes are implemented to resolve data anomalies. This data is further separated into training and testing sets that are used to test the models. The logistic regression classifier is employed to prognosticate the outcomes based on the pre-determined variables and ensures a high level of accuracy. Further, KNN is applicable to efficient model construction considering the closest neighbors and their distances, and thus reduces unnecessary calculations. Heat maps are implemented to visually clarify the association of different parameters in cardiovascular diseases to enhance a better perception of the interaction between them. The designations age, chest pain type (cp), steepness, and maximum heart rate attained (thalach), and the presence of angina during exercise (exang) are noted to be important predictors of heart disease, as shown in the heat. map analysis.

2. LITERATURE REVIEW:

At this point, much attention is paid to the analysis of cardiovascular disease prediction, and the application of different techniques to enhance the efficiency and accuracy of all parameters. Researchers investigated a heart disease prediction model on clinical datasets in a 2013 study that utilized decision tree

models. In their analysis, to enhance predictive performance, various machine learning and statistical methodologies were implemented, including ID3, Naive Bayes, Gain Ratio Decision Trees (DT), Adaptive Resonance Theory (ART) kernel density, bagging, and Support Vector Machines (SVM). However, achieving optimal estimation accuracy remained a challenge due to the misrepresentation of feature values inherent in Gain Ratio DT architectures. Nonetheless, it was difficult to make a more accurate estimation because of the misrepresentation of the values by Gain Ratio DTs. A later study, characterized as an investigation in the same year, came up with an estimation of heart disease through an amalgamation of SVM, DT, and logistic regression with the CHDD dataset. Although this approach involved the rule-based algorithm of DT, combining the aspects of classification, regression, and correlation, it also had drawbacks because SDL was unable to guarantee better results, being sometimes order-based. Machine learning is a new area that has been brought about by technological advancements. In 2018, a new machine learning algorithm for predicting heart diseases was introduced. The strategy utilised the MLP algorithm and delivered better efficiency and accuracy. Nevertheless, there were difficulties because MLP would have trouble in cases where the levels were too low or too high. A comparative study of various algorithms in a 2019 research led to an improved method of HRFLM on the basis of Random Forest and Linear systems.

