ISSN: 2633-4828 Vol. 5 No. S2, (Mar – Apr 2023)

International Journal of Applied Engineering & Technology

ENHANCING CORONARY ARTERY DISEASE PREDICTION THROUGH HYBRID FEATURE SELECTION METHOD

Sudipta Priyadarshinee[1] and Madhumita Panda[2*]
[1]School of Computer Science, G.M. University, Sambalpur, Odisha, India
E-mail:sudiptapatel88@gmail.com

[2]School of Computer Science, G.M. University, Sambalpur, Odisha, India
E-mail: mpanda.gmu@gmail.com
[*]Corresponding author : Madhumita Panda

ABSTRACT
Many recent research has prioritised early detection of coronary artery disease (CAD), which is the primary cause of cardiac-related fatalities worldwide. The effectiveness of essential illness diagnostic features has a substantial impact on machine learning system performance, allowing for quick and accurate therapy. Using the Z-Alizadeh Sani dataset, we used a hybrid feature selection framework with a machine learning method to categorise patients with CAD. Our methodology consists of outlier reduction, data synthesis with CTGAN, and class rebalancing with SMOTE. We develop FSRFE, a hybrid feature selection approach that integrates Forward Selection and Recursive Feature Elimination (RFE), and assess it against four machine learning methods: XGBoost, Random Forest and Extra Trees. When compared to traditional feature selection methods, our hybrid approach outperforms them. Our findings demonstrate the effectiveness of improved feature selection in enhancing predicting accuracy for heart disease, providing important insights into cardiovascular healthcare.

Keywords- CAD, FSRFE, SMOTE, CTGAN

1. INTRODUCTION
According to a World Health Organisation (WHO) research, coronary artery disease (CAD) is responsible for around 32% of global deaths, with a projected 23.6 million fatalities by 2030. CAD occurs when the big arterial vessels that supply the heart are damaged by atherosclerosis, a disorder characterised by the accumulation of fatty deposits known as atheroma within artery walls, resulting in vessel narrowing and congestion [1-3]. Detecting CAD early is difficult because it often appears after arteries are already clogged or during a cardiac attack. CAD and cardiac attacks are the most prevalent types of cardiovascular illnesses (CVD). CAD is usually diagnosed with specialised medical technology and skills, with popular approaches including blood testing and electrocardiograms (ECGs). However, these techniques are expensive and may not be economically viable, especially in low- and middle-income countries, where more than three-quarters of CAD cases occur. As a result, people in these areas may develop CAD at a younger age due to a delayed or poor diagnosis [4]. To overcome this issue, researchers investigated computer-aided systems in healthcare. By utilising information technologies, it is possible to design cost-effective CAD diagnosis procedures based on certain factors rather than expensive diagnostic devices. Such projects seek to improve access to CAD diagnosis, particularly in resource-limited areas, thereby reducing the disease's impact on affected communities.

Machine learning is widely applied in a variety of fields today, including healthcare. Medical data can be analysed using data mining and machine learning approaches to improve services and interactions with users, consumers, and patients [5-7]. These approaches have demonstrated promise in predicting illnesses such as Parkinson's disease, heart disease, liver disease, breast cancer, and lung cancer, with the goal of providing speedy and precise diagnosis with few errors. Overall, machine learning and data mining can improve diagnostic accuracy and provide high-quality healthcare services to patients [8]. Feature selection procedures are critical for model performance in complicated cardiac illness datasets, since they improve generalisation and predictability by reducing irrelevant or redundant characteristics. [9-10].

2 RELATED WORK

This review summarises current research on cardiovascular disease prediction and diagnosis with machine learning and data mining, including feature selection, classification methods, datasets, and performance measures.

