Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Machine Learning-Based Heart Disease Detection with ANOVA Feature Selection

Fatima Shaker[1], Rana Raad Shaker Alnaily[2], Saja Naeem Turky[3], Elham Kareem Wanas[4], Saja Sadiq Sadon[5]

[1] College of Computer Science and Information Technology, University of Al-Qadisiyah, Diwaniyah, Iraq ,fatima.s.hussian@qu.edu.iq

[2] Mathematics Department, College of Education, University of Al-Qadisiyah, Diwaniya, Iraq, rana.alnaily@qu.edu.iq

[3] College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq, saja.n@uokerbala.edu.iq

[4] College of Computer Science and Information Technology, University of Al-Qadisiyah, Diwaniyah, Iraq, elham.kareem.wanas@qu.edu.iq

[5] College of Computer Science and Information Technology, University of Al-Qadisiyah, Diwaniyah, Iraq, sajasadiq1993@gmail.com

ARTICLE INFO

Article history:
Received: 12 /07/2025
Rrevised form: 02 /09/2025
Accepted : 09/09/2025
Available online: 30/09/2025

Keywords:

Machine Learning, ANOVA Feature Selection,

ABSTRACT

Heart disease(HD) has emerged as one of the most critical health issues that significantly impact human existence. It has become one of the primary causes of mortality worldwide over the past decade. The World Health Organization announced in 2022 that heart disease was the cause of death for nearly one million people, equivalent to 33% of global mortality. In the current century, there is an increase in the use of non-surgical medical technologies, including artificial intelligence methods in the medical field. Machine learning employs many widely utilized algorithms and techniques that are essential in the rapid and efficient diagnosis of heart issues. However, diagnosing heart disease is a difficult task. The vast and expanding scale of medical datasets has hindered professionals' ability to comprehend the intricate correlations among variables and generate precise predictions. Accordingly, the proposed research aims to examine the role of feature selection techniques in supporting machine learning algorithms and improving model accuracy. A medical database of heart diseases with different features was relied upon. In the first stage, data analysis was conducted to understand the nature of the data and ensure its balance before the classification. This encompassed displaying statistical distributions of the data, identifying missing values, and analyzing the relationships between the variables that are independent and the target variable. This step was followed by implementing feature selection techniques, specifically using the ANOVA algorithm to identify the most pertinent features for heart disease detection. Finally, the machine learning algorithms were used on both the complete and reduced datasets to perform the classification. Accuracy, precision, recall, and F1-score were used to evaluate the trained classifiers. The results also show that when the number of features is reduced, the accuracy of classification models improves slightly compared to models trained on the entire set of features

MSC..

https://doi.org/10.29304/jqcsm.2025.17.32427

∗Corresponding author: Fatima Shaker

Email addresses: *fatima.s.hussian@qu.edu.iq*

Communicated by ‘sub etitor’

1. Introduction

Heart failure is a disorder that happens when the heart is incapable to pump enough blood to meet the body's needs. Heart disease has become one of the most prominent health problems globally, significantly impacting public health worldwide. Heart failure is a common and serious condition that affects millions of people, with recent statistics indicating that it affects approximately 26 million people. The causes of heart failure are divided into two main types: those related to the structure of the heart, such as a previous heart attack, and those related to the heart's function, such as high blood pressure. [1]

With the significant advancement in medical data collection systems and electronic patient records, there is an abundance of clinical data related to vital factors associated with heart disease, such as age, blood pressure, cholesterol levels, smoking history, and family history. However, this data is often unstructured and complex, making it difficult to analyze using traditional methods. In this context, exploratory data analysis (EDA) emerges as a vital first step to understanding data patterns, detecting outliers, and analyzing relationships between different variables**.[2]**

Machine learning(ML) is characterized by major transformative ability within the healthcare industry. This remarkable progress can be attributed to superior data processing capabilities, far exceeding what humans are capable of. This has led to the emergence of numerous AI applications in healthcare, leveraging the speed and accuracy of machine learning techniques, opening up new avenues for innovative solutions to various healthcare challenges.[3] Several machine learning methods have been applied to detect heart disease, and our most notable research contributions in this area include developing methods for detecting heart failure using machine learning, as follows:

• The study uses dataset of HD patients from kaggle.