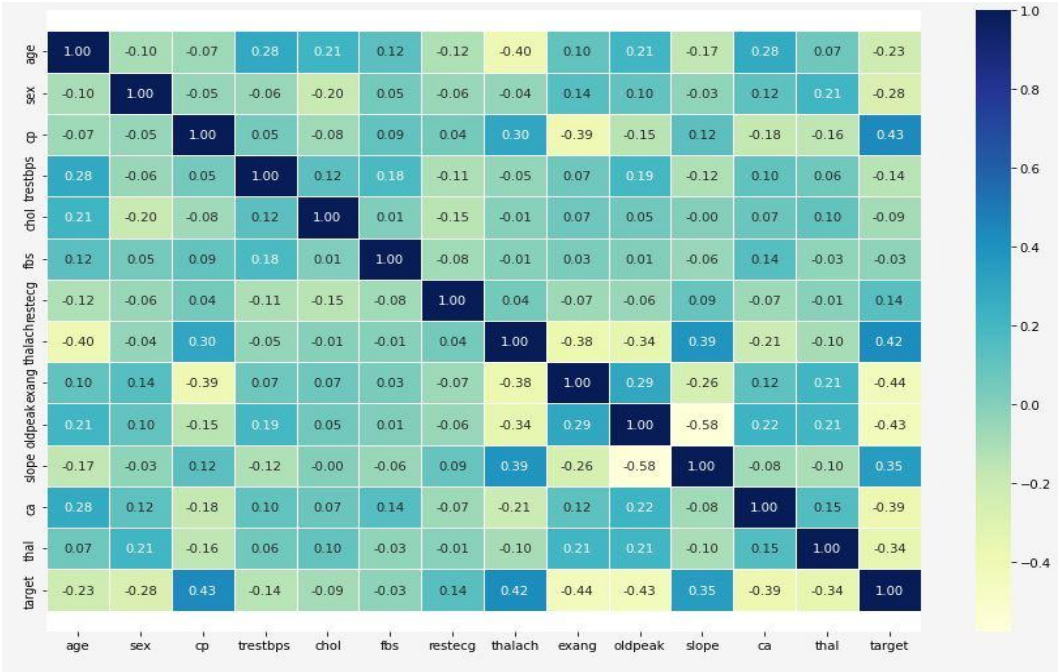

Figure 1. Heat Map

3. LITERATURE REVIEW:

At this point, much attention is paid to the analysis of cardiovascular disease prediction, and the application of different techniques to enhance the efficiency and accuracy of all parameters. Researchers investigated a heart disease prediction model on clinical datasets in a 2013 study that utilized decision tree models. In their analysis, to enhance predictive performance, various machine learning and statistical methodologies were implemented, including ID3, Naive Bayes, Gain Ratio Decision Trees (DT), Adaptive Resonance Theory (ART) kernel density,

bagging, and Support Vector Machines (SVM). However, achieving optimal estimation accuracy remained a challenge due to the misrepresentation of feature values inherent in Gain Ratio DT architectures. Nonetheless, it was difficult to make a more accurate estimation because of the misrepresentation of the values by Gain Ratio DTs. A later study, characterized as an investigation in the same year, came up with an estimation of heart disease through an amalgamation of SVM, DT, and logistic regression with the CHDD dataset. Although this approach involved the rule-based algorithm of DT, combining the aspects of classification, regression, and correlation, it also had drawbacks because SDL was unable to guarantee better results, being sometimes order-based. Machine learning is a new area that has been brought about by technological advancements. In 2018, a new machine learning algorithm for predicting heart diseases was introduced. The strategy utilised the MLP algorithm and delivered better efficiency and accuracy. Nevertheless, there were difficulties because MLP would have trouble in cases where the levels were too low or too high. A comparative study of various algorithms in a 2019 research led to an improved method of HRFLM on the basis of Random Forest and Linear systems.

The investigation was aimed at the implementation of practices based on raw evidence, though it was in theory. In one more recent study in 2019, SVM, DT, logistic regression, and Naive Bayes were separately tested on a rapid minor UCI dataset, in order to obtain a better precision than in the prior works. Nonetheless, the method enabled a sluggish model with numerous algorithms, which produced delayed outcomes. The predictive schemes heavily rely on the data utilized in the scheme. In a comparative study carried out in 2011, decision trees and Bayesian algorithms were more accurate than the others when applied to the same data using various algorithms. In 2013, the Chandigarh data set gathered was subjected to data mining and applied to the data, which proved to be efficient in predicting heart diseases. Also, in the same year, a prototype that trained nurses and physicians to predict diseases provided predictive features and clarification.

A structure based on the Adaptive Neuro-Fuzzy Inference Scheme was introduced in 2014 and is more precise compared to other methods, and is suggested to physicians as a prediction instrument. In 2011, another study also suggested a web-based, convenient, and precise prediction framework of heart disease on the UCI machine learning data, where a weighted related classifier algorithm was applied. Nonetheless, the research recommended using it in local data sets to have a practical application in light of different symptoms. Finally, in a cardiac prediction analysis study, the results were compared to different algorithms, and neural networks were the most precise, especially when applied together with a genetic algorithm.