This study [11] compares classification models on the Cleveland heart disease dataset, examining ten feature selection strategies such as ANOVA and Chi-square, as well as six approaches of machine learning such as decision trees and support vector machines. Among these, the backward feature selection methods have the maximum classification accuracy of 88.52% using a decision tree. This study [12] included classification approaches such as Logistic Regression and Random Forest, as well as feature selection approaches such as Pearson Correlation and Chi-Square, with the goal of increasing accuracy while decreasing execution time. Using the Statlog Heart Disease dataset, the study obtains an outstanding 84% accuracy using Logistic Regression, exceeding other methods. This paper [13] presents a hybrid Genetic-based Crow Search Algorithm (GCSA) for feature selection and classification that incorporates deep convolutional neural networks. The outcomes show that our GCSA model greatly improves classification performance, obtaining over 94% accuracy when compared to previous feature selection approaches on the Heart-C dataset, which has 15 attributes. This research [14] examines various data mining methods for predicting coronary heart disease utilising datasets from the UCI, such as Cleveland (303 samples, 14 characteristics) and Statlog Cleveland+Hungarian (1191 samples, 12 attributes). Feature selection techniques such as PCA and Chi Square improve classic machine learning methods. Random Forest with PCA has the highest accuracy of 92.85% in heart disease categorization. This study [15] used the Z-Alizadeh Sani dataset and five machine learning algorithms to predict coronary artery disease (CAD). Among the models, the multi-layer perceptron (MLP) had the highest accuracy at 90%. This study [16] examines heart failure survivors from a dataset of 299 patients in order to uncover key characteristics and successful data mining approaches for predicting patient survival. Nine classification models are used: Decision Tree, AdaBoost, Logistic Regression, SGD, Random Forest, GBM, ETC, G-NB, and SVM. SMOTE is used to remedy the imbalance in class. With SMOTE, ETC obtains the maximum accuracy of 0.9262, beating all other models. This study [17] uses the Cleveland heart disease dataset and employs data mining techniques such as regression and classification. It investigates machine learning techniques such as Random Forest and Decision Tree, as well as a novel hybrid model that combines the two approaches. Experimental results show that the hybrid model predicts heart disease with an accuracy of 88.7%. This study [18] employed machine learning approaches to detect CAD early in the Z-Alizadeh Sani dataset. After using Pearson correlation for feature selection, eight vital characteristics were identified. Among the six ML models examined, logistic regression and SVM both scored 95.45% accuracy, 95.91% sensitivity, 91.66% specificity, and 96.90% F1 score. This article [19] describes a sequential feature selection (SFS) approach for detecting deaths in heart disease patients during therapy. It compares the accuracy of the SFS approach against that of other machine learning approaches (LDA, GBC, RF, SVM, DT and KNN). The SFS technique, notably with the Random Forest Classifier _FS model, yields an accuracy of 86.67%, demonstrating its effectiveness in feature selection for this application. This work [20] developed a novel optimisation method called the N2Genetic optimizer, which aims to improve genetic training techniques. In trials using this approach, N2Genetic-nuSVM achieved a remarkable accuracy of 93.08% and an F1-score of 91.51% in predicting CAD outcomes on the Z-Alizadeh Sani dataset, a highly recognised benchmark dataset for such investigations. This study [21] employed the Stability Selection approach to identify key features of coronary artery disease. Age, blood pressure, and diabetes were all identified as important factors. Logistic regression surpassed other machine learning techniques, with 90.88% accuracy, 95.18% sensitivity, and 81.34% specificity. This paper [22] proposed a hybrid model (PSO-EmNN) for CAD diagnosis based on four feature selection approaches using the Z-Alizadeh Sani dataset. The top 22 characteristics picked by the WBSVM technique produced the best results, with accuracy, precision, sensitivity, specificity, and F1-score reaching 88.34%, 92.37%, 91.85%, 78.98%, and 92.12%, respectively.

Numerous studies have investigated machine learning-based cardiovascular disease (CAD) diagnosis, although standardisation is lacking due to dataset discrepancies. Recent research frequently uses datasets with little

features, making feature selection (FS) approaches ineffective. To overcome this gap, the proposed study would employ the Z-Alizadeh Sani dataset [23], which has 54 attributes. It strives to improve FS efficacy by using a hybrid FS method that combines old and unique methodologies.

3. PROPOSED METHODOLOGY

This work provides a thorough strategy for improving heart disease prediction using machine learning algorithms. It starts with comprehensive data preprocessing, using Isolation Forest [24] for outlier removal and CTGAN [25] for data synthesis. Class rebalancing strategies such as SMOTE [26] address class imbalances, resulting in a more representative training dataset. A new hybrid feature selection strategy, FSRFE, which combines Forward Selection and Recursive Feature Elimination (RFE) techniques, is given. This methodology outperforms existing methods, improving the projected accuracy of the Random Forest, XGBoost, and Extra Trees algorithms. A comparison with traditional feature selection methodologies demonstrates the effectiveness of the proposed method.

