Impact Factor 6.1

Journal of Cyber Security

ISSN:2096-1146

Scopus

DOI

Google Scholar

More Information
www.journalcybersecurity.com

Crossref doi Google Scholar scopus

Enhancing Cardiovascular Disease Prediction Through Machine Learning and Feature Selection: A Bagging Ensemble Approach

Dr. N. Hari Priya

Abstract

Cardiovascular disease is a leading cause of global mortality and predicting this disease is a crucial challenge in clinical data analysis. Effective prevention and treatment depend on early and accurate screening of affected persons. Making informed healthcare decisions relies on accurately identifying the people at risk of severe illness. The application of Machine Learning (ML) techniques and feature selection shows great potential in the early detection of cardiac disease. Machine learning algorithms are able to uncover patterns and cardiovascular health risk factors by successfully choosing the most significant factors from large datasets. This enables the creation of precise prediction models that support early intervention and detection. Thus, it improves patient outcomes in battling cardiovascular disease. In this study, I have determined the features that are most pertinent to disease prediction using wrapper feature selection technique known as Support Vector Machine - Recursive Feature Elimination (SVM-RFE). Then a Bagging Ensemble Machine Learning Classifier is used, which combines various base classifiers including k-Nearest Neighbors, Random Forests, and Decision Trees. Based on the experimental results, this ensemble approach which combines the predictions from several models attains an accuracy of 99% which outperforms other ML algorithms and similar works. The performance of the proposed ensemble learning model is further validated using the receiver operating characteristic curve and the value of ROC-AUC is 1.00. The findings demonstrate that the proposed ensemble model is capable of accurately predicting the risk of cardiac disease.

Keywords: Cardiovascular disease, Machine Learning, Prediction, Feature Selection, Ensemble Machine Learning, Bagging, SVM-RFE.

1. INTRODUCTION

Cardiovascular disease (CVD) remains a significant global health challenge, representing a leading cause of mortality worldwide [1]. Early detection and diagnosis are crucial for improving patient outcomes and reducing mortality rates. Accurate prediction of CVD is essential for effective prevention, timely intervention and in improving patient outcomes [2]. In recent years, the use of Machine Learning (ML) techniques has proven to be highly effective in improving the early detection and prediction of cardiovascular diseases. However, the efficacy of these models is contingent upon the selection of relevant features that capture the underlying patterns in the data while mitigating the impact of noise and irrelevant variables. By analyzing large datasets and identifying relevant features, ML algorithms can uncover intricate patterns and risk factors associated with CVD.

This research work explores the potential of combining machine learning with feature selection techniques to develop a robust and accurate model for predicting cardiovascular disease. Specifically, the study employs a wrapper feature selection technique known as Support Vector Machine - Recursive Feature Elimination (SVM-RFE), to identify the most pertinent features for disease prediction. Subsequently, a Bagging Ensemble Machine Learning Classifier is utilized, which integrates multiple base classifiers including k-Nearest Neighbors, Random Forests, and Decision Trees. This ensemble approach aims to capitalize on the strengths of diverse models, enhancing prediction accuracy and robustness.

The primary objective of this research is to demonstrate the efficacy of the proposed ensemble learning model in accurately predicting the risk of cardiac disease. Through comprehensive experimentation and performance evaluation, the study assesses the accuracy and reliability of the ensemble model in comparison to alternative ML algorithms and existing works in the field. Additionally, the validation of the model using the Receiver Operating Characteristic (ROC) curve further substantiates its predictive capabilities. The advancement of predictive analytics in healthcare is achieved by offering a sophisticated ML-based approach for cardiovascular disease prediction. The findings underscore the potential of ensemble learning techniques in improving risk assessment and facilitating early intervention strategies, thereby mitigating the burden of cardiovascular disease on public health.

The remaining section of this paper is structured as follows: Section 2 discusses previous studies in the realm of cardiovascular disease (CVD) prediction and feature selection. Section 3 describes the methodology, encompassing data pre-processing, the feature selection technique used, and the application of a Bagging Ensemble approach for machine learning. Section 4 details the experimental setup, evaluation metrics, and specifics of the implementation. Section 5 wraps up the paper with a summary of the results and their potential impact on clinical practice.

2. RELATED WORKS

Ibomoiye et al. [3] presented an enhanced ensemble learning method for forecasting the risk of heart disease. Their approach involves randomly partitioning the dataset into smaller subsets based on mean values. Utilizing a homogeneous CART model alongside an Accuracy-based Weighted Aging Classifier Ensemble, they assess the system's accuracy. Comparative analysis against previous machine learning algorithms reveals that this approach achieves superior accuracy compared to certain scholarly works.

Sharmila et al. [4] proposed the utilization of non-linear classification algorithms for heart disease prediction. Their approach involved employing big data tools such as Hadoop Distributed File System (HDFS) and MapReduce in conjunction with Support Vector Machines (SVM) to predict heart disease using an optimized attribute set. This research investigated the efficacy of various data mining

techniques for heart disease prediction. It recommended the use of HDFS for storing large datasets across different nodes and executing prediction algorithms using SVM in parallel fashion. The parallel implementation of SVM resulted in improved computation time compared to sequential execution.

Ramatenki et al. [5] introduced an ensemble technique to forecast heart disease and discern its presence or absence. Their approach involved employing four distinct algorithms: KNN, Modified KNN, SVM, and Decision Tree. Post-training, the performance of these algorithms underwent evaluation and comparison, employing metrics such as accuracy, recall, precision, and F1 score. Notably, the ensemble classifier demonstrated superior accuracy compared to all individual algorithms.

