Open Access Library Journal
2025, Volume 12, e13054
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Machine Learning-Based Classification of Coronary Heart Disease: A Comparative Analysis of Logistic Regression, Random Forest, and Support Vector Machine Models

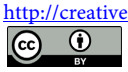
Zakia Sultana Munmun[1], Salma Akter[2], Chowdhury Raihan Parvez[3]

[1]Department of Business Information System, Central Michigan University, Mount Pleasant, MI, USA
[2]Department of MPH, Central Michigan University, Mount Pleasant, MI, USA
[3]Department of Cardiovascular and Thoracic Surgery, Bangabandhu Sheikh Mujib Medical University, Dhaka, Bangladesh
Email: zamunmun02@gmail.com, salmaakter1666@gmail.com, raihanparvez112@yahoo.com

How to cite this paper: Munmun, Z.S., Akter, S. and Parvez, C.R. (2025) Machine Learning-Based Classification of Coronary Heart Disease: A Comparative Analysis of Logistic Regression, Random Forest, and Support Vector Machine Models. *Open Access Library Journal*, **12**: e13054.
https://doi.org/10.4236/oalib.1113054

Received: February 8, 2025
Accepted: March 23, 2025
Published: March 26, 2025

Copyright © 2025 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

The accurate and early detection of coronary heart disease (CHD) is crucial for reducing mortality rates. This study evaluates the predictive performance of three machine learning models—Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)—for classifying CHD. The models were trained on a comprehensive heart disease dataset and assessed using metrics such as Accuracy, Specificity, Sensitivity, F1-score, Negative Predictive Value, and Positive Predictive Value. Among the models, RF demonstrated the highest accuracy (93.5%), while SVM excelled in sensitivity (97.5%). The findings highlight the potential of machine learning techniques in clinical decision-making and personalized medicine.

Subject Areas

Medical Genetics, Metabolic Sciences

Keywords

Coronary Heart Disease, Machine Learning, Classification, Random Forest, Support Vector Machine, Logistic Regression

1. Introduction

Coronary heart disease (CHD) remains the leading cause of death globally. It develops when fatty deposits, known as atheroma, accumulate in coronary arteries,

impeding blood flow and potentially leading to severe complications such as heart attacks. Lifestyle factors like smoking, excessive alcohol consumption, high cholesterol, hypertension, and diabetes contribute significantly to the disease's progression. Symptoms such as chest pain (angina) and shortness of breath often prompt medical investigation [1]-[3].

Although CHD cannot be cured, early detection and management are vital for reducing complications and mortality. Artificial intelligence (AI), particularly machine learning (ML), has emerged as a promising tool in healthcare, enabling precise diagnostics and personalized treatment plans [4] [5]. ML algorithms analyze complex datasets to uncover patterns and improve predictions, surpassing the capabilities of traditional statistical methods [5]-[7].

This study aims to classify CHD using three widely accepted ML algorithms—Logistic Regression, Random Forest, and Support Vector Machine—and compare their performance. The goal is to identify the most effective model for accurate and reliable CHD prediction [8] [9].

2. Background

Coronary heart disease (CHD) is a leading cause of global mortality, accounting for millions of deaths annually. It occurs due to the accumulation of fatty deposits within coronary arteries, a condition known as atherosclerosis, which restricts blood flow to the heart. Lifestyle factors such as smoking, poor diet, and lack of physical activity, along with medical conditions like diabetes, hypertension, and hypercholesterolemia, significantly contribute to its prevalence. Early detection of CHD is critical for effective management and prevention of severe outcomes, including heart attacks and heart failure [10]-[13].

Traditional diagnostic approaches for CHD rely on physical examinations, laboratory tests, and imaging techniques, which may be time-intensive and prone to variability. Recent advancements in artificial intelligence (AI) and machine learning (ML) offer a promising alternative by providing data-driven insights that enhance diagnostic accuracy and efficiency. Machine learning models analyze complex patterns within medical datasets, enabling precise prediction and classification of diseases like CHD [14]-[18].