3.1. Dataset

The Z-Alizadeh Sani dataset from Kaggle [23] comprises 303 patient records with 216 CAD patients and 87 healthy individuals, encompassing 54 variables. Its goal is to estimate coronary artery diameter stenosis based on angiography results, with CAD patients designated as 1 and healthy individuals labelled as 0 for machine learning classification.

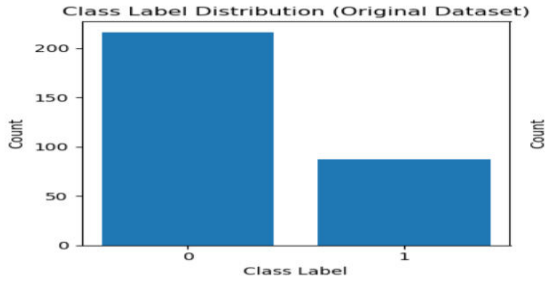


Fig. 1. Class Label Distribution of Dataset

3.2 Data Preprocessing

We used rigorous data preprocessing to improve dataset quality and integrity. We used cutting-edge approaches, including the Isolation Forest algorithm [24] to remove outliers and maintain dataset stability. Furthermore, we used the Conditional Tabular Generative Adversarial Network (CTGAN) [25] to produce 500 additional data points, increasing dataset variety. To address class imbalance concerns, we used the Synthetic Minority Over-sampling Technique (SMOTE) [26] to rebalance class distributions and assure proper representation of each class. These preprocessing procedures produced a high-quality dataset suitable for analysis and model training, allowing for accurate categorization of patients with coronary artery disease (CAD) using the Z-Alizadeh Sani dataset.

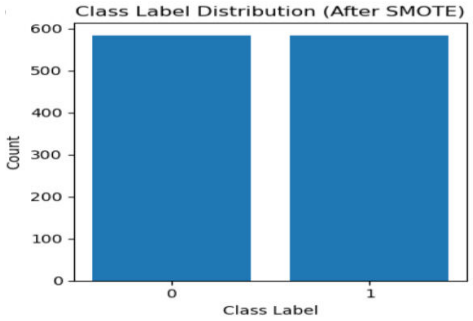


Fig. 2 Class Label Distribution of Dataset after SMOTE

3.3. Feature Selection

Feature selection is essential for reducing dataset dimensionality, which provides advantages such as increased learning performance and reduced processing costs. Strategies for categorical variables include the chi-square test [27], Recursive Feature Elimination (RFE) [28], forward selection [29], and Random Forest Importance, a tree-based feature selection approach [30]. These strategies assist in identifying the most crucial components of the topic.

3.3.1. Proposed Algorithm: Forward Selection and Recursive Feature Elimination (FSRFE)

The FSRFE (Forward Selection and Recursive Feature Elimination) algorithm combines forward selection and recursive feature elimination approaches to improve feature selection. It combines the advantages of both approaches: Forward Selection adds features sequentially to improve performance, whereas RFE systematically removes less important characteristics. This hybrid technique ensures a balanced collection of characteristics, which enhances model performance.

Inputs:
- X: Input feature matrix

- y: Target variable

- n_features_to_select: Number of features to select

- model: Machine learning model

- scoring: Scoring metric

Output:
- Selected feature indices

1. Perform forward selection:

a. Initialize an empty set to store selected feature indices: selected_features = { }

b. Add features sequentially to selected_features based on performance improvement.

c. Repeat until the desired number of features is selected.

2. Perform recursive feature elimination (RFE):

a. Initialize an empty set to store selected feature indices: selected_features_rfe = { }

b. Remove features iteratively to maintain performance.

c. Repeat until the desired number of features is selected.

3. Take the union of selected features from forward selection and RFE:

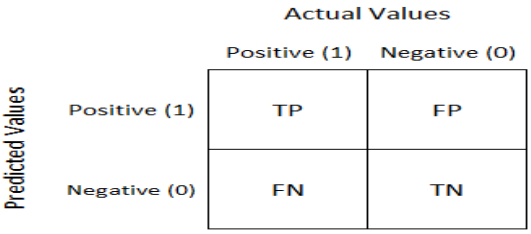
4. Return selected_features_union

4. EXPERIMENTAL RESULTS AND ANALYSIS

The work used the Z-Alizadeh Sani dataset to assess a proposed machine learning approach using four performance metrics: confusion matrix, accuracy, precision, recall, and F1 score (Equations 1-4) [31].