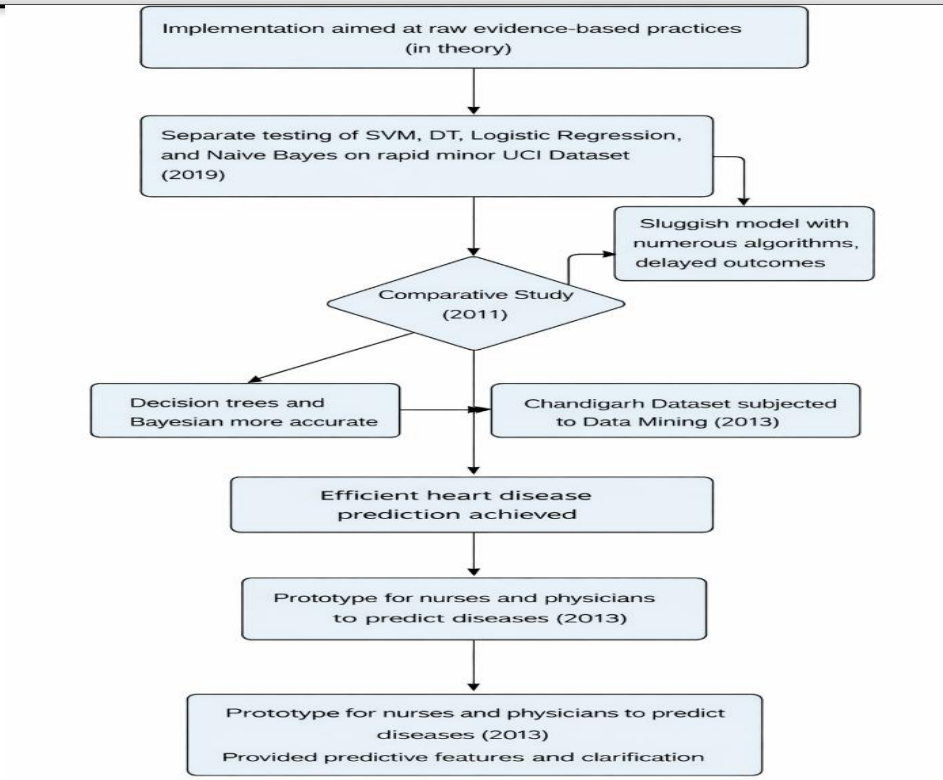

Figure 2. Model Flowchart

Table 1: Clinical Features Used for Heart Disease Prediction

Clinical Feature	Description
Num	Diagnosis of heart disease (target variable: 0 = no disease, 1 = disease present)
Age	Age of the patient in years
Sex	Gender of the patient (1 = male, 0 = female)
Cp	Chest pain type (e.g., typical angina, atypical angina, non-anginal pain, asymptomatic)
Trestbps	Resting blood pressure (in mm Hg)
Chol	Serum cholesterol level (mg/dl)
FBS	Fasting blood sugar (>120 mg/dl: 1 = true, 0 = false)
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina (1 = yes, 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
Ca	Number of major vessels (0–3) colored by fluoroscopy
Thal	Thalassemia status (3 = normal; 6 = fixed defect; 7 = reversible defect)

4. APPROACH:

The proposed methodology is based on the UCI Heart Disease Dataset [5] that is publicly available. This data is subjected to pre-processing phase which includes data cleaning and normalization to achieve reliability and consistency. After this, techniques of data visualization are used to extract the underlying trends and to determine dependencies between attributes. Data is then divided into training and testing data to aid in the assessment of the model. The training phase is where machine learning algorithms are applied to the training data, the models are then tested on the test data. Lastly, the classification algorithms are compared and their predictive performance evaluated. Fig. 2 represents the flowchart of the suggested framework, which shows the steps of the data preparation process, training the model, testing it, and evaluating the results in a sequence.

4.1. UCI Dataset:

The UCI dataset registry is an open-source library on which the analysis is based and contains various databases regarding diseases. These are open-source databases that are of scholarly use. In particular, this model uses the heart disease UCI dataset. It is a multivariate dataset, a type of existence category dataset and has 303 instances and 75 properties. In a database like this, we have valuable assets of knowledge and non-counting ones. Thus, the selection of the relevant data is made as a part of the preprocessing stage, and data cleaning operations are carried out to remove the null values.