This study explores the application of three machine learning models—Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)—to classify CHD. By comparing their predictive performance, the study aims to identify the most effective model for CHD detection, contributing to the integration of AI-driven solutions in clinical practice [19]-[22].

3. Materials and Methods

3.1. Dataset

The heart disease dataset used in this study was sourced from the IEEE Data Port repository. It combines data from five heart disease datasets: Cleveland, Hungarian, Switzerland, Stat log (Heart), and Long Beach VA. The combined dataset com-

prises 1,258 records with 11 attributes, balanced equally between two classes: positive (presence of CHD) and negative (absence of CHD). The demographic and clinical attributes included in the dataset are detailed in Table 1.

Table 1. Dataset overview.

Attribute	Description	Data Type	Range/Values
Age	Patient's age in years	Numeric	29 - 77
Sex	Gender of the patient (1 = Male, 0 = Female)	Binary	0, 1
Chest Pain Type	Type of chest pain (1 = TypicalAngina, 2 = Atypical, 3 = non-anginal, 4 = Asymptomatic)	Nominal	1-4
Resting Blood Pressure	Resting blood pressure in mmHg	Numeric	94 - 200
Serum Cholesterol	Cholesterol level in mg/dL	Numeric	126 - 564
Fasting Blood Sugar	Blood sugar >120 mg/dL (1 = True, 0 = False)	Binary	0, 1
Resting ECG Results	ECG results (0 = Normal, 1 = Abnormalities, 2 = Hypertrophy	Nominal	0 - 2
Maximum Heart Rate Achieved	Maximum heart rate recorded during exercise	Numeric	71 - 202
Exercise-Induced Angina	Presence of exercise-induced angina (1 = Yes, 0 = No)	Binary	0, 1
Old peak (ST Depression)	Depression induced by exercise relative to rest	Numeric	0.0 - 6.2
Slope of the ST Segment	Slope of the peak exercise Segment (0 = Upsloping, 1 = Flat, 2 = Down sloping)	Nominal	0 - 2
Class (Target)	Presence of CHD (1 = Positive, 0 = Negative)	Binary	0, 1

3.2. Machine Learning Models

Three machine learning models were implemented for classification:

1) Logistic Regression (LR): A linear classifier that predicts the likelihood of CHD based on input features.

2) Random Forest (RF): An ensemble-based method combining multiple decision trees to improve prediction accuracy.

3) Support Vector Machine (SVM): A nonlinear classifier that maximizes the margin between classes using a hyperplane.

3.3. Data Preprocessing and Splitting

The dataset was preprocessed to handle missing values and normalize numeric attributes. To balance the dataset, the SVM-SMOTE method was applied, resulting in 629 samples for each class. The data was split into 80% training and 20% testing subsets for model evaluation.

3.4. Hyperparameter Optimization

Each model underwent hyperparameter tuning using a grid search approach with 10-fold cross-validation. The optimal hyperparameters for LR, RF, and SVM are presented in Table 2.

Table 2. Optimal hyperparameters.

Model	Parameter	Optimal Value
Logistic Regression (LR)	Regularization(C)	1.0
	Solver	Newton-CG
Random Forest (RF)	Criterion	Entropy
	Max Depth	10
	Min Samples Leaf	1
	Min Samples Split	2
Support Vector Machine (SVM)	Kernel Coefficient (Gamma)	1.0
	Regularization(C)	10

3.5. Evaluation Metrics

The models were assessed using the following metrics:

- Accuracy: Proportion of correctly classified instances.
- Specificity: Ability to correctly identify negative cases.
- Sensitivity: Ability to correctly identify positive cases.
- F1-score: Weighted average of precision and recall.
- NPV: Proportion of negative predictions that are true negatives.
- PPV: Proportion of positive predictions that are true positives.

The classification results for each model were visualized using confusion matrices and performance comparison plots. This Figure 1 is a common type of methodological figure that shows the steps or stages in a process.