• ANOVA is employed for feature selection to identify the most relevant features for the target. This method also helps address the overfitting and underfitting problems that machine learning algorithms may face.

• Two machine learning classification models, Logistic Regression (LR) and Random Forest (RF), were applied to the databases to identify the most appropriate model to address the problem.

• Both the complete and reduced feature sets were used to test the classification models to examine how feature selection affected the models' performance

2. Related Work

Heart disease research has witnessed significant development in recent decades, considering that it is a primary cause of mortality and a global health burden. Numerous studies focus on analyzing the factors that contribute to the development of these diseases. Some studies have demonstrated the importance of early detection and accurate risk stratification in patients in improving treatment options and reducing complications and mortality rates.

S. S. Abu-Naser et al. [4] proposed to develop a model that use the classification techniques that which combines machine learning and deep learning technique. The proposed model was used to diagnose and predict heart disease. In this study the patient's medical history was used to predict individuals who might develop heart disease which including data such as smoking, stroke, alcohol consumption, physical and mental health, body mass index, and others. The model used was built using random forest classifier with accuracy 92.23%.

Md. Shaheenur Islam Sumon and colleagues [2025] [2] introduced a new model that employ medical tabular data for diagnosing and predicting heart disease . They combined a Transformer-based model into a framework called CardioTabNet. This research focuses on the data analysis step which included 11 clinical variables (7 numerical and 4 categorical), such as age, blood pressure, cholesterol level, and chest pain. They used Multi-Head Self-Attention (MHA) to capture the complex interactions between these variables. Random Forest model was applied to measure the significance of the features and choose the most useful ones .after 10 machine learning algorithms were tested , the Extra Trees model gave the best results with an accuracy of 94.1% and an AUC of 95.0%, showing that the model is very effective in predicting heart disease.

M. S. Guru Prasad et al. [2023] [5] developed a model for predicting heart disease by machine learning algorithms, based on data containing 1,025 cases and 14 clinical variables. The results indicated that the decision

tree algorithm outperformed with 100% accuracy, followed by Naïve Bayes and K-NN with 86.38% and 85.99% accuracy, respectively.

Rasool Reddy Kamireddy et al. [2024][6] developed a model aimed at predicting heart disease using machine learning algorithms, as it is one of the leading causes of death worldwide. The researchers used data from Kaggle and applied algorithms such as SVM, KNN, Random Forest, XGBoost, LightBoost, and SGD after data cleaning and parameter optimization using GridSearchCV. The results showed that the SGD algorithm achieved the highest accuracy of 92.76%, outperforming previous studies.

Zhicheng Wang et al [2024][7] These researchers developed an intelligent model that aims to use machine learning algorithms to predict heart disease, given the prevalence of heart disease and the difficulty of early diagnosis. The researchers used a database containing 303 records and 14 features from the Cleveland Clinic, and applied four algorithms: decision tree (DT), k-nearest neighbors (KNN), random forest (RF), and naive Bayes (NB). The results showed that the random forest algorithm outperformed with an accuracy of 91.80%, followed by KNN with 88.52%, while the NB algorithm was the least accurate with 85.25%.

TAHSEEN ULLAH et al [2024] [8] developed a system that employed machine learning classifiers to predict patients' heart conditions. Initially, features are extracted from ECG signals, then feature selection techniques like FCBF, MrMr, and relief, along with PSO-optimization, are used to select optimal features. Extra trees and Random Forest classifiers trained on the selected features have achieved perfect performance rates

3. Methodology

A schematic illustration of the proposed research framework's layout is shown in Fig (1). This figure offers a comprehensive summary of the components and structure of the suggested framework.