Benjamin et al. [6] scrutinized the influence of ensemble learning algorithms on precise heart disease prediction, aiming to pinpoint the most effective ensemble learning classification. They utilized the UCI data repository to evaluate the performance of stacking, Bagging, and Boosting ensemble methods. Metrics such as precision, recall, F-measure, and ROC were employed in this analysis. Following the comparative study, AdaBoost emerged as the top performer.

Muhammad et al. [7] devised a novel approach centered on significant features and ensemble learning models. In their methodology, they employed Random Forest with Stratified KFold for parameter tuning. The utilization of cross-validation aids in preventing both overfitting and underfitting of the data. By leveraging the bootstrapping technique alongside stratified KFold, Random Forest generates varying results by considering parameters such as train data, test data, and the number of estimators for each tree. Subsequently, the testing data undergoes a majority voting decision tree sampling process to determine its veracity. Notably, this method achieves high accuracy when compared to existing approaches.

Chala Beyene et al. [8] advocated for the prediction and analysis of heart disease occurrences through data mining techniques. Their primary goal was to facilitate early automatic diagnosis of the disease, thereby expediting treatment. The proposed methodology holds particular significance for healthcare organizations lacking in-depth expertise, as it leverages various medical attributes such as blood sugar levels, heart rate, age, and sex to identify the presence of heart disease. Dataset analyses were conducted using WEKA software.

Adithya et al. [9] devised an optimal multi-disease prediction framework employing hybrid machine learning techniques. They employed the genetic algorithm-based recursive feature elimination (GAE-RFE) method alongside AdaBoost to predict diseases. This feature selection process was applied to Random Forest, Decision Tree, XGBoost, and AdaBoost models. A 10-fold cross-validation approach was utilized to test the dataset. The GAE-RFE method significantly enhanced the performance of the machine learning algorithms. AdaBoost particularly exhibited remarkable precision, specificity, sensitivity, and F-measure values compared to benchmark techniques.

Bhanu et al. [10] introduced a soft voting classifier model for heart disease prediction, comparing the performance across three benchmark datasets. Data pre-processing was conducted to

eliminate noisy and missing values, while Min-max scalar was employed for data standardization. The voting classifier utilized Naive Bayes, Random Forest, SVM, and Gradient Boosting to construct the model. Leveraging the majority voting process, the proposed ensemble method demonstrated improved accuracy.

3. RESEARCH METHODOLOGY

3.1 Data Sources

The data set utilized in this study has been downloaded from Kaggle [11]. The dataset comprises 1025 instances and 14 features related to cardiovascular health, including demographic data (age and sex), cardiovascular indicators, and a target variable. The dependent variable "target" is a binary variable which represents 1 for presence of heart disease and 0 for absence of heart disease. The features contained in the dataset is shown in Table 1.

Table 1. List of all features

S. No	Attribute	Expansion
1	Age	Age of the patient
2	Sex	Sex (0 = Female; 1 = Male)
3	cp	Chest Pain Type (4 types)
4	trestbps	Resting Blood Pressure
5	chol	Serum Cholesterol in mg/dl
6	fbs	Fasting Blood Sugar > 120 m
7	restecg	Resting electrocardiographic results
8	thalach	Maximum Heart Rate achieved
9	exang	Exercise Induced Angina
10	oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) Coloured by Flourosopy
13	thal	Thalassemia
14	target	1 – Presence; 0 - Absence

3.2 Data Pre-processing

Data preprocessing is an essential phase in a data analysis or machine learning project. It involves converting raw data into a format that is appropriate for modeling. This process ensures that the data is clean and structured, facilitating effective analysis and predictive modeling. The dataset used in the study does not contain any missing values. For further analysis, outlier detection has been carried out using box plot and potential outliers were identified and it is clearly depicted in Figure 1.

Results of Outlier Detection:

Resting Blood Pressure (trestbps):

The maximum value of trestbps is 200 mm Hg, which is significantly higher than the 75th percentile (140 mm Hg). This could be an outlier indicating abnormally high blood pressure.

Serum Cholesterol (chol):

The maximum cholesterol level is 564 mg/dl, which is quite high compared to the 75th percentile of 275 mg/dl, suggesting a potential outlier.

Maximum Heart Rate Achieved (thalach):

The minimum value of thalach is 71 bpm, which is quite low compared to the 25th percentile (132 bpm), and could indicate an outlier.

ST Depression Induced by Exercise Relative to Rest (oldpeak):

The maximum value oldpeak is 6.2, significantly above the 75th percentile of 1.8, which might indicate the presence of outliers.

Number of Major Vessels Colored by Fluoroscopy (ca):

A value of 4 for ca is outside the normal range of 0-3 mentioned in the dataset documentation, indicating either a data entry error or an extreme value.

