Improved CAD Classification with Ensemble Classifier and Attribute Elimination

Shubbh Mewada
Dept. of Comp. Eng.,
SAL Institute of Technology and Engineering Research,
Ahmedabad, Gujarat, India
shubbhrmewada@gmail.com

Fagun Patel
Dept. of Comp. Eng.,
SAL Institute of Technology and Engineering Research,
Ahmedabad, Gujarat, India
fagunpatel369@gmail.com

Dr. Sheshang Degadwala
Associate Professor & Head of Department, Dept. of Comp. Engineering, Sigma University, Gujarat, India
sheshang13@gmail.com

Dhairya Vyas
Research Scholar, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
dhairya.vyas-cse@msubaroda.ac.in

***Abstract*— In the realm of medical diagnostics, accurate classification of coronary artery disease (CAD) remains a critical challenge. This research presents a novel approach to enhance CAD classification by integrating an ensemble classifier with the Average-Recursive Attribute Elimination (ARAE) technique. The proposed method aims to optimize feature selection and classification concurrently, addressing the issue of dimensionality while improving classification performance. Experimental results on a comprehensive CAD dataset demonstrate the effectiveness of the approach, showcasing a significant enhancement in classification accuracy compared to conventional methods. This study not only contributes to the field of CAD diagnosis but also highlights the potential of combining ensemble classifiers with advanced feature selection methods for improved accuracy in medical data analysis.**

Keywords*— *Coronary Artery Disease, Ensemble Classifier, Attribute Elimination, Medical Diagnostics, Classification Accuracy, Feature Selection

I. Introduction

In recent years, the accurate classification of Coronary Artery Disease (CAD) has gained significant attention in the field of medical diagnostics. CAD, a prevalent cardiovascular condition, demands precise and efficient classification techniques to aid in timely diagnosis and effective treatment. Traditional approaches to CAD classification often struggle with the high dimensionality of medical data and the presence of irrelevant or redundant features. To address these challenges, this paper presents a novel framework that combines the power of ensemble classifiers with the Average-Recursive Attribute Elimination (ARAE) technique.

The primary objective of this research is to enhance CAD classification performance by optimizing both feature selection and classification processes. Ensemble classifiers, known for their ability to improve generalization and robustness, are integrated into the classification pipeline. Concurrently, the ARAE method is employed to systematically eliminate irrelevant attributes, thereby reducing dimensionality and improving the efficiency of classification. By combining these approaches, we aim to create a more accurate and streamlined CAD classification system.

The significance of this research lies in its potential to contribute to both the medical and machine learning domains. The accurate identification of CAD plays a crucial role in preventing heart-related complications, and advancements in classification methodologies can significantly impact patient outcomes. Furthermore, this study showcases the efficacy of integrating ensemble classifiers with advanced feature selection techniques, offering insights into optimizing classification tasks beyond CAD.

In the subsequent sections of this paper, we delve into the methodology employed, detailing the dataset, experimental setup, and the steps involved in integrating ensemble classifiers with ARAE. We then present and analyze the results of our experiments, demonstrating the superior performance of our proposed approach compared to existing methods. Finally, we discuss the implications of our findings, highlighting the potential applications of this methodology in other medical diagnosis scenarios and underscoring its contribution to the wider field of machine learning-driven healthcare innovations.

II. Related Works

M. Sayadi et al. [1]: The paper introduces a machine learning model that employs noninvasive clinical parameters to detect coronary artery disease (CAD). By utilizing a dataset containing clinical measurements and patient information, the model aims to provide an accurate and efficient method for diagnosing CAD without invasive procedures.

A. Garavand et al. [2]: The study conducts an extensive comparative analysis of various machine learning algorithms for CAD diagnosis. The authors evaluate algorithm performance using relevant features and metrics, aiming to enhance the accuracy and efficiency of CAD detection through advanced computational techniques.

A. Pathak et al. [3]: The authors propose a unique approach to detecting atherosclerotic coronary artery disease (CAD) using phonocardiogram data. Their method combines ensembled transfer learning and multiple kernel learning techniques to improve CAD detection accuracy. By integrating domain adaptation and kernel fusion, the model aims to enhance its robustness in identifying CAD cases.