3.2 Pre-Processing:

The pre-processing step aims at deriving meaningful and valid information of the heart disease data. Raw medical data is frequently incomplete, inconsistent or noisy, thus pre-processing is necessary to guarantee that the dataset can be successfully used further in analysis. The University of California, Irvine (UCI) Heart Disease Dataset is used in this research and initially has 75 attributes. To better represent the health attributes of patients, 14 clinically relevant attributes are selected using a systematic feature selection method. Such features are blood pressure, gender, heart rate, and indicators of the chest, among others. The values of the attributes are all reduced to a normalized form and made numerical to make uniformity and easier computational analysis. This transformation ensures that consistency is ensured in the entire dataset and it is also easier to apply machine learning algorithms in the predictive modeling step.

3.3 Data Clean:

The quality of the data is of the utmost importance in this study, and the accuracy of the data is carefully regarded to be accurate. Data cleaning has been done in order to enhance the quality of our data. This is a crucial process since it involves the removal of any undesirable or meaningless attributes in the data set, therefore, improving the accuracy and accuracy of the data set. Specifically, at this step, the null (NaN) values are eliminated in the dataset, which can significantly decrease the utility of algorithms. In addition, the data cleaning phase involves normalization of the data in order to ensure consistency and eliminate any element of ambiguity following the cleaning.

3.4 Visualization

The information, represented in Table 1, might not be readable and comprehensible. In order to answer this, the graphical visualization technique is being employed, and the Fig.3 below indicates it. These are the visualizations that provide more insight into the trends of the data. The association of the many attributes of the data can be illustrated graphically, and complicated patterns are easy to comprehend. Such visualizations present a better insight into the trends of data. Graphical representation can be used to demonstrate the connection of the various attributes of

the data, and complex patterns can be easily understood. Bar charts and scatter plots are used in this analysis to present the clean data acquired in the process of pre-processing. These visualizations provide information about the behaviors of different data attributes, and it is easy to interpret the correlation between the data attributes. As highlighted above, such visualization is crucial in exploring the data, and one is able to understand the data further. Fig.3 is an example of such a visualization.

Figure 3. Exploratory Data Analysis Based On Age

Where various parameters of the dataset are plotted against the age of patients. Utilizing the Pairplot method from the Seaborn library in Python, it becomes evident that cardiac disorders, specifically cardiovascular disorders, are closely correlated with patient age. This graphical representation aids in identifying significant patterns and trends within the dataset.