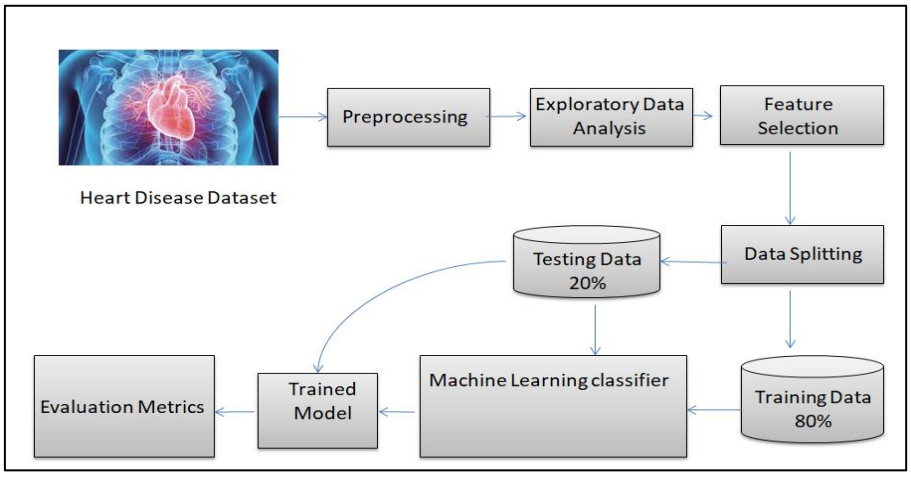

Fig. 1 - The proposed study sequences analysis for HD prediction

3-1 Dataset Description

The quality of the dataset used for statistical predictions has a significant impact on the classification metrics' accuracy. The following dataset was chosen for our study to demonstrate its importance and evaluate its generalizability. The dataset on heart diseases used in this study was obtained from the Kaggle repository. 1025 patient records pertaining to both healthy and heart failure patients are included in the dataset. [9] Age, gender, blood pressure, cholesterol, chest pain, and other clinical and laboratory measurements are just a few of the features in the datasets that can be used to predict heart disease, as shown in Table (1). The outcome variable ''Target'' has a binary value and indicates to the heart disease prediction feature (whether or not there is heart disease).

Table 1 - Heart Disease Dataset

Feature no	Feature name	Feature type	Description
1	Age	Numerical	Patient's age
2	Gender	Categorical	Patient sex (0 for female, 1 for male)
3	Chol	numerical	Assessment of the cholesterol levels of a patient
4	Trestbps	Numerical	Blood pressure at rest
5	Cp	Categorical	Chest pain types
6	Fbs	Categorical	blood sugar during a fast (1= True ,0= False)
7	Thalach	Numerical	Maximum heart rate
8	RestEcg	Categorical	resting electrocardiogram (0=no abnormalities, 1=normal, 2=left ventricular hypertrophy
9	Oldpeak	Numerical	ST depression in relation to the amount of rest
10	Exang	Categorical	Exercise-induced angina (0= no pain, 1= pain)
11	Ca	Categorical	number of main fluoroscopy-colored vessels (0-3)
12	Slope	Categorical	slope of the peak exercise ST section (0=up sloping, 1=flat, 2= down)

13	**thal**	**Categorical**	**Stress in the thallium (0=negative, 1=positive, 2=inconclusive)**
	Target		**Target variable refer to diagnosis of heart disease using angiographic disease status (1= heart disease,0= no heart disease)**

3-2 Preprocessing

This first phase is the primary step of the diagnostic procedure.The data were cleaned for consistency and completeness. The data were examined and found to contain no missing values. All categorical data were represented by numeric variables, e.g., categorical but coded as 1 and 0, thus eliminating the need for data coding. Next, all of the features are normalized to the relevant coefficient using StandardScaler, guaranteeing that each feature has a zero mean and a single variance. However, this step was implemented debending on the model. For example, it was implemented with Logistic Regression, because this algorithm relies on mathematical calculations(feature weighting). The model learns a linear relationship and then applies sigmoid to determine the probability,while the random forest donot need to scaling because it based on ranking not value.