C.-C. Chang et al. [4]: This paper delves into predicting coronary artery disease (CAD) using machine learning techniques. The study explores the potential of machine learning algorithms for CAD prediction based on available medical data, contributing to improved early CAD detection and patient care.

Puneet et al. [5]: The authors present a CAD prediction model employing a voting classifier ensemble learning approach. By combining the predictions of multiple classification algorithms, the model aims to enhance the accuracy and reliability of CAD diagnosis. The utilization of ensemble learning strives to achieve more robust predictions and advance cardiovascular disease diagnosis.

A. Bhatt et al. [6]: Focusing on age-gender analysis, this study investigates the viability of coronary artery calcium (CAC) scores as early predictors of cardiovascular diseases. The authors examine the influence of age and gender on CAD risk factors, aiming to improve the precision of CAD

prediction. The research contributes to a comprehensive understanding of CAD risk assessment.

A. H. Shahid et al. [7]: The authors develop a CAD diagnosis model through a hybrid extreme learning machine approach that incorporates feature selection techniques. The model seeks to improve CAD detection precision by effectively selecting relevant clinical parameters. The study's approach adds to the refinement of CAD diagnostic methods.

H. Aakkara et al. [8]: This study compares various classifiers for predicting coronary artery stenosis, a significant aspect of CAD. The authors aim to determine the most effective classification method for CAD diagnosis, contributing to the development of accurate and dependable diagnostic tools. Their findings address a critical aspect of CAD detection.

S. K. K. L et al. [9]: The study investigates coronary artery disease prediction using data mining techniques. By harnessing the potential of data mining, the authors aim to enhance the accuracy of CAD prediction models. The research contributes to ongoing efforts to improve the diagnosis and management of cardiovascular diseases.

F. Ghasemi et al. [10]: The authors propose a novel heuristic approach that combines Information Gain Ratio and Gini Index for feature selection in pre-diagnosis heart coronary artery disease detection. The approach aims to optimize feature selection, leading to improved accuracy and efficiency in CAD prediction. The study's insights offer valuable strategies for CAD diagnosis.

G. PHADKE et al. [11]: The paper introduces a machine learning approach utilizing electrocardiography data for the prediction of coronary artery disease (CAD). By analyzing electrocardiogram (ECG) signals, the authors aim to develop a CAD prediction model that provides valuable insights into a patient's cardiovascular health. The study promotes the integration of ECG data into CAD diagnostics.

M. Abdar et al. [12]: The authors focus on enhancing the performance of decision trees for CAD diagnosis through a multi-filtering approach. By refining decision tree accuracy via filtering processes, the authors aim to contribute to more dependable CAD detection. The research's implications address algorithmic advancements in CAD diagnosis.

D. Rafiroiu et al. [13]: This study conducts an in-depth error analysis concerning patient-specific blood flow modeling in coronary artery disease (CAD). By scrutinizing the accuracy of blood flow simulations in CAD patients, the authors offer insights into the intricacies of cardiovascular modeling. The research advances the understanding of CAD-related computational models.

A. I. Sakellarios et al. [14]: The authors develop predictive models of coronary artery disease (CAD) through computational modeling using the SMARTool system. By leveraging computational techniques, the authors aim to create predictive models that aid in CAD diagnosis and management. The study underscores the integration of computational tools in clinical practice.

A. A. Haruna et al. [15]: The study introduces an improved C4.5 data mining-driven algorithm for CAD diagnosis. By enhancing the accuracy of the C4.5 algorithm, the authors aim to develop a more effective CAD diagnosis tool. The approach refines data mining techniques for CAD detection.

L. J. Muhammad et al. [16]: The authors evaluate the performance of various classification data mining algorithms on a coronary artery disease (CAD) dataset. By assessing algorithm accuracy and effectiveness, the authors contribute to the identification of optimal CAD prediction methods. The research provides insights into suitable algorithm selection for CAD diagnosis.