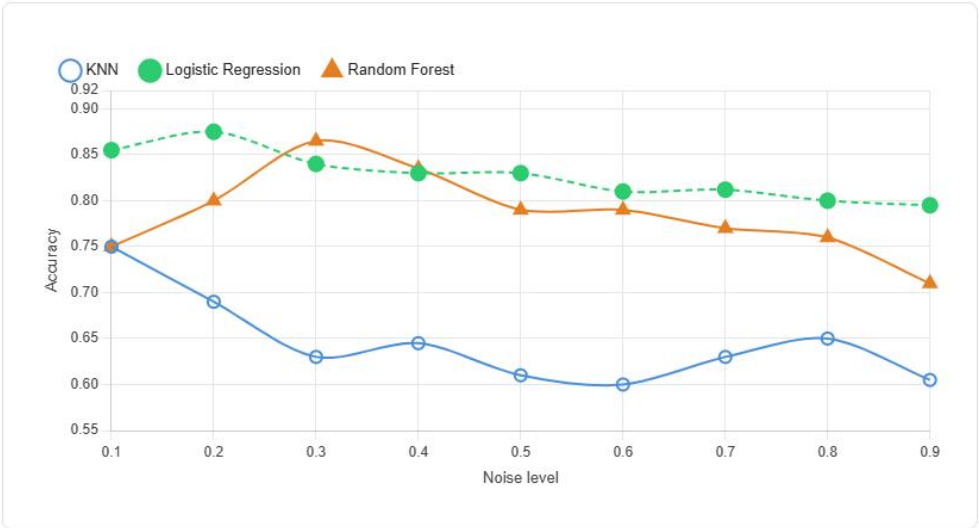

Figure 4: Testing/Training Accuracy Ratios

3.5 Training & Testing:

The idea of machine learning models is based on an input-output model wherein the training data is used to feed the algorithm allowing it to learn patterns which are then tested in terms of performance. In this study, the heart disease dataset was divided into the training and testing sets. The training part helps to create and fine tune predictive models using different machine learning algorithms in order to determine the best algorithm whereas the testing part validates the trained predictive models and measures the accuracy of prediction. Randomized data inputs are used in order to ensure reliability and stability of the evaluation. There are different paradigms of data-driven modeling that include machine learning supervised, semi-supervised, unsupervised, and reinforcement learning. Each paradigm has a variety of approaches to certain problems. In this study, three supervised classification algorithms, including: Random Forest, Logistic Regression and K-Nearest Neighbor (KNN) are used to predict the occurrence of heart diseases [13]. The selection of these algorithms was based on the fact that they are effective in classification problems and that they can work with varying data patterns.

3.6 **Random Forest:**

Random forest classifier is a machine learning algorithm that is a decision tree. It entails the development of more than one decision tree depending on the various attributes of the data. The average of the predicted performance of these trees defines the performance of the algorithm. Random forest constructs multiple decision trees and then employs them to identify the most favorable result. It uses a bootstrapping, aggregating, or bagging method of tree learning. In bootstrap aggregating, a dataset X =[x1, x2, x3, x4, and so on] is repeated on b= 1 to B; almost all the member trees are then used to predict x'. Besides, the standard deviation is also computed to ascertain the vagueness of prediction using such decision trees.

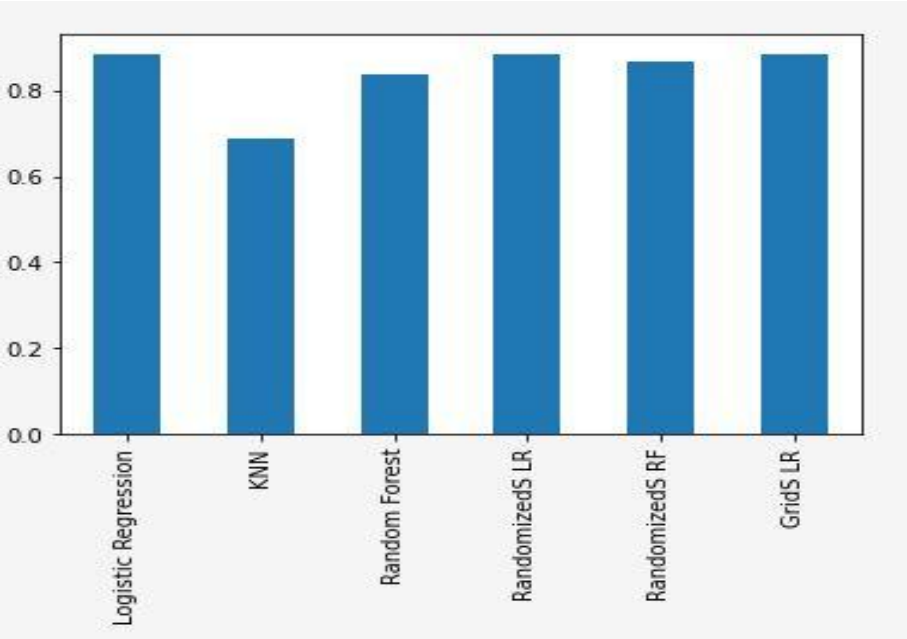

Figure 5. Hyper parameter tuning with Grid Search CV

3.7 Logistic regression:

An example of a statistical method, e.g., logistic regression (LR), is which deals with two potential outcomes of the variable under study. This implies that the interaction between the input and the output is linear. It involves estimation of the probability of the target variable taking the data, and the dependent variable is a binary outcome variable that is binary. LR is normally used in

predicting and estimating the chances of success. In LR, the formula is modeled to fit the required data format, and a simple equation will be written as follows: Y = b 0 + b 1x 1 + b 2x 2 +... + b nx n. The regression coefficients estimation is carried out with the help of the Maximum Likelihood Ratio (MLR), which will help to define the statistical significance of the dependent variables with respect to the independent variables. MLR considers the impact of independent variables, and the likelihood (p) of each case is subsequently provided by the Odds Ratio. It is a measure that is applied to determine the relationship and strength between two given events, P and Y.