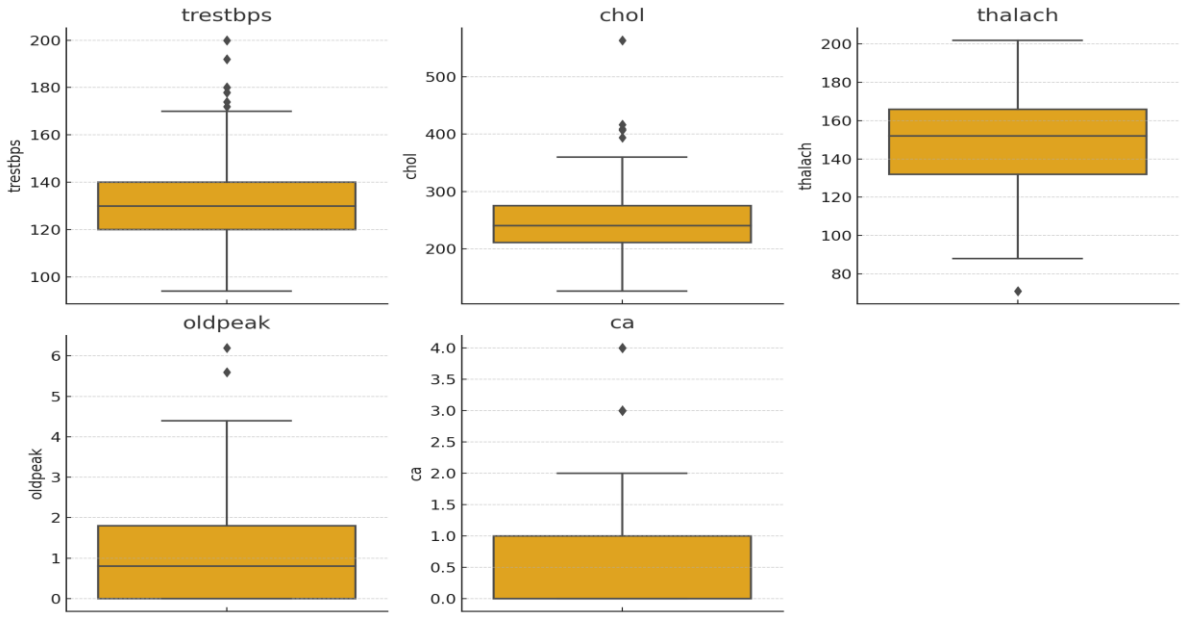

Figure 1. Outlier Detection using Box Plot

3.3 Feature Selection

Feature selection is a critical process in machine learning that involves selecting a subset of relevant features for use in model construction [12]. The goal is to improve the model's performance by eliminating redundant, irrelevant, or noisy data that can lead to overfitting or unnecessarily complex models. Effective feature selection can lead to simpler models, reduced computation time, and better generalization to new samples. SVM-RFE is a wrapper feature selection method that uses a Support

Vector Machine (SVM) to identify the most significant features [13]. This technique uses the coefficients of a support vector machine to rank the importance of features within a dataset. The rank obtained by the features using the SVM-RFE method is shown in Figure 2. Features such as thal, ca, and slope are positioned at the top of the chart, indicating their high relevance in predicting the target variable. Sex, age, and cp are positioned towards the bottom of the chart, suggesting they have less influence on the model's decision-making process. SVM-RFE technique and the corresponding visualization help in understanding which features contribute most to the predictive accuracy of the model, guiding efficient data processing and simplifying the model by focusing on the most relevant features.