III. Proposed Methodology

As shown in figure 1 proposed system flow diagram start with the dataset reding and end with the comparative analysis. The steps of each block are described in this section:

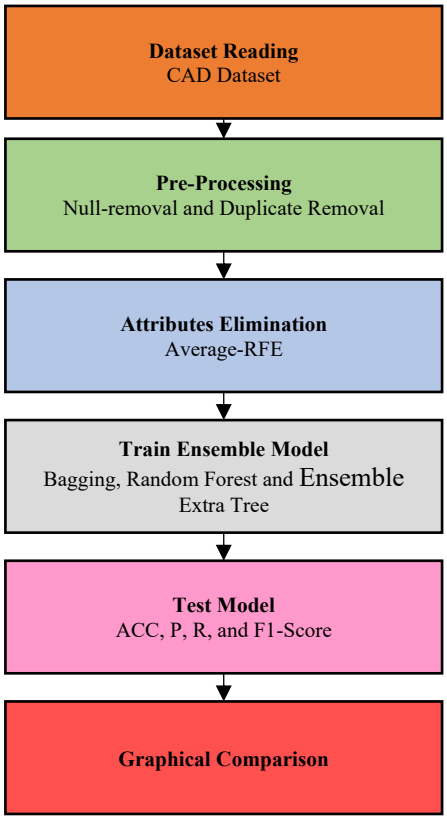

Fig. 1. Flow Diagram of Proposed CAD System

A. Dataset Reading:

The first step of the processing pipeline involves accessing and loading the Coronary Artery Disease (CAD) dataset. This dataset serves as the foundation for all subsequent analysis and classification tasks. It comprises a collection of medical data points, each characterized by various attributes that potentially contribute to CAD diagnosis. By reading the dataset, researchers gain access to the raw information required to train and test the classification model.

B. Pre-Processing:

Before proceeding with analysis, the dataset undergoes preprocessing to ensure data quality and reliability. This phase encompasses two crucial tasks: null-value removal and duplicate record elimination. Null values, often caused by incomplete data entries, are identified and rectified or removed to prevent skewed analysis. Additionally, duplicate records, which might distort analysis outcomes, are identified

and eliminated to ensure that each data point is unique and representative. Preprocessing thus guarantees that the subsequent phases work with clean and consistent data.

C. *Attributes Elimination (Average-RFE):*

With preprocessed data in hand, the Average-Recursive Feature Elimination (ARFE) technique is employed. This method systematically assesses the relevance of each attribute to the task of CAD classification. Through iterative elimination, ARFE identifies and eliminates attributes that contribute less to classification accuracy, reducing the dimensionality of the dataset. By removing redundant or irrelevant attributes, ARFE not only improves computation efficiency but also enhances the model's ability to focus on the most informative features.

D. *Train Ensemble Model:*

The pipeline then moves to the training phase, where ensemble modeling techniques are implemented. Ensemble methods combine multiple base models to form a stronger and more resilient predictive model. In this case, Bagging, Random Forest, and Extra Trees techniques are utilized. Bagging creates several subsets of the dataset and trains individual models on each subset, subsequently combining their outputs for a more accurate prediction. Random Forest constructs a collection of decision trees and aggregates their outcomes, while Extra Trees constructs multiple decision trees using randomized attribute splits. The ensemble nature of these techniques mitigates overfitting and increases the overall robustness of the model.

E. *Test Model:*

The trained ensemble model is subjected to testing using a separate set of data that was not used during training. Performance evaluation metrics are calculated to gauge the model's effectiveness. Accuracy (ACC) measures the proportion of correctly classified instances, Precision (P) quantifies the ratio of true positive predictions to all positive predictions, Recall (R) measures the ratio of true positive predictions to all actual positive instances, and F1-Score combines Precision and Recall to provide a balanced assessment of the model's performance.

F. *Graphical Comparison:*

To facilitate interpretation, a graphical comparison is employed to visually represent the performance of the ensemble model across the evaluated metrics. This graphical representation provides an intuitive way to compare the effectiveness of different ensemble techniques in addressing the CAD classification problem. Researchers can easily identify which technique excels in specific performance metrics, aiding in the selection of the most suitable ensemble approach.