.8 Nearest Neighbour:

K-Nearest Neighbors (KNN) is an algorithm that retrieves the data from the dataset and predicts the closest output. It is an effective method since it boasts of high predictive accuracy. KNN is particularly well applied to pattern recognition, such as in numerous cases in the heart disease dataset. KNN operates based on logic and knowledge retrieval based on the Euclidean distance between the samples, based on the following function of d(xi, xj) and most of the neighbors, based on K. Mathematically, it can be expressed as follows.

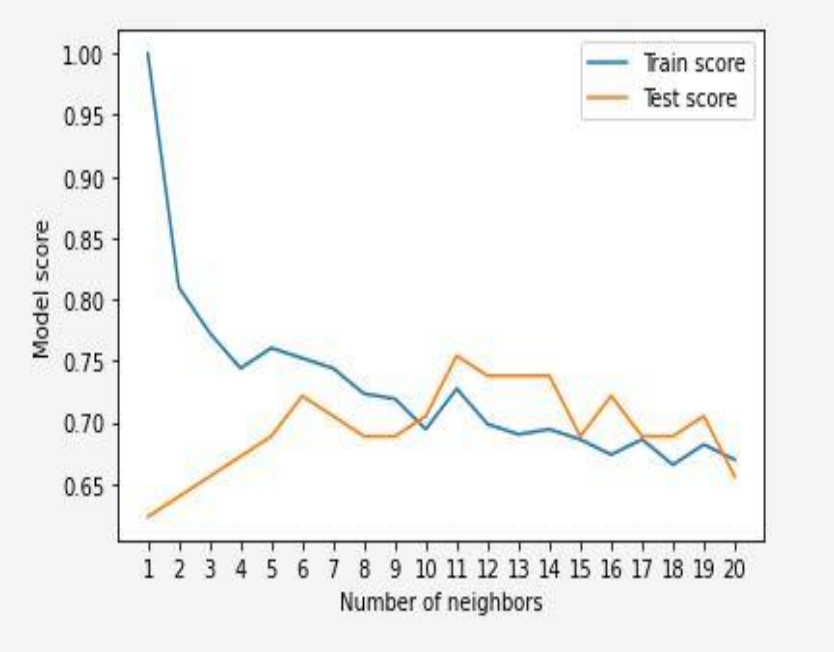

Figure 6 .Maximum KNN score on the test data: 75.41%

5. RESULTS:

The UCI Heart Disease Dataset [12] serves as the basis for applying Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest classification methods. The outputs vary depending on the choice of algorithm and the proportion of test data utilized. For instance, when the test data proportion is set to 20%, the maximum accuracy achieved by Logistic Regression is 87%, as reported in Table 3. Table 3 provides a detailed overview of the performance metrics, including accuracy and error rates, across the different models. The UCI Heart Disease Dataset has been extensively employed in prior studies due to its accessibility and comprehensive patient attributes, enabling diverse methodological approaches. In this research, the dataset serves as the basis for evaluating the predictive capabilities of the selected algorithms. Furthermore, Table 3 presents a comparative study between the models

implemented in this paper and those reported in previous research, thereby highlighting similarities, differences, and improvements in predictive performance.