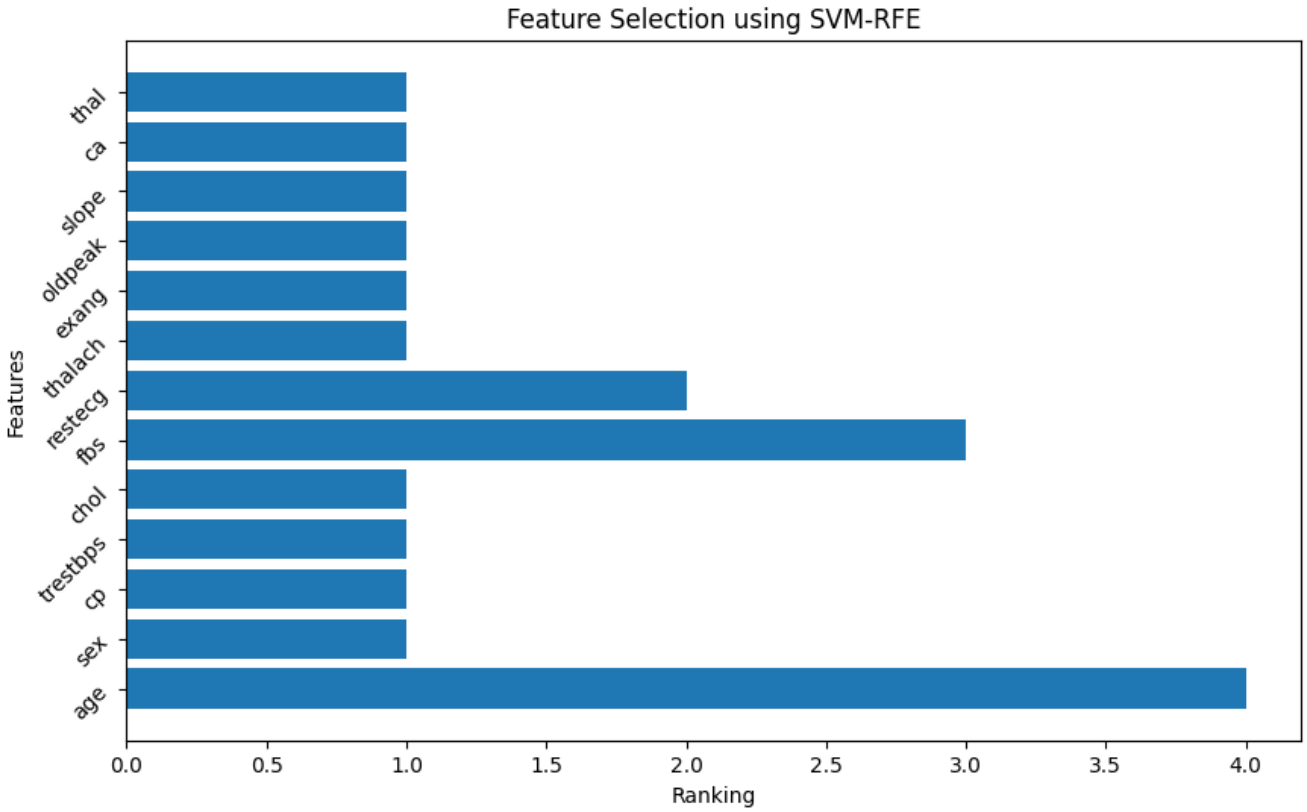

Figure 2. Feature Selection using SVM - RFE

3.4 Descriptive Analytics

Comprehensive descriptive analytics will include the following components.

A. Summary Statistics

B. Distribution of features

C. Target Distribution

D. Correlation Matrix

A. Summary Statistics

Each of the 1025 records in the dataset comprises 14 attributes, inclusive of the target variable. The count denotes the number of valid data points present. The mean offers insight into the central tendency of the dataset. Standard deviation quantifies the dispersion or variability in attribute values.

The range, spanning from the minimum to the maximum value, showcases the extent of data values for each attribute. For a quick grasp of data distribution and potential skewness, the 25th, 50th (median), and 75th percentiles serve as handy indicators. The brief overview of the dataset, focusing on central tendencies, standard and deviations is shown in Figure 3.

	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.434146	9.072290	29.0	48.0	56.0	61.0	77.0
sex	1025.0	0.695610	0.460373	0.0	0.0	1.0	1.0	1.0
cp	1025.0	0.942439	1.029641	0.0	0.0	1.0	2.0	3.0
trestbps	1025.0	131.611707	17.516718	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.000000	51.592510	126.0	211.0	240.0	275.0	564.0
fbs	1025.0	0.149268	0.356527	0.0	0.0	0.0	0.0	1.0
restecg	1025.0	0.529756	0.527878	0.0	0.0	1.0	1.0	2.0
thalach	1025.0	149.114146	23.005724	71.0	132.0	152.0	166.0	202.0
exang	1025.0	0.336585	0.472772	0.0	0.0	0.0	1.0	1.0
oldpeak	1025.0	1.071512	1.175053	0.0	0.0	0.8	1.8	6.2
slope	1025.0	1.385366	0.617755	0.0	1.0	1.0	2.0	2.0
ca	1025.0	0.754146	1.030798	0.0	0.0	0.0	1.0	4.0
thal	1025.0	2.323902	0.620660	0.0	2.0	2.0	3.0	3.0
target	1025.0	0.513171	0.500070	0.0	0.0	1.0	1.0	1.0

Figure 3. Summary Statistics of the dataset

B. Distribution of Features

The histograms depicted in Figure 4 illustrates the distribution patterns of variables, highlighting key characteristics such as skewness, potential outliers, and overall distribution shape.

Chest Pain Type (cp):

The histogram for chest pain type (cp) indicates a categorical distribution with varying frequencies across types. The data suggests a non-uniform distribution, which could be useful in assessing the prevalence and significance of different chest pain types in a clinical setting.

Resting Blood Pressure (trestbps):

This histogram displays a quasi-normal distribution with a slight right skew, suggesting most patients have blood pressure in a moderate range with fewer instances of very high values, potentially indicating hypertensive outliers.

Cholesterol Level (chol):

Cholesterol levels are depicted with a distribution that approximates a normal curve but with a tail extending to higher values. This skewness to the right suggests the presence of individuals with significantly elevated cholesterol levels, which are critical in cardiovascular risk assessments.

Fasting Blood Sugar (fbs):

Displaying a binary distribution, this histogram categorizes individuals based on whether their fasting blood sugar is above or below 120 mg/dL, highlighting a significant divide in metabolic health within the population.

Resting Electrocardiographic Results (restecg):

The data shows a categorical distribution with a prominent mode, suggesting one category significantly dominates, which may be indicative of a common cardiac condition within the studied group.

Maximum Heart Rate Achieved (thalach):

This variable shows a left-skewed distribution, with most data clustered at higher heart rates and tailing off as heart rates decrease. This pattern might reflect the aerobic capacity or cardiovascular health of the population.

Exercise Induced Angina (exang):

Another binary histogram, which categorizes individuals into those who do and do not experience angina induced by exercise, useful for understanding the prevalence of exercise-related cardiac symptoms.

ST Depression Induced by Exercise Relative to Rest (oldpeak):

The histogram for oldpeak shows a right-skewed distribution, with most values clustering near zero and extending to higher values, potentially indicating cases with severe exercise-induced cardiac stress.

Slope of the Peak Exercise ST Segment (slope):

The histogram for slope reflects a categorical distribution with some categories more prevalent than others, possibly correlating with specific cardiac outcomes post-exercise.

Number of Major Vessels Coloured by Fluoroscopy (ca):

This histogram is highly skewed right, with a majority of observations showing zero visible vessels, and fewer cases showing more, which might suggest a lower prevalence of severe coronary artery disease.