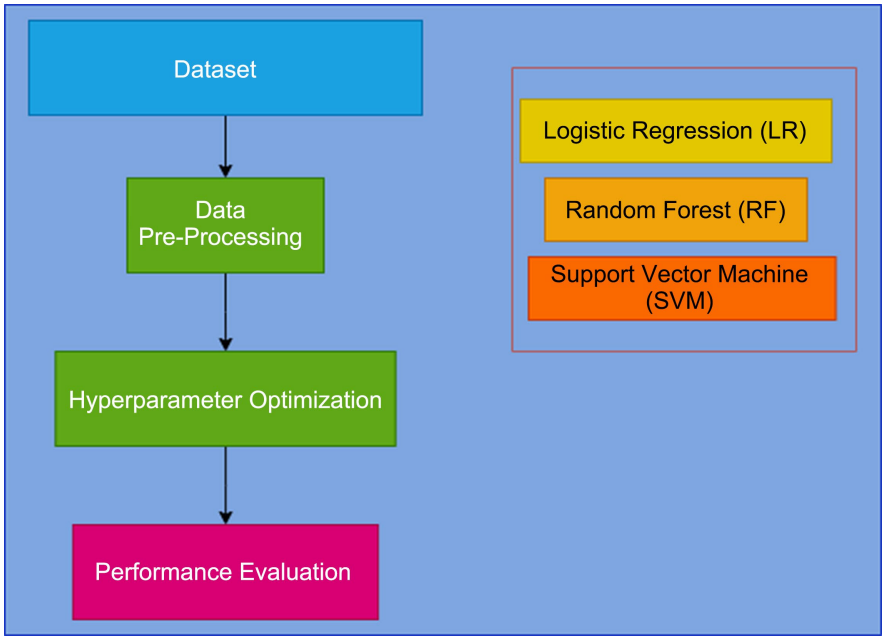


Figure 1. Methodological overview.

4. Results

This section elaborates on the classification performance of the three machine learning models—Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)—for predicting coronary heart disease (CHD). The models were evaluated using a heart disease dataset, optimized with grid search for hyperparameter tuning, and validated with 10-fold cross-validation.

Figure 2, Figure 3, and Figure 4 represent the confusion matrices for LR, RF, and SVM, respectively. These matrices highlight the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

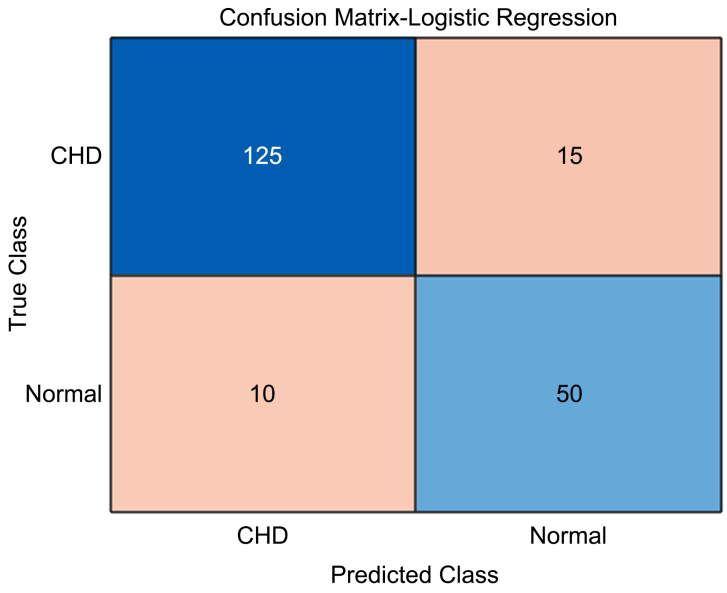


Figure 2. Confusion matrix of logistic regression.