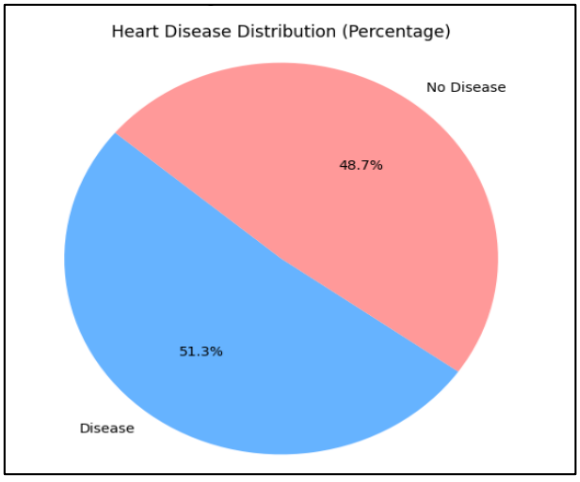

Fig. 2 The Distribution Analysis of the Dataset According to the Target Column

3-3 Feature Selection

The machine learning process requires the extraction of the optimal features for classification, which makes feature selection strategies crucial. This aids in cutting down on execution time as well.[10] Machine learning employs five feature selection strategies: Analysis of Variance (ANOVA), Fast Correlation-Based Filter Solution (FCBF), Least Absolute Shrinkage and Selection Operator (LASSO), Relief, and Minimum Redundancy Maximum Relevance (MrMr). These are common machine learning feature selection techniques.

The ANOVA algorithm was employed in this study. ANOVA is a statistical method that identifies features which show significant variation across different classes of the target variable.[8]

The ANOVA-F test can be implemented in python language using the f_classif() function provided by scikit-learn library. The f_classif() function is used in selecting the most important features (features with largest values) via the SelectKBest class. SelectKBest is a method made available in scikit-learn that grades the features according to a scoring function. The following is the formula to determine ANOVA-F values:

$$\text{variance\_between\_groups} = \frac{\sum_{i=1}^{j} j_i \, (\bar{k}_ı - \bar{k})^2}{(S - 1)}$$

$$variance\_within\_groups = \frac{\sum_{i=1}^{S}\sum_{p=1}^{j_i}(k_{i\,p} - \bar{k}_i)^2}{(N - S)}$$

$$\mathbf{f - value} = \frac{\mathbf{variance\_between\_groups}}{\mathbf{variance\_within\_groups}}$$

where N is the overall sample size, S is the number of groups, $\mathbf{j_i}$ is the number of observations in the jth group, $\bar{\mathbf{k}}_\mathbf{i}$ is the ith group sample mean, $\bar{\mathbf{k}}$ is the overall mean of the data, $\mathbf{k_{i\,p}}$ is the pth observation in the ith out of S groups [12].

3-4 Exploratory Data Analysis

Exploratory data analysis was applied to the study dataset to extract valuable statistics. This analysis relies on graphs and charts that highlight patterns and relationships within the data. Figure (2) illustrates the distribution analysis of the data set using a pie chart based on the target property .The analysis shows that the dataset contains 51.32٪ healthy patients and 48.68٪ have not been infected with the disease. Furthermore, 713males and 312 females are found with heart failure disease in the dataset as shown in figure(3).The dataset used to construct the machine learning models is nearly balanced, as this analysis shows.