Thalassemia (thal):

A bimodal distribution is shown with peaks at fixed defect and reversible defect types of Thalassemia, offering insight into the types of this genetic disorder prevalent in the population studied.

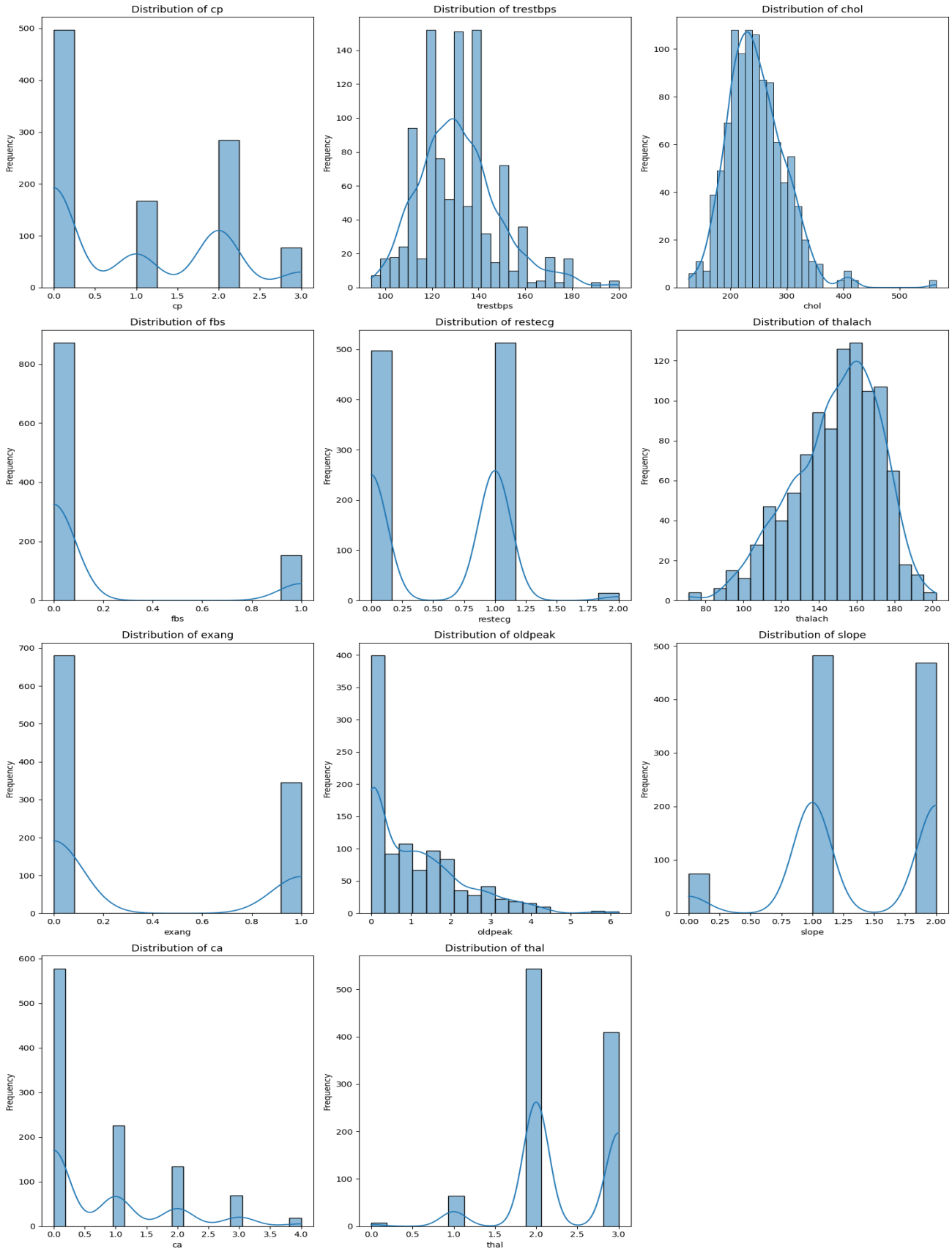

Figure 4. Distribution of Features using Histogram

C. Target Distribution:

Figure 5 illustrates the distribution of the target variable in the dataset, indicating individuals with heart disease (Class 1) and without heart disease (Class 0).

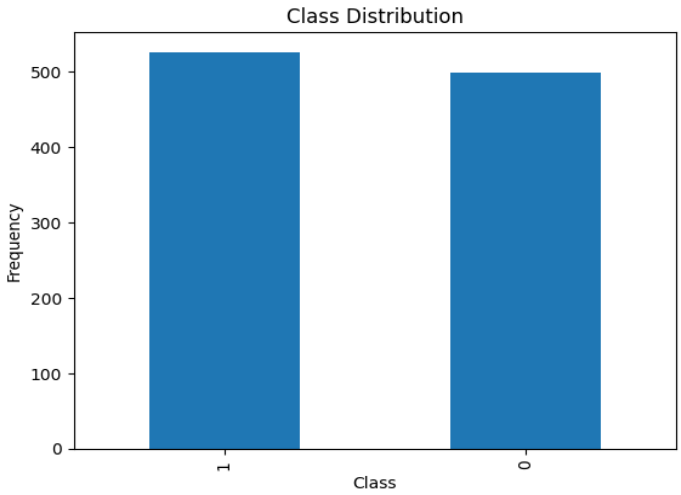

Figure 5. Distribution of Target Variable

D. Correlation Matrix:

The correlation matrix is presented in the form of a heatmap in Figure 6, depicting the pairwise correlation coefficients between the features and the target variable. The heatmap is a powerful visualization tool for understanding the relationships between multiple variables in a dataset. Each cell in the heatmap represents the correlation coefficient between two variables. The value of the correlation coefficient ranges from -1 to 1, indicating the strength and direction of the linear relationship between the variables.