4.1 Confusion Matrix

The performance of the models was examined using a confusion matrix. In this context, True Positive (TP) means correctly diagnosed sickness, True Negative (TN) indicates accurately identified lack of disease, False Positive (FP) denotes mistakenly classified positives, and False Negative (FN) denotes incorrectly categorised negatives.



The classifier's accuracy (Acc) is computed by dividing the number of accurate predictions. it made during the testing phase by the total number of predictions it made.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} * 100\% \quad (1)$$

Precision (Pre) is the percentage of events correctly classified as positive by a classifier compared to the total expected positives, indicating the classifier's accuracy.

$$\text{Pre} = \frac{TP}{(TP+FP)} * 100\% \quad (2)$$

Recall is computed by dividing the total number of relevant samples by the number of true positive results.

$$Recall = \frac{TP}{(TP+FN)} * 100\% \quad (3)$$

F1-score is the harmonic mean of precision and recall, indicating a balanced combination.

$$F1 - Score = \frac{2*TP}{2*(TP+FN+FP)} * 100\% \quad (4)$$

4.2. Results and Discussion

The experiments were carried out using the Python3 programming language, Anaconda backend, and Jupyter Notebook environment.

We used our proposed hybrid feature selection FSRFE in conjunction with machine learning approaches such as Random Forest [32] Extra tree [33] and Xgboost [34] to improve performance, evaluating accuracy, precision, recall, F1 score, and confusion matrix. The dataset was divided into 80:20 training and testing sets to facilitate model training and evaluation. The results in Table 1 summarise the performance outcomes of machine learning methods after data preprocessing, whereas Table 2 shows FSRFE's superiority over standard feature selection algorithms. These findings highlight our methodology's ability to effectively categorise individuals with coronary artery disease, pointing to potential advances in cardiovascular healthcare through enhanced prediction accuracy.

Table 1. Performance of ML Classifier after Data Preprocessing

Parameters	Xgboost	Extratree	Random Forest
Testing Acc	94.02	94.87	94.02
Precision	96	97	95
Recall	93	94	94
F1-Score	94	95	94

Table 1 shows that the Extra Tree classifier outperformed all other models in terms of accuracy (94.87%), precision (97%), recall (94%), and F1 score (95%).

Table 2. Performance Comparison of ML Classifier with Hybrid vs. Traditional Feature Selection Methods

Parameters	Random forest					Extratree					Xgboost				
	Chi-square test	RFE	FS	RF	Proposed Feature Selection	Chi-square test	RFE	FS	RF	Proposed Feature Selection	Chi-square test	RFE	FS	RF	Proposed Feature Selection
Testing Acc	94.8	96.5	96	95.7	97.01	96.1	96.1	94	94.8	97.44	95.3	95.7	92	94.8	97.01
Precision	93	99	98	99	98	96	100	97	98	98.4	96	98	97	97	96.88
Recall	98	94	94	93	95.28	97	93	91	92	96.06	95	94	88	94	97.64
F1-Score	95	97	96	96	97.19	96	96	94	95	97.6	96	96	92	95	97.55
No of feature	40	38	13	10	50	40	38	13	10	50	40	38	13	10	50

In our study, novel feature selection approaches such as FSRFE considerably improved CAD prediction accuracy, especially when combined with the Extra Trees algorithm, providing remarkable metrics: 97.44% accuracy, 98.4% precision, 96.06% recall, and 97.6% F1 score, as shown in Table 2.

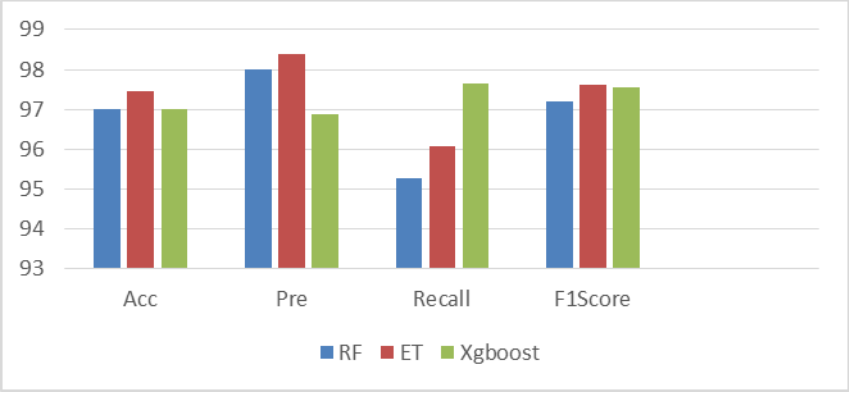


Fig. 3 performance comparison of ML classifiers that use hybrid feature selection methods

Figure 3 depicts a performance comparison of three machine learning classifiers that use hybrid feature selection methods. Classifier performance is measured using metrics such as accuracy, precision, recall, and F1 score.