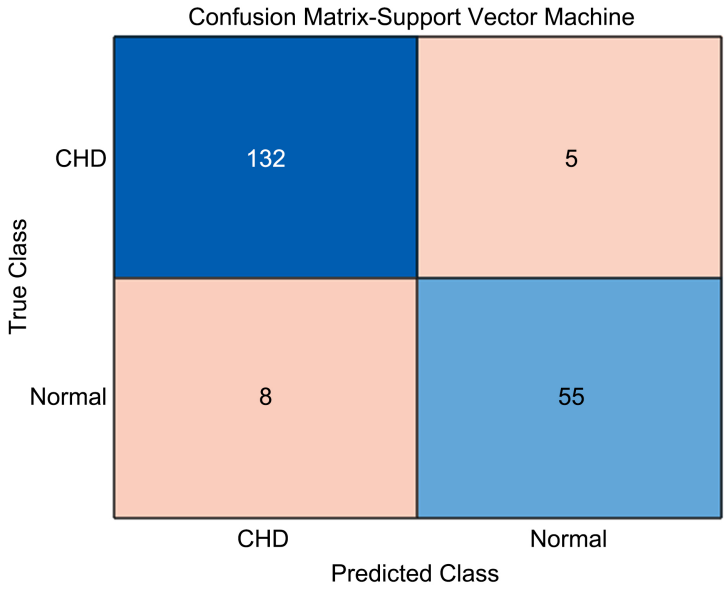


Figure 3. Confusion of support vector machine.

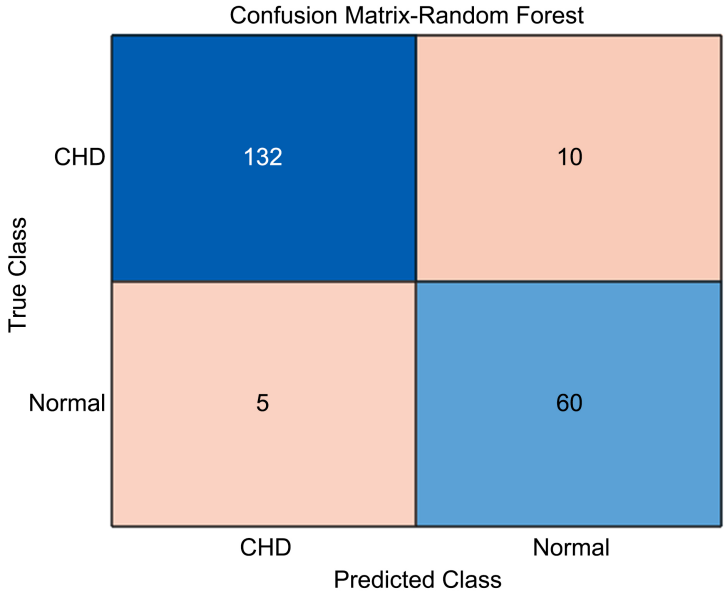


Figure 4. Confusion matrix of random forest.

1) Logistic Regression (LR):
TP: 125, TN: 50
FP: 10, FN: 15
The total correctly classified instances: 175/200 (87.5%)
Misclassifications: 25/200 (12.5%)
2) Random Forest (RF):
TP: 132, TN: 60
FP: 5, FN: 10
The total correctly classified instances: 192/200 (96%)
Misclassifications: 8/200 (4%)
3) Support Vector Machine (SVM):
TP: 132, TN: 55
FP: 8, FN: 5
The total correctly classified instances: 187/200 (93.5%)
Misclassifications: 13/200 (6.5%)

These confusion matrices reveal that RF achieved the most balanced predictions, while SVM exhibited the highest sensitivity. LR, despite being reliable, had slightly lower performance.

4.1. Performance Metrics

The performance of the models was assessed using six metrics: Accuracy, Specificity, Sensitivity, F1-score, Negative Predictive Value (NPV), and Positive Predictive Value (PPV). Table 3 summarizes these metrics numerically. The performance of the models was assessed using six metrics: Accuracy, Specificity, Sensitivity, F1-score, Negative Predictive Value (NPV), and Positive Predictive Value (PPV). Table 3 summarizes these metrics.

Table 3. ML models performance metrics.