Heatmap Analysis:

Exang (Exercise Induced Angina) and Oldpeak: These have a negative correlation of -0.58, indicating that as exercise-induced angina increases, ST depression decreases, which might suggest differential responses to exercise stress tests among individuals.

Thalach and Slope: The positive correlation of 0.31 suggests that individuals with a higher heart rate tend to have a more positive slope of peak exercise ST segment, which could be indicative of better heart function during exercise.

CA (Number of Major Vessels) and Age: A correlation of 0.27 suggests that older individuals might have more visible major vessels during fluoroscopy, a possible indicator of aging effects on cardiovascular health.

In summary, this heatmap provides a comprehensive overview of how various medical and physiological variables interrelate, offering insights that can guide further analysis and decision-making in a healthcare context.

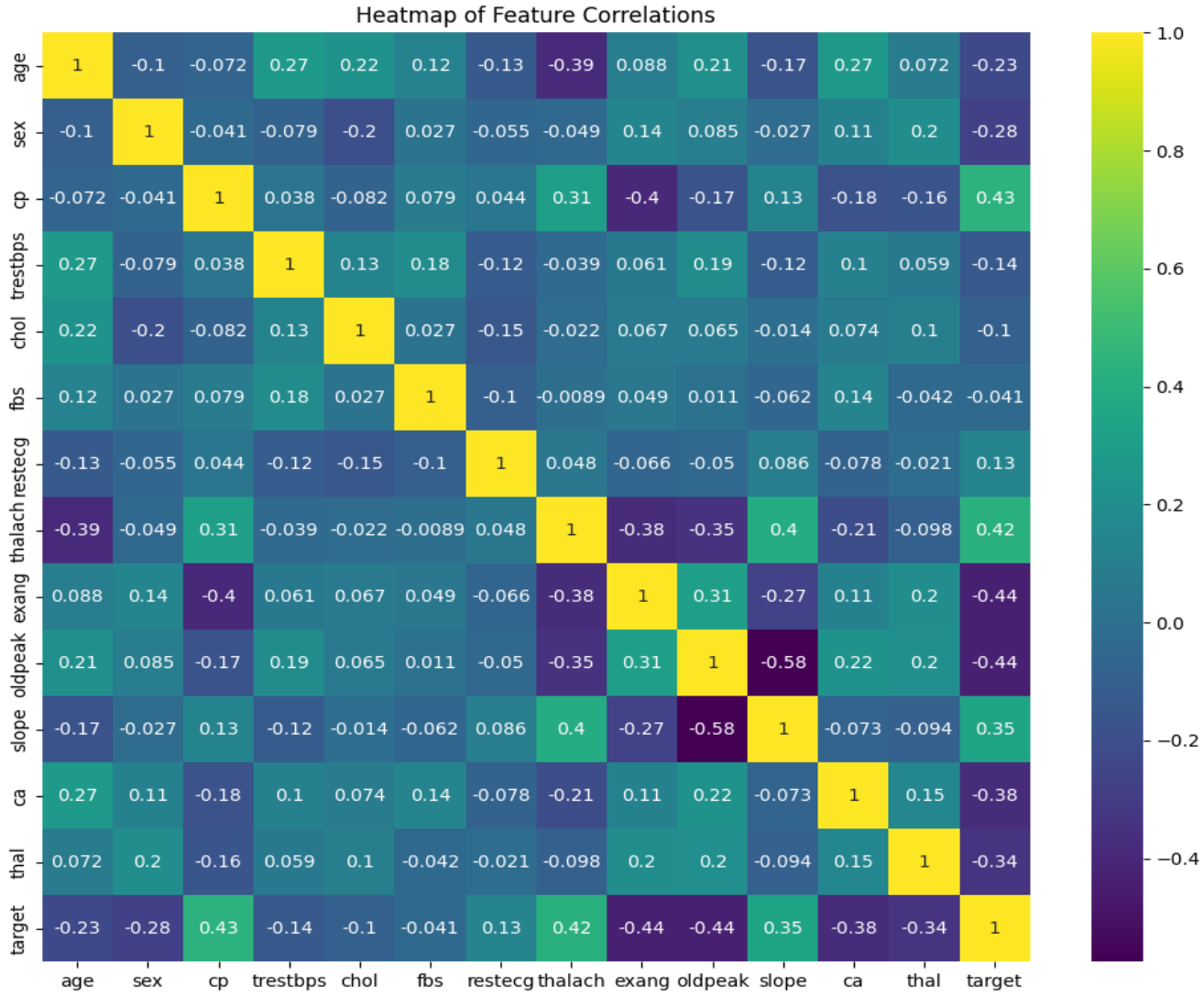

Figure 6. Correlation Matrix using Heatmap

3.5 Bagging Ensemble Machine Learning Approach

In the context of machine learning, the Bagging Ensemble Classifier is a powerful method designed to improve the stability and accuracy of machine learning algorithms, particularly through reducing variance, which often leads to better performance on unseen data. The text describes the application of this technique using three different types of base classifiers, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors (KNN) Classifier.