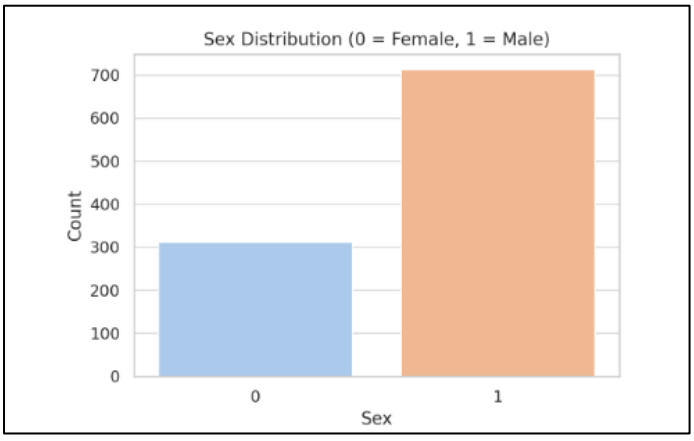

Fig.3 Sex Distribution of dataset

The significance of features provides insight into each feature's applicability and predictive power within dataset. Using the ANOVA f-test technique, the features exang, oldpeak, cp, and thalach thal have the top four high f-scores, as illustrated in Fig.4. The analysis determines that the remaining dataset attributes exhibit a robust association.

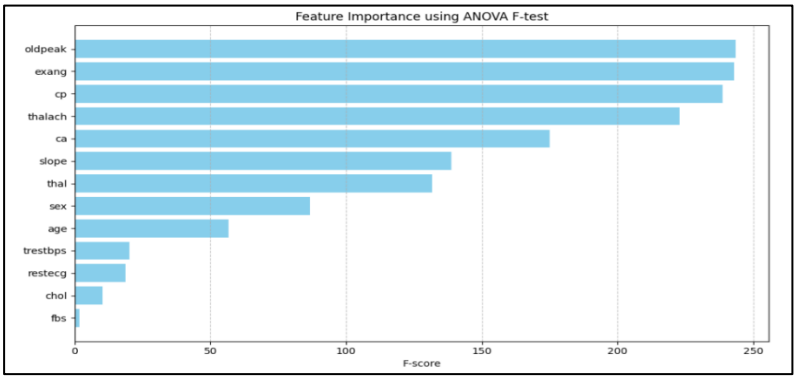

Fig.4 Feature Importance using ANOVA-F-test

3-5- Applied Machine Learning Techniques

Model construction is automated by a specialized method called machine learning. Machines can find insights and hidden patterns in datasets by using algorithms. Crucially, in machine learning, humans don't specifically tell machines where to look for insights; rather, the algorithms allow the machines to pick up new information and scenarios and modify their methods and outputs accordingly. Machine learning is an effective technique for processing and interpreting complicated datasets because of its iterative nature, which enables constant adaptation and development.**[3]**

A- Logistic Regression

logistic regression (LR) is a customarily used supervised machine-learning technique that can be used for both regression and classification issues. The logistic regression method utilize probability to predict the labeling of categorical data. [1] LR analysis is a statistical method that predicts the result of a class-dependent variable (or class of variables) from a collection of anticipated variables. A class with two categories is used in logistic regression, which is mostly used to compute and forecast the likelihood of a particular outcome. [13] Comparable to our study dataset's target column, which contains two distinct binary numbers. The dataset's presented heart patients are represented by a one, whereas patients with no likelihood of heart illness are represented by a zero..

$$P(y = 1|x) = 1 + \frac{1}{1 + e^{-z}}$$

where

$z = \beta 0 + \beta 1x1 + \cdot \cdot \cdot + \beta pxp$

B- Random Forest

Random Forest is a strategy for classifying data that makes use of an extensive collection of decision trees. By combining bagging and feature randomization, an uncorrelated forest of trees is generated, where the accuracy of the committee prediction surpasses that of any individual tree frequently employed in classification and regression tasks. To attain the most favorable result, this classification methodology generates and merges numerous decision trees. For tree learning, it primarily employs bootstrap aggregation or bagging.[14**]**

4. Performance Metrics

We employed four widely recognized performance evaluation metrics: accuracy, precision, recall, and F1-score, to evaluate the performance of machine learning classification models.[15]
The assessment was conducted utilizing the confusion matrix , which illustrates the outcomes of the classification process concerning false positives (FN), true negatives (TP), true positives (TN), and false positives (FP). TN denotes instances that were accurately forecasted as belonging to the negative class, whereas TP signifies instances that were accurately predicted as belonging to the positive class. Positive-class inaccurate predictions (FP) and negative-class incorrect predictions (FN) are denoted as FP and FN, respectively.