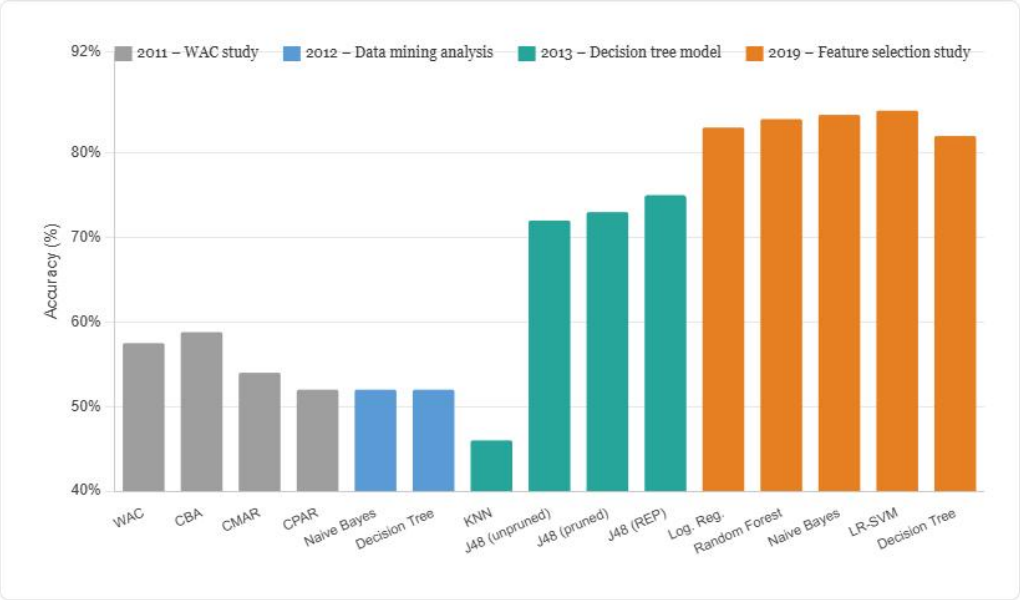

Figure 7: H. K-Nearest Neighbour Result Table

6. CONCLUSION:

The given paper contains an in-depth and most useful machine learning-based model that can help medical practitioners to identify heart diseases early to enable patients to take precautions in time. By using three independent classifiers as illustrated in Table 3, one can clearly see that the ratio of test and training data has a great influence on the operation of the classification model.

IEER
Institute for Excellence in Education & Research

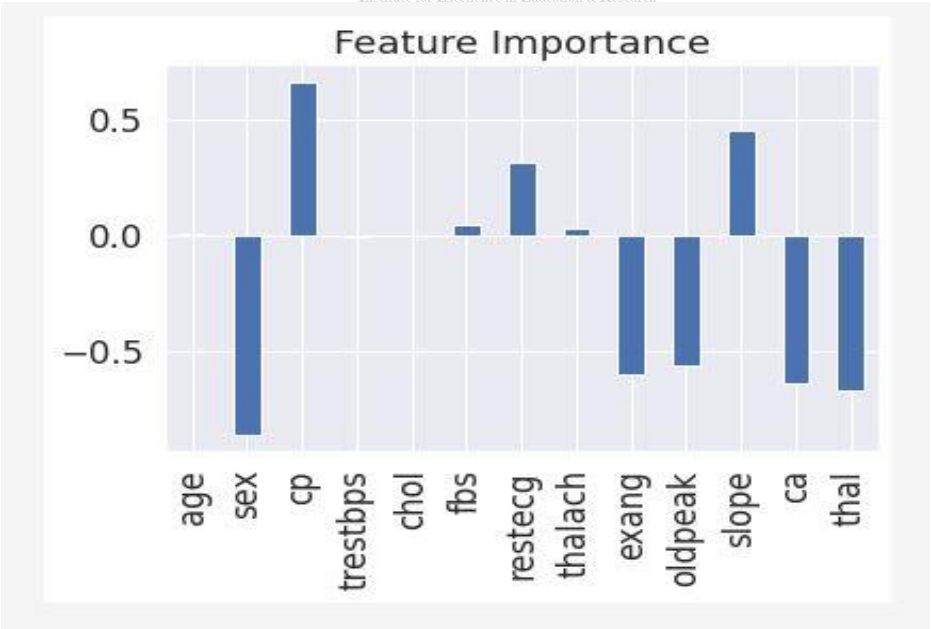

Figure 8. Visual Comparison between the Accuracies of Algorithms

Based on the outcomes of the experiment, one can conclude that logistic regression provides the most successful results with a test data size of 0.2. Nonetheless, it is worth mentioning that at a size of 0.3 random forest performed rather well. Thus, logistic

