

An Enhanced Comprehensive Study Towards Predicting Cardiovascular Disease Using Machine Learning Techniques.

Anil Kumar Prajapati*, Umesh Kumar Singh², Ramji Yadav³
¹²³Institute of Computer Science, Vikram University, Ujjain, Madhya Pradesh, India

*anilprajapatiujjain@gmail.com

Abstract. In recent years, heart and cardiovascular diseases have become more common, causing a significant increase in mortality rates worldwide. Data from various organisations highlights the severity of heart disease, which remains a major concern. Accurately and quickly identifying severe conditions, such as heart disease, is vital for effective prevention. Techniques such as data mining, machine learning, and deep learning have been used in medicine to reliably detect heart disease. However, these methods depend on data that can change over time. To ensure accurate detection, proper use of historical data is essential; otherwise, results can be inaccurate. Machine learning techniques produce outcomes based on mathematical calculations, so data cleaning and refinement are necessary. Disease-related data can include text, numbers, and images, which may vary widely, requiring extensive stratification, normalisation, cleaning, encoding, and randomisation; otherwise, results may be biased. Our previous review article addressed a specific challenge related to the CVD Prediction Model. This Enhanced review primarily examines how machine learning techniques operate on medical datasets and their effectiveness in predicting cardiovascular diseases (CVD). It also aims to analyse datasets, features, and machine learning methods used in CVD prediction.

Keywords: Healthcare, Machine Learning, Cardiovascular Disease, ML Algorithm, Ensemble Techniques

How to cite:

Anil Kumar Prajapati, Singh, U. K., & Ramji Yadav. (2026). An Enhanced Comprehensive Study Towards Predicting Cardiovascular Disease Using Machine Learning Techniques. Journal of Systems Engineering and Information Technology (JOSEIT), 5(1), 1-9. <https://doi.org/10.29207/joseit.v3i3.7585>

Received by the Editor: 2026-02-14

Final Revision: 2026-04-26

Published: 28-04-26

SDGs contributed:



© The Author(s) 2025 | This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)
Publisher: [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatikaIndonesia.org/)



1. Introduction

The proper functioning of all human organs relies on blood vessels. When these blood vessels are affected, it is called cardiovascular disease, which impacts specific body organs. The main causes of CVD include an unhealthy diet, physical inactivity, dyslipidemia, hyperglycemia, high blood pressure, obesity, gender, age, race/ethnicity, thrombosis/smoking, kidney dysfunction, diabetes, genetics (family history), body mass index, and hypercholesterolemia[1][2][3]. Cardiovascular disease can be classified into four main types. The first is coronary heart disease, caused by blockages in the blood vessels that supply the heart muscle. The second type is stroke and transient ischemic attack, which result from blockages in the brain's blood vessels that temporarily disrupt blood flow. The third type is peripheral artery disease, which involves blockages in blood vessels that supply other organs. The fourth type is aortic disease, affecting the aorta, the body's largest blood vessel[4][5].

Our previous article, a comprehensive study and analysis of the working process of the machine for CVD Prediction, considered that machine learning and/or deep learning techniques use statistics, probability, computation, and data mining. These approaches are highly effective for data testing due to their exceptional data analysis capabilities and computational power. To assess the effectiveness of heart disease detection, it is essential to study existing data related to heart disease thoroughly, enabling accurate diagnosis. This data can be complex, normal, high-end, or low-end, and detailed analysis is possible through advanced techniques such as machine learning.

The article highlights specific issues related to cardiovascular disease (CVD) prediction, such as dataset selection, feature selection, and variations in accuracy[6]. Machine learning techniques are capable of predicting disease through basic symptoms and historical data, but there are several challenges in using the technology, such as effective dataset collection, dataset preprocessing, feature selection, algorithm selection, model training, etc. Now, the extended paper addresses key challenges and research gaps in predicting CVD using machine learning techniques.

2. Related Work

A review is a thorough analysis of existing research, which considers the data, techniques, and future possibilities related to the subject. Below, we discuss several previous reviews that focus on different aspects of cardiovascular disease (CVD) prediction using machine learning. A Systematic review has been done in the cardiovascular domain by Dudchenko et al [7]. In this review, twenty-seven research articles were examined, which used classification techniques and a high percentage of secondary datasets. The PRISMA statement has been used for systematic reviews. The research articles in the review have been divided into four different categories of the cardiology domain, and some have incorporated review questions based on machine learning methods. In this review, the author did not address the issue of dataset size and source, nor consider dataset preprocessing issues.