Table 3 Confusion Matrix for the RF Algorithm during testing

Actual	Predicted	
	Normal	CAD
Normal	**106**	**1**
CAD	**6**	**121**

Table 4 Confusion Matrix for the ET Algorithm during testing

Actual	Predicted	
	Normal	CAD
Normal	**106**	**1**
CAD	**5**	**122**

Table 5 Confusion Matrix for the Xgboost Algorithm during testing

Actual	Predicted	
	Normal	CAD
Normal	103	4
CAD	3	124

After analysing the confusion matrix for all classifiers, it was found that the Extra Trees classifier had only one false negative and five false positives, demonstrating its superior performance in correctly identifying genuine negatives while minimizing false positives and false negatives. This confirms the model's reliability in identifying individuals with coronary artery disease (CAD).

5. COMPARATIVE STUDY WITH OTHER RECENTLY SCHOLARLY WORKS

In this section, we compare our proposed work to existing research, demonstrating that our method obtained more accuracy than others.

Table 6. A comparison of the proposed model with various existing approaches

Source	Year	Dataset	Approach	Parameters
Abdar et al. [20]	2019	Z-Alizadeh Sani dataset	N2Genetic-nuSVM	Acc-93.08
Nazlı, Bahar et al. [15]	2020	Z-Alizadeh Sani dataset	MLP	Acc-90
Shahid, Afzal Hussain et al. [22]	2020	Z-Alizadeh Sani dataset	PSO-EmNN	Acc-88.34
Akyol, Kemal et al [21]	2021	Z-Alizadeh Sani dataset	Logistic Regression	Acc-90.88
Sayadi, Mohammadjavad, et al. [18]	2022	Z-Alizadeh Sani dataset	Logistic Regression &SVM	Acc-95.45
Proposed Work	**2023**	**Z-Alizadeh Sani dataset**	**Extra Tree with FSRFE feature selection**	**Acc-97.44**

6. CONCLUSION

In the conclusion, we employ machine learning to improve the prediction accuracy of coronary artery disease (CAD). We introduced a new hybrid feature selection approach, FSRFE, and compared it to established methods. Our results reveal that FSRFE paired with the Extra Trees algorithm has the maximum accuracy of 97.44%. This emphasises the relevance of improved feature selection in improving CAD prediction and provides useful insights for cardiovascular healthcare. Combining hybrid feature selection approaches with machine learning can result in more accurate forecasts and improved treatment for CAD patients.

REFERENCES

1. Wah, T. Y., Gopal Raj, R., & Iqbal, U. (2018). Automated Diagnosis of Coronary Artery Disease: A Review and Workflow. Cardiology research and practice, 2018.
2. Cardiovascular diseases (CVDs), http://www.who.int/news-room/factsheets/detail/cardiovascular- diseases-(cvds), [accessed on 11/8/2018].