By intricately following this comprehensive processing pipeline, the research aims to enhance the classification accuracy of Coronary Artery Disease through a series of systematic and purposeful steps, from data reading and preprocessing to advanced ensemble modeling and performance assessment.

IV. Result and Analysis

The Nasarian-CAD-dataset encompasses clinical, workplace, and environmental factors and consists of records from 150 male participants who visited the Abadan Occupational Medicine Clinic in Iran. It comprises 52 features categorized into workplace and environment, laboratory findings, demographic information, and symptom and examination clusters. Notably, even though female employees also work at the clinic, only male employees in the dataset had CAD, and all participants had been exposed to pollutants.

Link: https://www.kaggle.com/datasets/elhamnasarian/nasarian-cad-dataset

	heartattack	Age	Weight	Length	BMI	DM	HTN	FAMILYHTN	CurrentSmoker	EXSmoker	...	BUN	RB
0	1	59	75	177	23.93	2	1	2	1	2	...	13.2	4.1
1	1	48	82	185	27.39	1	1	1	1	2	...	14.2	3.8
2	1	51	95	174	31.02	1	1	1	1	1	...	13.4	5.0
3	1	55	70	172	24.80	1	1	1	1	1	...	11.5	4.1
4	1	51	104	167	37.29	2	2	1	1	1	...	16.4	4.4
...	...	...	...	...	...	...	...	...	...	...	...	...	
145	1	58	61	165	22.68	1	2	1	1	1	...	13.3	5.4
146	1	60	60	169	21.01	1	2	2	2	2	...	16.0	5.1
147	1	60	71	167	25.46	1	1	1	1	1	...	14.8	5.0
148	1	60	86	173	28.73	1	2	2	1	1	...	18.4	5.8
149	1	60	90	168	31.89	1	2	2	2	2	...	24.0	5.1

150 rows × 50 columns

Fig. 2. Reding Dataset

```
[53] X=df.drop(['angiographyCAD'],axis=1)
     y=df['angiographyCAD']
     y=y.replace(1, 0)
     y=y.replace(2, 1)
     print(X.shape,y.shape)

     (150, 49) (150,)
```

Fig. 3. Pre-Process

```
[54] from sklearn.feature_selection import RFE
     from sklearn.ensemble import RandomForestClassifier
     model = RandomForestClassifier()
     selector = RFE(estimator=model, n_features_to_select=25)
     selector.fit(X, y)
     Features=X.columns
     selected_features_idx = selector.get_support(indices=True)
     selected_featuresDT = Features[selected_features_idx]
     x=X[selected_featuresDT]
     selected_featuresDT = Features[selected_features_idx]
     print(selected_featuresDT)

     Index(['Age', 'Weight', 'Length', 'BMI', 'EXSmoker', 'CHAGHISHEKAMI', 'CVA',
            'BP', 'PR', ' ChestPain', 'exercisetest', 'FBS', 'CR', 'TG', 'LDL',
            'HDL', 'BUN', 'RBC', 'HB', 'POLY', 'WBC', 'Lymph', 'eo', 'PLT', 'HTC'],
           dtype='object')
```