Accuracy expresses the ability of the proposed model to determine the percentage of samples that are correctly classified, and is calculated according to the following equation:

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Recall: recall refers to the model's ability to accurately detect all positive cases. It is defined as the percentage of positive samples that are correctly identified and is calculated using the following equation:

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp+Fn}} \text{ ……(2)}$$

Precision: It identifies the accuracy of the classifier and can be computed using the provided data. By contrasting actual TP versus predicted TP..

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp+Fp}} \text{ ……..(3)}$$

F-measure: This measure is a statistical tool for assessing the efficiency of a classification model. It is based on calculating the harmonic mean between precision and recall, giving them equal weight. It allows the model's performance to be summarized and compared with a single score that combines prediction accuracy and recall. It is calculated using the following equation:

$$\text{F} - \text{measure} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \text{ … … …. . (4)}$$

5- Expert Results and Discussion

This part illustrates the classification models' accuracy from two perspectives. First, machine learning algorithms were implemented on the full-featured dataset to identify the best-performing model. After that, the algorithms were applied to the selected feature dataset to assess the impact of feature reduction techniques on the classifiers. Several metrics were used in the analysis, such as precision, accuracy, recall, and f-score.

Feature selection techniques are a fundamental tool in machine learning and data analysis because they enable the selection of the most important and instructional features for the development of predictive models. The size and characteristics of the dataset, the type of features, and the feature selection algorithm applied are important factors on which the effectiveness of feature selection techniques depends. In this research, The ANOVA F-test was selected for feature selection because the dataset contains numerical input features and a categorical target variable (binary: presence or absence of heart disease). ANOVA is well-suited for identifying features that have statistically significant differences in mean values across the target class. Among 13 features, 8 were selected to forecast the presence or absence of heart disease.the feature selection technique also it is benefit to avoid the overfitting.

5-1 Results of Classification using the Full Feature Set

This part details the evaluation of machine learning models on datasets utilizing the complete collection of attributes to forecast the binary disease outcome. All prediction models were trained on the complete dataset, utilizing 80% for training and 20% for testing subsets. Table (3) display the binary classification outcomes of the machine learning model to forecast heart disease for the dataset . The findings revealed that the LR model demonstrated the greatest degree of accuracy, measuring at 0.99. Along with RF the other classifier represented by LR worked well and gave acceptable prediction accuracy using the entire feature set. The enhanced accuracy attained by RF can be attributed to its proficiency in identifying patterns within intricate medical datasets.

Table (3)Classification outcomes of ML model for the dataset with full features.

Model	Accuracy	Precision	Recall	F-score
Random forest	**0.98**	**0.99**	**0.99**	**0.99**
Logistic Regression	**0.79**	**0.80**	**0.79**	**0.79**

5.2 Results of Classification using the Reduced Feature Set

Based on individual feature scores, we chose the most notable features from the entire feature space to discover potential biomarkers and examine the effect of the feature selection technique on classification accuracy. The ANOVA-F test was used to determine which features had the greatest influence on the dataset's outcome, as seen in Fig().eight features were selected, taking into account the feature weights determined by the ANOVA-F test. We used only the chosen features as inputs to assess each classification model's performance. Table (4) display the classification performance of all model when using the reduced feature set from the data. Analyses showed that machine learning algorithms performed better after reducing the number of features compared to using the full feature set. The Random Forest (RF) model achieved the highest accuracy of 1.0, with precision, recall, and F1 score values of 1.0, using only 8 features as input. Models trained on the smaller feature set also took less computational time.

Table (4) classification outcomes of ML models for a dataset with a lesser feature set.