Decision Tree Classifier

A Decision Tree Classifier is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It's easy to interpret and understand. It can handle both numerical and categorical data and can model complex relationships with simple rules [14].

Random Forest Classifier

Random Forest is an ensemble of Decision Trees, typically trained via the bagging method. Each tree in the forest is built from a sample drawn with replacement from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of them. It reduces the variance/overfitting problem of Decision Trees without increasing error due to bias, and it handles unbalanced data well. It is also very robust to outliers and can handle irrelevant features efficiently [15].

K-Nearest Neighbors (KNN) Classifier

KNN is a simple, instance-based learning algorithm where the function is only approximated locally and all computation is deferred until function evaluation. It is a type of lazy learning where the function is approximated locally and all computation is deferred until classification. KNN is very simple to implement and understand. It's robust to noisy training data and effective if the training data is large.

Usage of classifiers in Bagging Ensembles

Each of these classifiers can be used as a base estimator in a bagging ensemble. The idea is to create multiple instances of these classifiers, each trained on random subsets of the training data, and then aggregate their individual predictions to form a final verdict. This approach harnesses their strengths while mitigating weaknesses such as overfitting and variance. In the bagging context, multiple versions of these models are trained independently on different bootstrapped subsets of data, and their predictions are then combined through voting (for classification). This typically results in a performance improvement and more robust predictions. In essence, bagging ensembles provide a systematic way of merging multiple predictions to improve the decision-making process. This technique leverages the strength of each individual classifier and mitigates their weaknesses, leading to a more accurate and stable model.

4. RESULTS AND DISCUSSIONS

The Research work was carried out using Python which performs a series of steps in a machine learning workflow, using a dataset presumably related to a classification problem. The steps include data pre-processing, feature selection, model training, evaluation, and visualization of results. The data is split into training and testing sets, with 80% of data used for training and 20% of the data reserved for testing. This helps in evaluating the model on unseen data. The features are scaled using standard scaling. This is crucial for many machine learning algorithms that are sensitive to the scale of input data, like SVM. A linear kernel SVM is used within a Recursive Feature Elimination (RFE) wrapper to select the top 10 features. RFE works by recursively removing the least important features based on the model weights and retraining the model. A Bagging Classifier,

which is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset, is trained on the selected features. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1 score and it is depicted in Figure 7. These metrics provide a comprehensive understanding of the model's performance, especially in classification tasks. The model's probability estimates for the positive class are used to plot a Receiver Operating Characteristic (ROC) curve, which illustrates the diagnostic ability of a binary classifier. The Area Under the Curve (AUC) score for the ROC curve is calculated, which measures the entire two-dimensional area underneath the ROC curve and provides an aggregate measure of performance across all possible classification thresholds. Finally, the ROC curve is plotted with the AUC score and it is shown in Figure 8.

According to the experimental outcomes, the ensemble method, which consolidates predictions from multiple models, achieves an achieves of 99%, surpassing other machine learning algorithms and related studies [2,3]. The effectiveness of this proposed ensemble learning model is additionally confirmed through the receiver operating characteristic curve, with an ROC-AUC value of 1.00. These results show that the proposed ensemble model can precisely predict the risk of cardiac disease.