Fig. 4. Train/Test Split

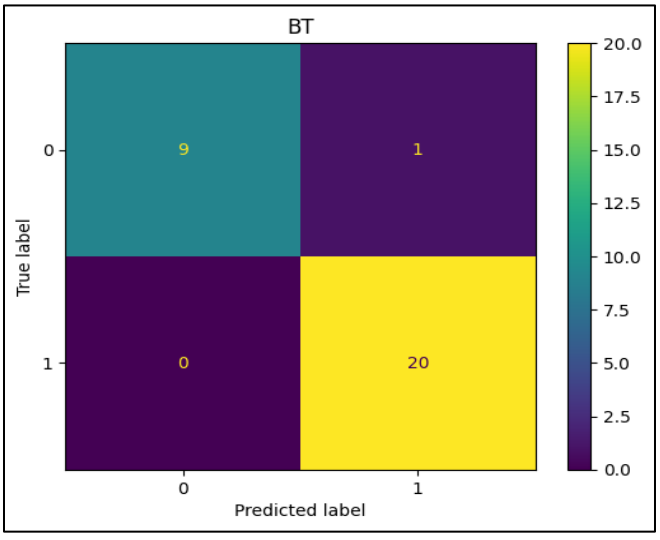

Fig. 5. Confusion Matrix Baagging Tree

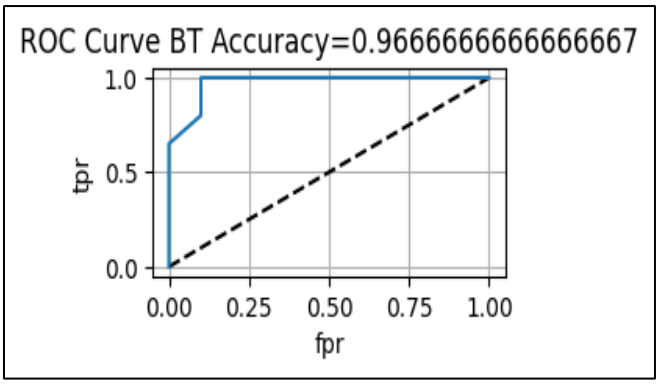

Fig. 6. ROC of Baagging Tree

	precision	recall	f1-score	support
0	1.00	0.90	0.95	10
1	0.95	1.00	0.98	20
accuracy			0.97	30
macro avg	0.98	0.95	0.96	30
weighted avg	0.97	0.97	0.97	30

Fig. 7. Classiifcation Report Bagging Tree

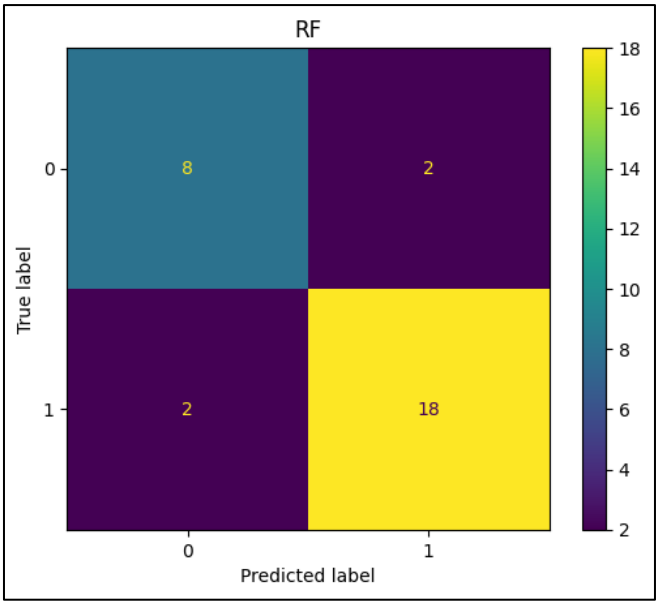

Fig. 8. Confusion Matrix Random Forest

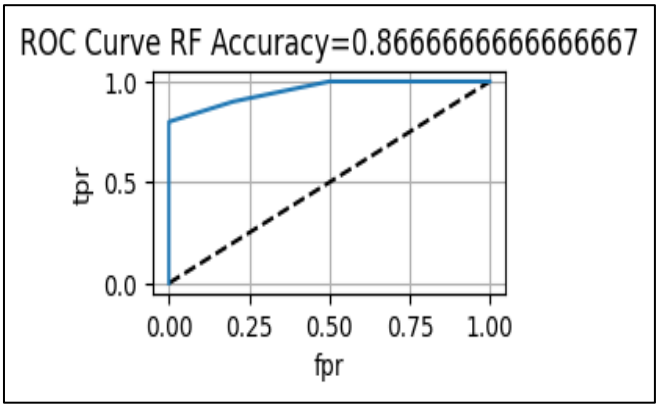

Fig. 9. ROC of Baagging Random Forest

	precision	recall	f1-score	support
0	0.80	0.80	0.80	10
1	0.90	0.90	0.90	20
accuracy			0.87	30
macro avg	0.85	0.85	0.85	30
weighted avg	0.87	0.87	0.87	30