3. Cüvitoğlu, Ali, and Zerrin Işik. "Classification of CAD dataset by using principal component analysis and machine learning approaches." *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*. IEEE, 2018.
4. Alizadehsani, Roohallah, et al. "Machine learning-based coronary artery disease diagnosis: A comprehensive review." *Computers in biology and medicine* 111 (2019): 103346.
5. Parikh, R. B., Kakad, M., & Bates, D. W. (2016). Integrating predictive analytics into high-value care: the dawn of precision delivery. Jama, 315(7), 651-652.
6. Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine learning and the profession of medicine.Jama, 315(6), 551-552.
7. Ahsan, Md Manjurul, and Zahed Siddique. "Machine learning-based heart disease diagnosis: A systematic literature review." *Artificial Intelligence in Medicine* 128 (2022): 102289.
8. Verma, Luxmi, Sangeet Srivastava, and P. C. Negi. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data." *Journal of medical systems* 40 (2016): 1-7.
9. Shilaskar S, Ghatol A. Feature selection for medical diagnosis:Evaluation for cardiovascular diseases. Expert Systems withApplications 2013;40(10):4146–53
10. 11.Shao YE, Hou C-D, Chiu C-C. Hybrid intelligent modelling schemes for heart disease classification. Applied Soft Computing 2014; 14:47–52.
11. Dissanayake, Kaushalya, and Md Gapar Md Johar. "Comparative study on heart disease prediction using feature selection techniques on classification algorithms." *Applied Computational Intelligence and Soft Computing* 2021 (2021): 1-17.
12. Singh, Priya, Gyanendra Kumar Pal, and Sanjeev Gangwar. "Prediction of cardiovascular disease using feature selection techniques." *International Journal of Computer Theory and Engineering* 14.3 (2022): 97-103.
13. Nagarajan, Senthil Murugan, et al. "Innovative feature selection and classification model for heart disease prediction." *Journal of Reliable Intelligent Environments* 8.4 (2022): 333-343.
14. Tasnim, Farzana, and Sultana Umme Habiba. "A comparative study on heart disease prediction using data mining techniques and feature selection." *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE, 2021.
15. Nazlı, Bahar, Yasemin Gültepe, and Hayriye Altural. "Classification of Coronary Artery Disease Using Different Machine Learning Algorithms." *International Journal of Education and Management Engineering (IJEME)* 10.4 (2020): 1-7.
16. Ishaq, Abid, et al. *"Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques." IEEE access* 9 (2021): 39707-39716.
17. Kavitha, M., et al. "Heart disease prediction using hybrid machine learning model." *2021 6th international conference on inventive computation technologies (ICICT)*. IEEE, 2021.
18. Sayadi, Mohammadjavad, et al. "A machine learning model for detection of coronary artery disease using noninvasive clinical parameters." *Life* 12.11 (2022): 1933.
19. Aggrawal, Ritu, and Saurabh Pal. "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease." *SN Computer Science* 1.6 (2020): 344.

20. Abdar, Moloud, et al. "A new machine learning technique for an accurate diagnosis of coronary artery disease." *Computer methods and programs in biomedicine* 179 (2019): 104992.
21. Akyol, Kemal. "Feature selection based data mining approach for coronary artery disease diagnosis." *Academic Platform-Journal of Engineering and Science* 9.3 (2021): 451-459.
22. Shahid, Afzal Hussain, and M. P. Singh. "A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network." *Biocybernetics and Biomedical Engineering* 40.4 (2020): 1568-1585.
23. https://www.kaggle.com/datasets/tanyachi99/zalizadeh-sani-dataset-2csv
24. Cheng, Zhangyu, Chengming Zou, and Jianwei Dong. "Outlier detection using isolation forest and local outlier factor." *Proceedings of the conference on research in adaptive and convergent systems*. 2019.
25. Shourmasti, Elaheh Shahmir. *Generating Synthetic Health Data Using Machine Learning GAN Methods*. MS thesis. 2022.
26. Priyadarshinee, Sudipta, and Madhumita Panda. "Improving prediction of chronic heart failure using smote and machine learning." *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*. IEEE, 2022.
27. Spencer, Robinson, et al. "Exploring feature selection and classification methods for predicting heart disease." *Digital health* 6 (2020): 2055207620914777.
28. Theerthagiri, Prasannavenkatesan, and Jyothiprakash Vidya. "Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques." *Expert systems* 39.9 (2022): e13064.
29. Khaire, Utkarsh Mahadeo, and R. Dhanalakshmi. "Stability of feature selection algorithm: A review." *Journal of King Saud University-Computer and Information Sciences* 34.4 (2022): 1060-1073.
30. Akhiat, Yassine, et al. "A new noisy random forest based method for feature selection." *Cybernetics and Information Technologies* 21.2 (2021): 10-28.
31. Terrada, Oumaima, et al. "Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques." *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*. IEEE, 2020.
32. Singh, Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh. "Heart disease prediction system using random forest." *Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11-12, 2016, Revised Selected Papers 1*. Springer Singapore, 2017.
33. Shafique, Rahman, Arif Mehmood, and Gyu Sang Choi. "Cardiovascular disease prediction system using extra trees classifier." (2019).
34. Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective prediction of heart disease." *Journal of King Saud University-Computer and Information Sciences* 34.7 (2022): 4514-4523.