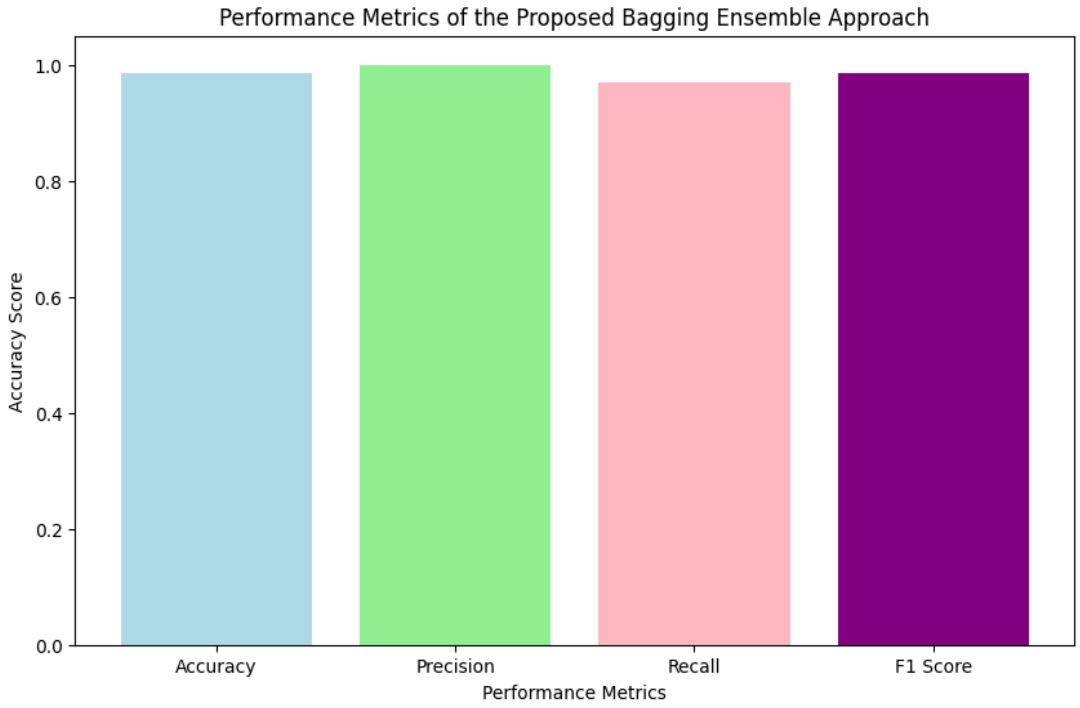

Figure 7. Performance Metrics of Bagging Ensemble Approach

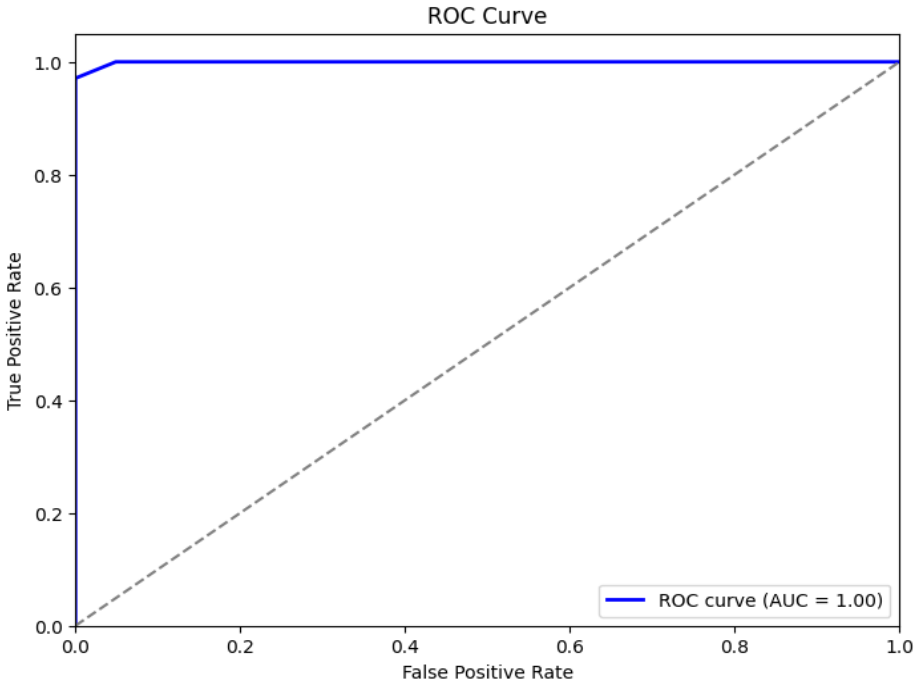

Figure 8. ROC – AUC Curve of Bagging Ensemble Approach

5. CONCLUSION

The research work underscores the effectiveness of integrating advanced machine learning techniques with ensemble modelling to enhance the predictive accuracy of cardiovascular disease diagnosis. The application of the Support Vector Machine - Recursive Feature Elimination (SVM-RFE) for feature selection effectively isolated the most significant predictors of cardiovascular risk. Following feature selection, the deployment of a Bagging Ensemble Classifier, leveraging diverse algorithms like k-Nearest Neighbors, Random Forests, and Decision Trees, significantly boosted the model's performance. The ensemble model achieved an accuracy of 99%, surpassing other comparable machine learning models and traditional analytical methods.

Additionally, the model's robustness was confirmed by an impressive ROC-AUC score of 1.00, validating its accuracy and reliability in clinical environments. These results advocate for the broader application of ensemble machine learning approaches in healthcare, promising substantial improvements in early detection and intervention strategies for cardiovascular disease, thereby potentially saving lives and enhancing patient outcomes.

References

[1] Muthulakshmi, P., Parveen, M., & Rajeswari, P. (2023). Prediction of Heart Disease using Ensemble Learning. Indian Journal of Science and Technology, 16(20), 1469-1476.

[2] Majumder, A. B., Gupta, S., & Singh, D. (2022, July). An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. In Journal of Physics: Conference Series (Vol. 2286, No. 1, p. 012017). IOP Publishing.

[3] Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20, 100402.

[4] Sharmila, R., & Chellammal, S. (2018). A conceptual method to enhance the prediction of heart diseases using the data techniques. International Journal of Computer Science and Engineering, 6(4), 21-25.

[5] Kumar, R., Fatima, S. S., & Thomas, A. (2020). Heart disease prediction using ensemble learning method. International Journal of Recent Technology and Engineering (IJRTE), 9.

[6] David, H. B. F. (2020). Impact of ensemble learning algorithms towards accurate heart disease prediction. ICTACT Journal on Soft Computing, 10(3), 2084-2089.

[7] Alim, M. A., Habib, S., Farooq, Y., & Rafay, A. (2020, January). Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model. In 2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET) (pp. 1-5). IEEE.

[8] Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(8), 165-174.

[9] Gupta, A., & Singh, A. (2022). An optimal multi-disease prediction framework using hybrid machine learning techniques: 10.48129/kjs. splml. 19321. Kuwait Journal of Science.

[10] Doppala, B. P., Bhattacharyya, D., Janarthanan, M., & Baik, N. (2022). A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. Journal of Healthcare Engineering, 2022.

[11] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download

[12] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & electrical engineering, 40(1), 16-28.

[13] Hari Priya, N., & Rajeswari, S. Covid-19 Prediction Using Enhanced KNN Imputation for Data Pre-Processing.

[14] Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences, 12(3), 246-269.

[15] Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: Random Forest classifier approach. In Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021 (pp. 219-244). Springer Singapore.