regression and random forest classification algorithms appear to be the best combination in terms of predicting heart attacks with the highest level of accuracy. The transparent and clear machine learning process will be able to reinvigorate past research work in this area, as it is outdated. The experiments held in this research show that by changing several parameters and adapting them to research requirements, the accuracy rates could be increased. The earlier studies were imprecise, with an accuracy of 72 to 84 percent according to decision trees, SVM, logistic regression, and feature selection techniques, as shown in Fig.6. Additionally, the comparative analysis done in Table 3 shows a very high effectiveness of the method used in this research paper over the previous methods on the same dataset. In this way, it may be concluded that the methodology found in this paper provides a substantial improvement in predicting heart diseases in comparison with the past studies.

7. FUTURE WORK:

In order to improve the model presented in the paper, more features may be included in the dataset to make it more varied. The attributes could be more than the number of attributes that have been used to make predictions, and this could help in enhancing the accuracy. In addition, the accuracy can also be enhanced by increasing the size of the dataset. The analysis of other datasets, other than the UCI dataset, can be insightful. Although the given model employs three algorithms, KNN, logistic regression, and random forest, the inclusion of other algorithms may make this model more efficient. The algorithms, e.g., Support Vector Machine (SVM) and Linear Regression, etc., might provide the results with more accuracy. The algorithms can be positively utilized in a range of disorders, which is not only heart disease, and the methods covered in the paper apply to a wider range of problems. A comparative analysis of the performance of this model in various diseases and algorithms would also help in explaining its efficiency. The comparative analysis may help determine the versatility and strength of the model in different medical conditions and computation methods.

References

1. Pandey, A.K., Pandey, P., Jaiswal, K.L., Sen, A.K.: A heart disease prediction model using decision tree. IOSR Journal of Computer Engineering (IOSR-JCE) 12(6), 83–86 (2013)
2. Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. In: IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012072 (2021)
3. Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A., Bashir, K.: Improving heart disease prediction using feature selection approaches. In: Proceedings of the 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 619–623 (2019)
4. Kumar Mohan, C.T.S.: Effective heart disease prediction using hybrid machine learning techniques. IEEE Access (Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things) (2019)
5. Gavhane, A., K.: Prediction of heart disease using machine learning. In: Proceedings of the 2nd International Conference on Electronics, Communication, and Aerospace Technology (ICECA), pp. 1275–1278 (2018)
6. Arumugam, K., Naved, M., Shinde, P.P., Leiva-Chauca, O., Huaman-Osorio, A., Gonzales-Yanac, T.: Multiple disease prediction using machine learning algorithms. Materials Today: Proceedings 80, 3682–3685 (2023)
7. Taneja, A.: Heart disease prediction system using data mining techniques.

Oriental Journal of Computer Science & Technology, 457–466 (2013)
8. Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L.: Effective heart disease prediction using machine learning techniques. Algorithms 16(2), 88 (2023)
9. Ziasabounchi, N., A., I.: ANFIS based classification model for heart disease prediction. International Journal of Engineering & Computer Science, 7–12 (2014)
10. Soni, J., A., U.: Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering, 2385–2392 (2011)
11. Masethe, H.D., A., M.: Prediction of heart disease using classification algorithms. In: Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA (2014)
12. Ahmad, W., Mehmood, A., Zaidi, H. R., Khan, S. A., Adil, M., Zainoor, M., ... & Shaukat, Z. (2025). A Robust Deep Learning Model for Early Glaucoma Detection Using Retinal Imaging. International Journal of Innovations in Science & Technology, 7(4), 2513-2526.
13. Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L.: Effective heart disease prediction using machine learning techniques. Algorithms 16(2), 88 (2023)

IEER
Institute for Excellence in Education & Research