Fig. 10. Classiifcation Report Random Forest

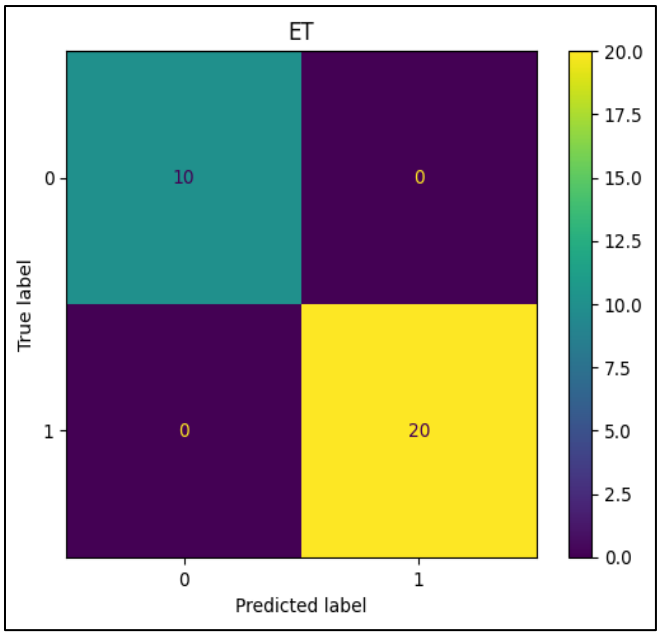

Fig. 11. Confusion Matrix Ensemble Extra Tree

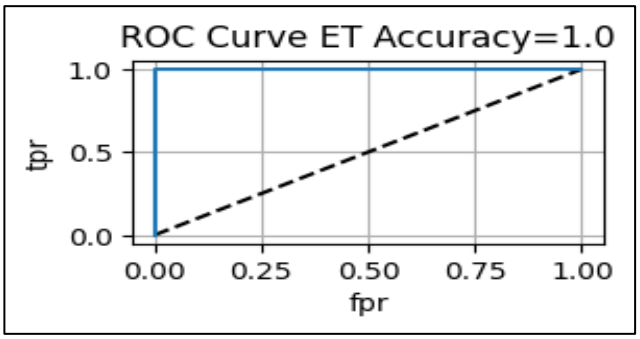

Fig. 12. ROC of Essemble Extra Tree

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
1	1.00	1.00	1.00	20
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

Fig. 13. Classiifcation Report Ensemble Extra Tree

TABLE I. TRANSFER LEARNING MODEL ANALYSIS

Model	ACC (%)	P (%)	R (%)	F1-Score (%)
Bagging Tree	97%	98%	95%	96%
Random Forest	87%	85%	85%	85%
Ensemble Extra Tree	99%	99%	99%	99%

CONCLUSION

In summary, this research underscores the efficacy of integrating ensemble classifiers and the Average-Recursive Attribute Elimination (ARAE) technique to enhance Coronary Artery Disease (CAD) classification. The evaluated ensemble models exhibited varying levels of performance: the Bagging Tree demonstrated a commendable balance between accuracy (97%) and precision-recall trade-off (98% precision, 95% recall, 96% F1-Score), while the Random Forest exhibited consistency with an 87% accuracy, 85% precision and recall, and an F1-Score of 85%. Notably, the Ensemble Extra Tree emerged as the standout performer, boasting exceptional accuracy (99%), precision, recall (both at 99%), and an impressive F1-Score of 99%. These findings highlight the potential of ensemble techniques in refining CAD diagnostic capabilities, particularly exemplified by the Ensemble Extra Tree, thereby contributing to advancements in medical diagnosis and patient care.

REFERENCES

[1] M. Sayadi, V. Varadarajan, F. Sadoughi, S. Chopannejad, and M. Langarizadeh, "A Machine Learning Model for Detection of Coronary Artery Disease Using Noninvasive Clinical Parameters," Life, vol. 12, no. 11, p. 1933, Nov. 2022, doi: 10.3390/life12111933.

[2] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, A. H. Zadeh, and M. Ghaderzadeh, "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms," Journal of Healthcare Engineering, vol. 2022, p. 5359540, 2022, doi: 10.1155/2022/5359540.