An analytical survey on cardiovascular disease (CVD) prediction using machine learning techniques was conducted by Kresoja et al [8]. The review highlights the extensive use of datasets in medical science. Machine learning techniques rely on data labelling and feature selection, so the data must be up-to-date and accurate. Incomplete or unrepresentative data can lead to bias in the predictions. Additionally, the evolving nature of the disease presents challenges when working with omics data. The authors express concerns about making predictions based on a limited number of fixed datasets, as new developments continuously introduce different characteristics of the disease. However, the review does not adequately address significant challenges in machine learning, such as the sources and size of datasets, model training, and parameter tuning.

A meta-analysis on machine learning techniques for predicting CVD was conducted by Krittanawong et al [9]. The review also followed the PRISMA statement for systematic reviews. The study included 55 research articles on CVD prediction based on ML techniques. They found that support vector machine algorithms and boosting algorithms are the best for predicting CVD. They suggested that SVM gives better results on small or large datasets by reducing the high dimensionality of the dataset. They state that the algorithms and datasets used are arbitrary and heterogeneous, and there is no fixed threshold for dataset labelling. This research discusses the dataset but lacks comments on feature selection, cross-fold validation, mathematical modelling, and the optimal dataset size.

A comprehensive survey has been conducted by Naser et al [10]. To predict CVD through machine learning techniques. This survey classified cardiovascular diseases into ten different types, and the literature was reviewed accordingly. Most of the research articles have used only the UCI repository secondary dataset for predicting CVD. They found that bagging and boosting methods provide better performance than traditional ML classifiers. They also suggested that more reliable and realistic datasets of CVD would deliver significant breakthroughs to prediction. This research discusses nearly all aspects of Machine Learning and Deep Learning; however, it completely lacks defined research questions and research gaps.

Rao et al [11] organised a survey based on various techniques for heart disease prediction. In this survey, machine learning, deep learning, ensemble, data mining, and IoT-based prediction techniques were analysed. Secondary data and a limited number of attributes are utilised in well-thought-out approaches for prediction. They suggested that heterogeneous and different demographic datasets are more prone to improving forecasting accuracy. This review lacks details about the dataset (size, source) and omits preprocessing methods (scaling, encoding, randomisation).

A meta-study in the field of social determinants of cardiovascular disease was conducted by Zhao et al [12]. This study reviewed 48 state-of-the-art research articles closest to social determinants. The survey is based on five different parameters of social determinants, which are demographic, economic, neighbourhood, environmental, and social & cultural. The entire review is based on eight different ML classifiers and their performance on the UCI dataset. Additionally, hospital datasets and neural networks have also been used to predict CVD. This review addresses social determinants but neglects various aspects of machine learning methods. Additionally, it lacks discussion on how social determinants can be integrated into ML technology.

The number of deaths from cardiovascular diseases, including heart disease, is steadily rising. Mortality reports from various institutions and agencies emphasise the increasing impact of cardiovascular disease [13][14][15]. The recent Global Burden of Cardiovascular Disease report, covering the years from 1900 to 2022, indicates that ischemic heart disease is the leading cause of death, especially in low-income countries in Asia [16]. The report's data provides important insights into the severity of the disease. Although researchers have proposed several models for cardiovascular disease prediction.

Advanced techniques such as machine learning can predict and identify diseases. However, developing these techniques involves several key challenges, including dataset quality, dataset size, dataset source, feature determination, dataset preprocessing, parameter tuning, and accurate model selection [7][8][9][10][11][12]. Therefore, this scoping review aims to identify specific research problems, research gaps, and uncover unexplored aspects of research that are not covered in existing reviews and are important for the identification/prediction of heart disease. The goal of this proposed research review is to raise research questions that may lead to new directions in the field of cardiovascular disease prediction.

3. Methodology

This enhanced review is based on current trends in technology enhancements, specifically in machine learning for predicting cardiovascular and/or heart diseases. The review focuses on methods of machine learning in the cardiology domain and the role of the dataset. Previously, we have analysed the 21 best research articles of the cardiovascular disease and/or heart disease domain between 2019 to 2023. This meta study focused on the datasets and features used for model training and looked at the impact of machine learning dataset size and selected features on accuracy scores [6].