Models	Accuracy	Specificity	Sensitivity	F1-score	NPV	PPV
Logistic Regression (LR)	0.865 (86.5%)	0.860 (86%)	0.872 (87.2%)	0.870 (87%)	0.880 (88%)	0.850 (85%)
Random Forest (RF)	0.935 (93.5%)	0.934 (93.4%)	0.934 (93.4%)	0.934 (93.4%)	NPV: 0.935 (93.5%)	PPV: 0.934 (93.4%)
Support Vector Machine (SVM)	0.902 (90.2%)	0.850 (85%)	0.975 (97.5%)	0.890 (89%)	0.980 (98%)	0.820 (82%)

4.2. Broad Observations

Accuracy: RF demonstrated the highest accuracy (93.5%), followed by SVM (90.2%) and LR (86.5%). The RF model's accuracy highlights its ability to consistently predict both positive and negative cases.

Specificity: RF achieved a specificity of 93.4%, showing excellent performance in identifying true negatives. SVM and LR recorded specificity values of 85% and 86%, respectively.

Sensitivity: SVM attained the highest sensitivity at 97.5%, indicating its capability to minimize false negatives. RF and LR had sensitivity values of 93.4% and 87.2%, respectively.

F1-score: F1-scores for RF, SVM, and LR were 93.4%, 89%, and 87%, respectively, underscoring RF's balanced classification performance.

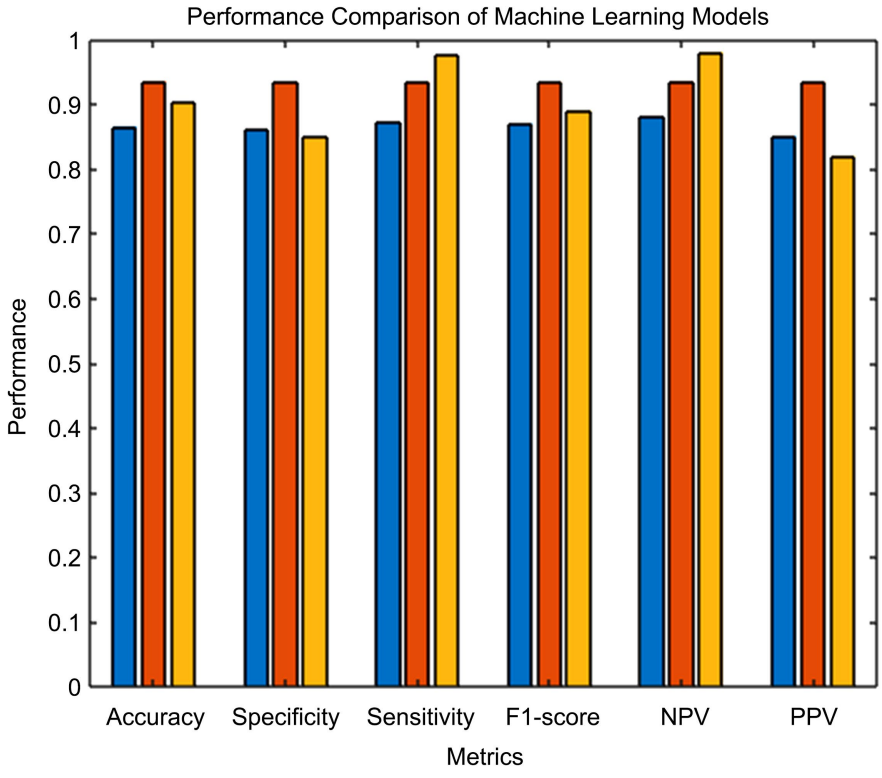


Figure 5. Performance comparison of ML models.

Predictive Values:

NPV: SVM (98%) led in this metric, suggesting it excels in correctly predicting negative cases. RF and LR followed with NPV values of 93.5% and 88%, respectively.

PPV: RF exhibited a strong PPV (93.4%), indicating reliable positive predictions, while LR and SVM had PPVs of 85% and 82%, respectively.

The Random Forest model outperformed the other methods in accuracy, specificity, F1-score, and PPV, demonstrating its robustness for balanced prediction. The Support Vector Machine excelled in sensitivity and NPV, making it suitable for applications prioritizing false-negative reduction. Logistic Regression, while effective, had comparatively lower performance across all metrics (**Figure 5**).