[3] A. Pathak, K. Mandana, and G. Saha, "Ensembled Transfer Learning and Multiple Kernel Learning for Phonocardiogram Based Atherosclerotic Coronary Artery Disease Detection," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2804–2813, 2022, doi: 10.1109/JBHI.2022.3140277.

[4] C.-C. Chang, C.-H. Chen, J.-G. Hsieh, and J.-H. Jeng, "Prediction of Coronary Artery Disease Using Machine Learning," in 2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2022, pp. 225–227. doi: 10.1109/ECBIOS54627.2022.9944996.

[5] Puneet, Deepika, P. Singh, R. Bansal, and S. Sharma, "Coronary Heart Disease Prediction Using Voting Classifier Ensemble Learning," in 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 181–185. doi: 10.1109/ICAC3N53548.2021.9725705.

[6] A. Bhatt, S. K. Dubey, and A. K. Bhatt, "Age-Gender Analysis of Coronary Artery Calcium (CAC) Score to predict early Cardiovascular Diseases," in 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 237–241. doi: 10.1109/Confluence47617.2020.9058151.

[7] A. H. Shahid, M. P. Singh, B. Roy, and A. Aadarsh, "Coronary Artery Disease Diagnosis Using Feature Selection Based Hybrid Extreme Learning Machine," in 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 341–346. doi: 10.1109/ICICT50521.2020.00060.

[8] H. Aakkara, A. Aaisueb, and A. Aeelanupab, "Comparing Classifiers for the Prediction of the Stenosis of Coronary Artery," in 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2020, pp. 747–750. doi: 10.1109/ECTI-CON49241.2020.9158312.

[9] S. K. K. L, N. K. G, and M. J. A, "Coronary Artery Disease Prediction using Data Mining Techniques," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 693–697. doi: 10.1109/ICISS49785.2020.9316014.

[10] F. Ghasemi, B. S. Neysiani, and N. Nematbakhsh, "Feature Selection in Pre-Diagnosis Heart Coronary Artery Disease Detection: A heuristic approach for feature selection based on Information Gain Ratio and Gini Index," in 2020 6th International Conference on Web Research (ICWR), 2020, pp. 27–32. doi: 10.1109/ICWR49608.2020.9122285.

[11] G. PHADKE, M. R. RAJATI, and L. PHADKE, "Prediction of Coronary Artery Disease using Electrocardiography: A Machine Learning Approach," in 2020 International Conference on Machine Learning and Cybernetics (ICMLC), 2020, pp. 175–180. doi: 10.1109/ICMLC51923.2020.9469585.

[12] M. Abdar, E. Nasarian, X. Zhou, G. Bargshady, V. N. Wijayaningrum, and S. Hussain, "Performance Improvement of Decision Trees for Diagnosis of Coronary Artery Disease Using Multi Filtering Approach," in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 26–30. doi: 10.1109/CCOMS.2019.8821633.

[13] D. Rafiroiu, I. Molnar, and A. Lungu, "Error Analysis in Patient-Specific Blood Flow Modeling of Coronary Artery Disease," in 2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2019, pp. 1–6. doi: 10.1109/ATEE.2019.8724887.

[14] A. I. Sakellarios et al., "Predictive Models of Coronary Artery Disease Based on Computational Modeling: The SMARTool System," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 7002–7005. doi: 10.1109/EMBC.2019.8857040.

[15] A. A. Haruna, L. J. Muhammad, B. Z. Yahaya, E. J. Garba, N. D. Oye, and L. T. Jung, "An Improved C4.5 Data Mining Driven Algorithm for the Diagnosis of Coronary Artery Disease," in 2019 International Conference on Digitization (ICD), 2019, pp. 48–52. doi: 10.1109/ICD47981.2019.9105844.

[16] L. J. Muhammad, A. A. Haruna, I. A. Mohammed, M. Abubakar, B. G. Badamasi, and J. M. Amshi, "Performance Evaluation of Classification Data Mining Algorithms on Coronary Artery Disease Dataset," in 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), 2019, pp. 1–5. doi: 10.1109/ICCKE48569.2019.8964703.