In light of the previous study, there is a need to understand the current trends, techniques, and methods in the machine learning field. Now this detailed review covers 21 research articles from the previous study as well as five research papers from 2024 and 2025, which are shown in Table 1.0. The complete methodology incorporates the mechanisms of machine learning techniques, their applications, the strengths of machine learning, the computational power of mathematical models, technological enhancements, and their limitations. Review explores research gap between machine learning and cardiology domains based on the existing research articles. All research papers have been assessed based on the review question, that is, for CVD prediction, which methods are focused on, which aspects of the dataset have been used, what results do they reflect, and how effective are they?

Table 1. Dataset Description and Comparison Between 2024 to 2025

Author	ML Classifier/ Techniques	Dataset description	Accuracy Score
DeGroat et al. [17] Scientific Reports 2024	The Model used two ensemble techniques, XGB and SVC, and three ML classifiers, such as RF, SVM, and KNN.	751 Clinically integrated genomic and transcriptomic datasets have been used for prediction, which have been collected from different clinics.	The XGBoost method achieved 96% accuracy and is best suited for predicting.
Ogunpola et al. [18] MDPI Indexed 2024	Four ML classifiers, two ensemble techniques, and a neural network have been used to predict CVD with Grid Search CV.	The Mendeley and Cleveland datasets have been used with 1,000 and 300 instances, respectively, and 14 features.	The XGBoost model optimisation techniques achieved 98.50% accuracy.
Olawade et al [19] Elsevier 2025	Six traditional ML algorithms have been used, such as SVM with linear, KNN, logistic regression, random forest, polynomial, and RBF kernel	Two datasets online, Framingham with 4200 instances, 15 features, and Z-Alizadeh Sani with 304 instances, 55 features, have been used.	90% accuracy achieved by the RF and SVM algorithms on the Framingham Z-Alizadeh Sani dataset, respectively
Alsabhan et al [20] Scientific Reports 2025	Basic ML and/or DL classifiers were used to predict cardiovascular disease	Four UCI repository datasets, i.e., Cleveland, Hungary, Switzerland, and VA Long Beach, have been used with 14 features	The Gradient Boosting Machine (GBM) algorithm achieved the highest accuracy of 90.20% compared to others
Rohan et al [21] Scientific Reports 2025	A variety of machine learning classifiers, including bagging and boosting, have been utilised for predicting cardiovascular disease (CVD).	An integrated dataset has been created from five UCI datasets with 11 features and 1190 instances	The Extreme Gradient Boost algorithm achieved the highest accuracy of 97% compared to other algorithms

4. Result and Discussion

Table 1.0 provide a detailed description of recent trends in ML techniques. A detailed analysis of studies reveals critical insights into CVD prediction using machine learning techniques. These insights aim to synthesize findings, presenting a cohesive view of current trends in CVD prediction and ML methods.

4.1 Role of the ML Algorithm

The ML algorithm and/or mathematical model is a process that analyzes data to extract valuable insights by identifying relationships and correlations among datasets towards computation. Machine learning algorithms classify data and uncover latent and hidden patterns within it [22]. These reviews observed that the underlying concepts and functionality of each mathematical model are different from each other. The ML algorithms utilize various classification techniques, including binary classification (most common), multiclass classification, multilabel classification, and imbalanced classification.

Decision Tree and Random Forest are algorithms based on tree structures for data splitting, while Support Vector Machine, k-Nearest Neighbors, and logistic regression are non-tree-based methods, and Naive Bayes uses probability concepts. To minimize bias and address issues of overfitting or underfitting, it is essential to select an appropriate classifier based on the dataset and the characteristics of the disease [23]. In light of Table 22.1[6] and on behalf of Table 1.0, the review observed that the UCI dataset is frequently used for CVD prediction. This raises the question of how to choose a machine learning classifier that is free from bias when working with the UCI dataset.

Additionally, real-time medical datasets tend to be more complex and multidimensional. They often include a wider range of latent values, both high and low variance. Such datasets require careful refinement and intention. Therefore, an important question arises: what criteria will be used to determine the algorithm, considering that most research relies on UCI datasets?

4.2 Role of the Dataset

The underlying concept behind machine learning techniques is to produce outcomes based on historical datasets related to events by performing calculations on that data. These datasets can vary in type, including numerical, textual, and image data. In disease prediction, all data types are relevant, making their role and application essential. It's important to note that disease symptoms are often interconnected and can influence each other; even if these dependencies are minor, they should not be ignored. The dataset has low-high and/or complex-general values, which affect the result of the prediction.