Model	Accuracy	Precision	Recall	F-score
Random forest	**1.0**	**1.0**	**1.0**	**1.0**
Logistic Regression	**0.80**	**0.82**	**0.81**	**0.81**

The performance comparisons of the past proposed studies on our dataset are analyzed in Table 5. The parameter used for comparison is accuracy, which is usually considered the most important technique for evaluating machine learning algorithms. The analysis demonstrates that the previous study [10] employed the same dataset and applied four classifiers (Random Forest, Logistic Regression, Extra Trees, and Gradient Boosting), achieving an accuracy of 94%. In contrast, our study utilized the same dataset but focused only on two classifiers (Random Forest and Logistic Regression). Despite employing fewer classifiers, our models achieved superior accuracy, reaching 100%. Focusing on two classifiers with proper parameter optimization avoided the overfitting risk associated with employing multiple models without adequate tuning**.**

Table (5)Performance comparisons with previous study

References	Techniques	Performance accuracy
[10]	Rf,LR, ET,GB	94%
Our study	RF, LR	100%

6- Conclusion

This work suggested a model for classifying heart illness via machine learning algorithms and appropriate feature selection methods. The suggested model exhibited a substantial influence of feature selection on improving the performance of machine learning algorithms for HD prediction. The dataset of 1025 patient records is utilized to construct the applied models. The ANOVA feature selection method is suggested, which improves performance by choosing the most relevant features with target. In addition, both complete and reduced feature sets were used in the classification experiments to examine the influence of feature selection on the performance of different ML prediction model. The highest accuracy obtained with the complete feature set was 0.98 for the dataset on heart disease. After applied the reduced feature set, the accuracy reached to 1.00. The analysis presented that reducing the number of features led to better results compared to using the complete feature set. Empirical results show that even with a limited number of features, we can correctly classify HD by using a feature selection technique. We can conclude that by employing feature selection, only the most significant features related to heart disease are chosen, which lessen computational complexity and increases the accuracy of the prediction model.

References

[1] Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. S. (2023). Effective feature engineering technique for heart disease prediction with machine learning. *IEEE Access*, *11*, 56214-56224.

[2] Sumon, M. S. I., Islam, M. S. B., Rahman, M. S., Hossain, M. S. A., Khandakar, A., Hasan, A., ... & Chowdhury, M. E. (2025). CardioTabNet: a novel hybrid transformer model for heart disease prediction using tabular medical data. *Health Information Science and Systems*, *13*(1), 44.

[3] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, *14*(2), 144.

[4] Abu-Naser, S. S., Obaid, T., Abumandil, M. S., & Mahmoud, A. Y. (2022, November). Heart Disease Prediction Using a Group of Machine and Deep Learning Algorithms. In *The International Conference of Advanced Computing and Informatics* (pp. 181-196). Cham: Springer International Publishing.

[5] Prasad, M. G., Kumar, D. S., Pratap, M. S., Kiran, J., Chandrappa, S., & Kotiyal, A. (2023, June). Enhanced Prediction of Heart Disease Using Machine Learning and Deep Learning. In *International Conference on Advanced Communication and Intelligent Systems* (pp. 1-12). Cham: Springer Nature Switzerland.
[6] Kamireddy, R. R., & Darapureddy, N. (2023). A Machine Learning-Based Approach for the Prediction of Cardiovascular Diseases. *Engineering Proceedings*, *56*(1), 140.
[7] Wang, Z., Gu, Y., Huang, L., Liu, S., Chen, Q., Yang, Y., ... & Ning, W. (2024). Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data. *Cardiovascular Diabetology*, *23*(1), 351.
[8] Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., & Algarni, A. (2024). Machine learning-based cardiovascular disease detection using optimal feature selection. *IEEE Access*, *12*, 16431-16446.
[9] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data
[10] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, *9*, 19304-19326.
[11] Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., & Algarni, A. (2024). Machine learning-based cardiovascular disease detection using optimal feature selection. *IEEE Access*, *12*, 16431-16446.
[12] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, 100060.
[13] Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific reports*, *13*(1), 22588.
[14] Ahmed, M., & Husien, I. (2024). Heart disease prediction using hybrid machine learning: A brief review. *Journal of Robotics and Control (JRC)*, *5*(3), 884-892.
[15] El-Sofany, H. F. (2024). Predicting heart diseases using machine learning and different data classification techniques. *IEEE Access*