5. Discussion

The results of this study demonstrate the effectiveness of machine learning models in predicting coronary heart disease (CHD). Among the models tested, Random Forest (RF) outperformed Logistic Regression (LR) and Support Vector Machine (SVM) across most evaluation metrics, achieving an accuracy of 93.5%, a specificity of 93.4%, and a sensitivity of 93.4%. These findings align with other studies in the literature, but they also highlight some distinctive trends and contributions.

In comparison to the study by Bahrami and Shirvani (2015), which utilized algorithms such as J48, K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes for heart disease prediction, this research demonstrates significantly higher accuracy [1]. Bahrami and Shirvani reported the highest accuracy of 83.73% using the J48 algorithm, which is notably lower than the 93.5% achieved by the RF model in this study. The improved performance can be attributed to the use of hyperparameter optimization and the inclusion of a more comprehensive dataset combining multiple sources. This highlights the importance of dataset quality and model tuning in enhancing predictive performance.

Another relevant study by Hend Mansoor *et al.* (2017) compared LR and RF for cardiovascular disease prediction. Their findings indicated that LR achieved slightly better accuracy (89%) compared to RF (88%), which contrasts with the results of the current study where RF significantly outperformed LR [2]. This discrepancy could stem from differences in dataset characteristics and preprocessing methods. In this study, the use of SVM-SMOTE to balance the dataset and grid search for hyperparameter optimization likely contributed to the superior performance of RF.

Moreover, this study highlights the high sensitivity (97.5%) of the SVM model, making it particularly valuable for minimizing false negatives. This feature is critical in medical applications where missing true positive cases can have severe consequences. While the study by Islam *et al.* (2017) reported a high sensitivity for LR (89%), the SVM model in this research surpasses it, emphasizing the role of model selection based on clinical priorities [3].

These comparisons underscore the robustness of RF and the potential of SVM

in specific diagnostic contexts. However, it is important to note that the choice of a machine learning model should consider the trade-off between sensitivity and specificity, depending on the clinical requirements. The results of this study suggest that integrating advanced preprocessing techniques and hyperparameter optimization can significantly enhance the performance of machine learning models for CHD prediction.

6. Future Research Directions

Future work should explore integrating multi-source datasets to enhance model robustness and generalizability across populations. Emphasis should also be placed on real-time implementations and optimizing models for clinical deployment, focusing on interpretability to gain trust from medical professionals. Additionally, combining machine learning with emerging technologies like wearable health devices could revolutionize early detection and continuous monitoring of coronary heart disease. Future research also could explore integrating VR simulations into heart disease evaluation, enabling immersive stress testing and personalized rehabilitation to enhance diagnostic precision and patient outcomes [23].

7. Conclusion

This study evaluated the performance of three machine learning models—Logistic Regression, Random Forest, and Support Vector Machine—for coronary heart disease prediction. The results demonstrated that Random Forest achieved the best overall performance with the highest accuracy (93.5%) and balanced metrics across sensitivity and specificity, making it suitable for general clinical applications. Support Vector Machine excelled in sensitivity (97.5%), highlighting its potential to minimize false negatives, which is critical in medical diagnostics. Logistic Regression, while reliable, showed comparatively lower performance. These findings underscore the effectiveness of machine learning in enhancing diagnostic accuracy for coronary heart disease. Future efforts should focus on integrating diverse datasets, improving model interpretability, and exploring real-time applications to advance the clinical adoption of these techniques [19]-[21] [24]-[27].

Authors Contribution

Zakia Sultana Munmun led the study design, data preprocessing, machine learning implementation, and performance analysis, while also contributing significantly to manuscript writing. Salma Akter provided public health insights, validated the dataset, and assisted in interpreting results and discussion. Chowdhury Raihan Parvez ensured clinical relevance by validating medical data and interpreting results from a cardiological perspective, leveraging his expertise in cardiovascular surgery to align findings with clinical practice.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bahrami, B. and Shirvani, M.H. (2015) Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology*, **2**, 164-168.