In Table 1, we have observed and analyzed such a type of dataset, and we observed that the maximum number of UCI and/or predefined datasets were utilized for CVD predictions, which are structured, balanced, have low diversity, and low dimension, and become outdated over time. The review observed that most researchers have used small datasets and/or a limited number of instances. Considering a low-volume dataset does not capture the wide diversity, such as demographic and socioeconomic factors. Small data sets can create issues with cross-validation and randomization during data preprocessing. For instance, when randomizing the dataset, the training set can include data solely from all cardiovascular disease (CVD) patients, while the testing set includes data only from non-CVD patients. In this scenario, the model may demonstrate 100% accuracy, but it would not reflect its true predictive performance. We suggest that the size of the dataset should be appropriate, as large datasets can lead to issues with processing, execution, and memory.

4.3 Role of Features

The ML model works on selected features and/or disease symptoms, and the whole prediction depends on selected features. It is more important to select features that are valuable and highly correlated from the perspective of cardiovascular disease (CVD) to develop a bias-free model[24]. The review noted that most researchers utilize 13 features related to cardiovascular disease (CVD) and achieve an accuracy of around 90%. Some researchers employ 11 features and also reach 90% accuracy.

Additionally, studies are using 15, 16, 18, or even 19, features that attain the same level of accuracy. This raises the question: which features are the most useful, and how many features should be included for model training? Because using fewer features may result in an underfitting issue, while incorporating more features could lead to an overfitting issue.

4.4 Role of Ensemble techniques

The next steps for ML classifiers involve ensemble techniques, which integrate and combine single or weak classifiers. On behalf of Table 1.0 and Table 22.1[6], three major ensemble techniques are encountered: Bagging, Boosting, and Stacking, which have different functionality. The stacking technique relies on probabilities, the bagging technique utilizes Bootstrap aggregating, and the boosting technique is based on majority voting. Stacking ensembles are appropriate for large datasets; while Bagging and Boosting are better suited for smaller datasets. Unlike Bagging, which does not require cross-validation, both Boosting and Stacking do need it. In the case of Boosting, there is a risk of overfitting on small datasets due to low variance, whereas Bagging tends to have a lower risk of overfitting. This raises the issue: which technique will prove to be the most effective for the CVD dataset and features?

4.5. Limitations

The proposed comprehensive review is influenced by the core idea of how machine learning operates on medical datasets and how to generate predictions from highly scalable datasets through a mathematical model. The major limitations, loopholes, and/or gaps considered in the review findings are as follows. Most of the papers in our study involve researchers validating secondary and structured datasets sourced from UCI, Kaggle, and other repositories. Common issues with these structured datasets include limited size, low diversity, outdated content, and the fact that they are often pre-cleaned, low-dimensional, and balanced. The main drawback of UCI and/or structured datasets is that machine learning models can achieve high accuracy on these training datasets, but often perform poorly on real-world data, which is often chaotic and high-dimensional. This situation exemplifies the overfitting problem in machine learning models.

The data preprocessing step is crucial in model training, involving cleaning, encoding, and scaling to resolve issues like dataset imbalance. The main challenge of preprocessing is to maintain the original form of the data. Typically, it is believed that transformations such as cleaning, encoding, and scaling alter the true nature of the data, which can introduce bias into the model. The dataset cleaning techniques enhance accuracy, consistency, and reduce noise. This approach employs methods like Data Imputation, such as creating a new feature, using Mean/Median/Mode, forward/backward fill, and Predictive models to balance the dataset. However, a significant drawback is that small datasets can lose their original structure, leading to the loss of valuable information.

Encoding is the process of converting categorical (non-numeric) data into a numerical format. A major concern is that some encoding methods, such as label encoding, can overly simplify complex categories, leading algorithms to mistakenly assume a hierarchical relationship between labels and/or assume that one label is greater than another. Furthermore, scaling refers to normalizing or standardizing features to ensure they are on a similar scale.

Methods such as min-max scaling, Z-score scaling, interquartile range (IQR), and maximum absolute value are used for scaling and/or normalization. These methods are also more effective for large datasets, and the biggest risk associated with them is the loss of the actual boundaries and the true nature of the dataset.

Selecting suitable and correlated datasets and features for a particular algorithm is also a challenging task. Various statistical methods, including Recursive Feature Elimination, Lasso regression, ANOVA, the Chi-square test, and correlation analysis, are used to identify more effective features. These methods select valuable features from the original dataset. The review reveals significant variation in the features used for model training, yet their accuracy scores remain nearly identical.

Therefore, the critical question is which features are most relevant. Selecting an algorithm for disease prediction is also an important task. Each classifier has unique functionalities, and a wide range of machine learning (ML) algorithms and ensemble techniques are available. Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Decision Trees work best for small to medium-sized datasets. In contrast, Naive Bayes and Random Forest are more suitable for medium to large datasets. Moreover, ensemble techniques work on the combination of algorithms. Almost all the proposed research papers use algorithms and/or ensemble techniques without knowing the nature of the data.

7. Conclusion

The presence or absence of any disease is determined by the symptoms associated with that disease; accurate identification is possible by selecting appropriate correlated symptoms. Machine learning techniques also work on this basis, in which results are produced by performing computations on the old dataset of the disease. In the extended review, we observed that the selection of the dataset and the right ML methods is important in model training. Small datasets bring the problem of overfitting and underfitting, while large datasets have the problem of complexity and processing. Therefore, the dataset must be ideal. Dataset processing is a vital part of machine learning techniques. This process involves tasks like cleaning, encoding, and scaling data using various methods. Applying these methods is necessary to address issues such as dataset imbalance. However, techniques like IQR, Min-Max scaling, Z-score normalization, constant value imputation, and mean/mode imputation can modify the data's original form, significantly influencing the results. Real-time medical datasets tend to be dense and high-dimensional.

Our review shows that datasets from online repositories are often used, which may be sparse, low-dimensional, and stratified based on the findings. In such cases, accuracy can surpass 90%. The issue of dataset imbalance and randomness can be mitigated by employing K-fold cross-validation during model training. Additionally, consulting medical experts for feature selection can be beneficial. Important tasks such as parameter tuning will also contribute to more robust results, depending on the algorithm's functionality. In conclusion, significant results can be achieved through the selection of real-time medical, demographic, and/or social determinants, and/or socioeconomic datasets, the application of precise dataset preprocessing methods, the consistent choice of mathematical models, and effective feature selection.

References

- [1] H. E. Bays *et al.*, "Ten things to know about ten cardiovascular disease risk factors," *Am J Prev Cardiol*, vol. 5, p. 100149, Mar. 2021, doi: 10.1016/J.AJPC.2021.100149.
- [2] L. Ciumămean *et al.*, "Cardiovascular Risk Factors and Physical Activity for the Prevention of Cardiovascular Diseases in the Elderly," *Res. Public Health*, vol. 19, p. 207, 2022, doi: 10.3390/ijerph19010207.
- [3] L. S. Mehta *et al.*, "Cardiovascular Disease Risk Factors in Women: The Impact of Race and Ethnicity: A Scientific Statement from the American Heart Association," *Circulation*, vol. 147, no. 19, pp. 1471–1487, May 2023, doi: 10.1161/CIR.0000000000001139/ASSET/65A233F7-3218-40CA-A84B-DA0CB4BAA410/ASSETS/GRAPHIC/CIR.0000000000001139.FIG02.JPG.
- [4] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.ict.2021.08.021.
- [5] F. Zardawi, S. Gul, A. Abdulkareem, A. Sha, and J. Yates, "Association Between Periodontal Disease and Atherosclerotic Cardiovascular Diseases: Revisited," *Front Cardiovasc Med*, vol. 7, p. 625579, Jan. 2021, doi: 10.3389/FCVM.2020.625579/BIBTEX.
- [6] A. K. Prajapati, U. K. Singh, and D. S. Yadav, "A Comprehensive Study and Analysis of the Working Process of Machine Learning (ML) Techniques to Predict Cardiovascular Disease," pp. 293–306, 2025, doi: 10.1007/978-981-96-2182-8_22.
- [7] A. Dudchenko, M. Ganzinger, and G. Kopanitsa, "Machine Learning Algorithms in Cardiology Domain: A Systematic Review," *Open Bioinforma J*, vol. 13, no. 1, pp. 25–40, Apr. 2020, doi: 10.2174/1875036202013010025.
- [8] K. P. Kresoja, M. Unterhuber, R. Wachter, H. Thiele, and P. Lurz, "A cardiologist's guide to machine learning in cardiovascular disease prognosis prediction," *Basic Res Cardiol*, vol. 118, no. 1, pp. 1–12, Dec. 2023, doi: 10.1007/S00395-023-00982-7/TABLES/2.
- [9] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports* 2020 10:1, vol. 10, no. 1, pp. 1–11, Sep. 2020, doi: 10.1038/s41598-020-72685-1.
- [10] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," *Algorithms* 2024, Vol. 17, Page 78, vol. 17, no. 2, p. 78, Feb. 2024, doi: 10.3390/A17020078.
- [11] G. S. Rao and G. Muneeswari, "A Review: Machine Learning and Data Mining Approaches for Cardiovascular Disease Diagnosis and Prediction," *EAI Endorsed Trans Pervasive Health Technol*, vol. 10, Mar. 2024, doi: 10.4108/EETPHT.10.5411.
- [12] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, "Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review," *Am J Prev Med*, vol. 61, no. 4, pp. 596–605, Oct. 2021, doi: 10.1016/j.amepre.2021.04.016.