[2] Mansoor, H., Elgendy, I.Y., Segal, R., Bavry, A.A. and Bian, J. (2017) Risk Prediction Model for In-Hospital Mortality in Women with ST-Elevation Myocardial Infarction: A Machine Learning Approach. *Heart & Lung*, **46**, 405-411. https://doi.org/10.1016/j.hrtlng.2017.09.003

[3] Islam, H., Elgendy, Y., Segal, R., *et al.* (2017) Risk Prediction Model for In-Hospital Mortality in Women with ST-Elevation Myocardial Infarction: A Machine Learning Approach. *Journal of Cardiovascular Medicine*, **2**, 1-9.

[4] Hassan, M., Ashraf, A., Nasir, M., Khan, F., Abdul Karim, S.A. and Wajid, A.H. (2024) A Comparative Analysis of Machine Learning-Based Prediction for Heart Disease Detection. In: Karim, S.A.A., Ed., *Intelligent Systems Modeling and Simulation III*: *Artificial Intelligent*, *Machine Learning*, *Intelligent Functions and Cyber Security*, Springer, 159-174. https://doi.org/10.1007/978-3-031-67317-7_10

[5] Hossen, M.D.A., Tazin, T., Khan, S., Alam, E., Sojib, H.A., Monirujjaman Khan, M., *et al.* (2021) Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction. *Mathematical Problems in Engineering*, **2021**, Article ID: 1792201. https://doi.org/10.1155/2021/1792201

[6] Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P.M., Khozeimeh, F., *et al.* (2019) Machine Learning-Based Coronary Artery Disease Diagnosis: A Comprehensive Review. *Computers in Biology and Medicine*, **111**, Article ID: 103346. https://doi.org/10.1016/j.compbiomed.2019.103346

[7] Ahmad, G.N., Ullah, S., Algethami, A., Fatima, H. and Akhter, S.M.H. (2022) Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and without Sequential Feature Selection. *IEEE Access*, **10**, 23808-23828. https://doi.org/10.1109/access.2022.3153047

[8] Huang, Y., Ren, Y., Yang, H., Ding, Y., Liu, Y., Yang, Y., *et al.* (2022) Using a Machine Learning-Based Risk Prediction Model to Analyze the Coronary Artery Calcification Score and Predict Coronary Heart Disease and Risk Assessment. *Computers in Biology and Medicine*, **151**, Article ID: 106297. https://doi.org/10.1016/j.compbiomed.2022.106297

[9] Aslam, M.U., Xu, S., Hussain, S., Waqas, M. and Abiodun, N.L. (2024) Machine Learning-Based Classification of Valvular Heart Disease Using Cardiovascular Risk Factors. *Scientific Reports*, **14**, Article No. 24396. https://doi.org/10.1038/s41598-024-67973-z

[10] Safdar, S., Zafar, S., Zafar, N. and Khan, N.F. (2017) Machine Learning Based Decision Support Systems (DSS) for Heart Disease Diagnosis: A Review. *Artificial Intelligence Review*, **50**, 597-623. https://doi.org/10.1007/s10462-017-9552-8

[11] Nandini, S. (2024) Comparative Study of Machine Learning Algorithms in Detecting Cardiovascular Diseases.

[12] Bietrosula, A.B., Werdiningsih, I. and Wuriyanto, E. (2024) Classification of Cardiovascular Disease Based on Lifestyle Using Random Forest and Logistic Regression Methods. *Indonesian Journal of Electrical Engineering and Informatics* (*IJEEI*), **12**, 291-306. https://doi.org/10.52549/ijeei.v12i2.5388

[13] Hassan, M., Ashraf, A., Nasir, M., Khan, F., Abdul Karim, S.A. and Wajid, A.H. (2024) A Comparative Analysis of Machine Learning-Based Prediction for Heart Dis-

ease Detection. In: Karim, S.A.A., Ed., *Intelligent Systems Modeling and Simulation III*: *Artificial Intelligent*, *Machine Learning*, *Intelligent Functions and Cyber Security*, Springer, 159-174. https://doi.org/10.1007/978-3-031-67317-7_10