- [13] R. L. Sacco *et al.*, “The Heart of 25 by 25: Achieving the Goal of Reducing Global and Regional Premature Deaths From Cardiovascular Diseases and Stroke: A Modeling Study From the American Heart Association and World Heart Federation,” *Circulation*, vol. 133, no. 23, pp. e674–e690, Jun. 2016, doi: 10.1161/CIR.0000000000000395/ASSET/E2F46EAA-3505-4141-BF71-3505B83573E6/ASSETS/GRAPHIC/E674FIG16.JPEG.
- [14] S. C. W. Tan, B.-B. Zheng, M.-L. Tang, H. Chu, Y.-T. Zhao, and C. Weng, “Global Burden of Cardiovascular Diseases and its Risk Factors, 1990-2021: A Systematic Analysis for the Global Burden of Disease Study 2021,” *QJM*, Jan. 2025, doi: 10.1093/QJMED/HCAF022.
- [15] S. S. Martin *et al.*, “2025 Heart Disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association,” *Circulation*, vol. 151, no. 8, pp. e41–e660, Feb. 2025, doi: 10.1161/CIR.0000000000001303.
- [16] G. A. Mensah *et al.*, “Global Burden of Cardiovascular Diseases and Risks, 1990-2022,” *J Am Coll Cardiol*, vol. 82, no. 25, p. 2350, Dec. 2023, doi: 10.1016/J.JACC.2023.11.007.
- [17] W. DeGroat, H. Abdelhalim, K. Patel, D. Mendhe, S. Zeeshan, and Z. Ahmed, “Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine,” *Scientific Reports* 2024 14:1, vol. 14, no. 1, pp. 1–13, Jan. 2024, doi: 10.1038/s41598-023-50600-8.
- [18] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarak, and S. N. Qasem, “Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases,” *Diagnostics* 2024, Vol. 14, Page 144, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/DIAGNOSTICS14020144.
- [19] D. B. Olawade, A. A. Soladoye, B. A. Omodunbi, N. Aderinto, and I. A. Adeyanju, “Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection,” *Int J Cardiol*, vol. 436, p. 133443, Oct. 2025, doi: 10.1016/J.IJCARD.2025.133443.
- [20] W. Alsabhan and A. Alfadhly, “Effectiveness of machine learning models in diagnosis of heart disease: a comparative study,” *Sci Rep*, vol. 15, no. 1, pp. 1–19, Dec. 2025, doi: 10.1038/S41598-025-09423-Y;SUBJMETA=166,639,985;KWRD=BIOMEDICAL+ENGINEERING,ENGINEERING.
- [21] D. Rohan, G. P. Reddy, Y. V. P. Kumar, K. P. Prakash, and C. P. Reddy, “An extensive experimental analysis for heart disease prediction using artificial intelligence techniques,” *Sci Rep*, vol. 15, no. 1, pp. 1–47, Dec. 2025, doi: 10.1038/S41598-025-90530-1;SUBJMETA=166,230,639,692,699,75,985;KWRD=BIOMEDICAL+ENGINEERING,HEART+FAILURE.
- [22] A. K. Prajapati and U. K. Singh, “An empirical analysis of ML techniques and/or algorithms for disease diagnosis prediction from the perspective of cardiovascular disease (CVD),” *International Journal of Computing Algorithm*, vol. 11, no. 2, Dec. 2022, doi: 10.20894/IJCOA.101.011.002.002.
- [23] A. K. P. Anil and U. K. Singh, “An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models,” *Journal of Systems Engineering and Information Technology (JOSEIT)*, vol. 2, no. 2, pp. 77–84, Oct. 2023, doi: 10.29207/JOSEIT.V2I2.5460.
- [24] A. K. Prajapati, U. K. Singh, R. Singh, and A. Shukla, “Study and analysis of feature selection problems and impact of bias in machine learning disease prediction models,” *International Journal of Engineering & Technology*, vol. 13, no. 2, pp. 182–188, May 2024, doi: 10.14419/HN1N3H15.