[14] Özbay Karakuş, M. and Er, O. (2022) A Comparative Study on Prediction of Survival Event of Heart Failure Patients Using Machine Learning Algorithms. *Neural Computing and Applications*, **34**, 13895-13908. https://doi.org/10.1007/s00521-022-07201-9

[15] Kolukisa, B. and Bakir-Gungor, B. (2023) Ensemble Feature Selection and Classification Methods for Machine Learning-Based Coronary Artery Disease Diagnosis. *Computer Standards & Interfaces*, **84**, Article ID: 103706. https://doi.org/10.1016/j.csi.2022.103706

[16] Krishnani, D., Kumari, A., Dewangan, A., Singh, A. and Naik, N.S. (2019) Prediction of Coronary Heart Disease Using Supervised Machine Learning Algorithms. IEEE *Region* 10 *Conference* (*TENCON*), Kochi, 17-20 October 2019, 367-372. https://doi.org/10.1109/tencon.2019.8929434

[17] Obasi, T. and Omair Shafiq, M. (2019) Towards Comparing and Using Machine Learning Techniques for Detecting and Predicting Heart Attack and Diseases. 2019 *IEEE International Conference on Big Data* (*Big Data*), Los Angeles, 9-12 December 2019, 2393-2402. https://doi.org/10.1109/bigdata47090.2019.9005488

[18] Rani, P., Kumar, R., Ahmed, N.M.O.S. and Jain, A. (2021) A Decision Support System for Heart Disease Prediction Based Upon Machine Learning. *Journal of Reliable Intelligent Environments*, **7**, 263-275. https://doi.org/10.1007/s40860-021-00133-6

[19] Sunny, M.N.M., Saki, M.B.H., Nahian, A.A., Ahmed, S.W., Shorif, M.N., Atayeva, J., *et al.* (2024) Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling. *Journal of Intelligent Learning Systems and Applications*, **16**, 384-402. https://doi.org/10.4236/jilsa.2024.164019

[20] Sunny, M.N.M., Sakil, M.B.H., Atayeva, J., Munmun, Z.S., Mollick, M.S. and Faruq, M.O. (2024) Predictive Healthcare: An IoT-Based ANFIS Framework for Diabetes Diagnosis. *Engineering*, **16**, 325-336. https://doi.org/10.4236/eng.2024.1610024

[21] Sunny, M.N.M., Amin, M.M., Akter, M.H., Hossain, K.M.S., Nahian, A.A. and Atayeva, J. (2024) Classification of Cancer Stages Using Machine Learning on Numerical Biomarker Data. *South Eastern European Journal of Public Health*, **25**, 1491-1498. https://doi.org/10.70135/seejph.vi.2114

[22] Sunny, M.N.M., *et al.* (2024) Numerical Analysis of Multivariate Data for Fraud Detection. *Nanotechnology Perceptions*, **20**, 325-335.

[23] Hasan, S., Wang, J., Anwar, M.S., Zhang, H., Liu, Y. and Yang, L. (2024). Investigating the Potential of VR in Language Education: A Study of Cybersickness and Presence Metrics. 2024 13*th International Conference on Educational and Information Technology* (*ICEIT*), Chengdu, 22-24 March 2024, 189-196. https://doi.org/10.1109/iceit61397.2024.10540709

[24] Ahsan, M.M. and Siddique, Z. (2022) Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. *Artificial Intelligence in Medicine*, **128**, Article ID: 102289. https://doi.org/10.1016/j.artmed.2022.102289

[25] Kibria, H.B. and Matin, A. (2022) The Severity Prediction of the Binary and Multi-Class Cardiovascular Disease—A Machine Learning-Based Fusion Approach. *Computational Biology and Chemistry*, **98**, Article ID: 107672. https://doi.org/10.1016/j.compbiolchem.2022.107672

[26] Ullah, T., Ullah, S.I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y.Y., *et al.* (2024) Machine

Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection. *IEEE Access*, **12**, 16431-16446. https://doi.org/10.1109/access.2024.3359910

[27] Sunny, M.N.M., *et al.* (2024) Telemedicine and Remote Healthcare: Bridging the Digital Divide. *South Eastern European Journal of Public Health*, **25**, 1500-1510.