Journal of Pericardiology
Vol.01, No.02
www.jopjournal.com

Development and Validation of a Machine Learning Framework for Cardiovascular Disease Risk Stratification Using Clinical, Anthropometric, and Lifestyle Factors

Dr Rajkiran Tiku (PT)

HOD/Professor, Physiotherapy, Cardiorespiratory Physiotherapy, Suryadatta Institute of Health Sciences - College of Physiotherapy, Pune /411021, ORCID ID : 0009-0008-8201-6976
Email ID: rajkiran.tiku@gmail.com

Received: 14/06/2026 **Revised:** 15/07/2026 **Accepted:** 21/07/2026 **Published:** 29/07/2026

Abstract

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, emphasizing the need for accurate and interpretable predictive models to support early risk assessment and preventive healthcare. This study developed and compared five supervised machine learning algorithms Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, and XGBoost for multiclass CVD risk prediction using demographic, anthropometric, biochemical, lifestyle, and clinical variables. A comprehensive preprocessing pipeline, including missing-value imputation, feature standardization, categorical encoding, and elimination of target leakage variables, was implemented before model development. Model performance was evaluated using accuracy, balanced accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). Among the evaluated algorithms, XGBoost achieved the best predictive performance, obtaining an accuracy of 65.69% and a macro-ROC-AUC of 0.8095, demonstrating superior discrimination compared with the remaining models. Explainable artificial intelligence using SHAP analysis revealed that total cholesterol, HDL cholesterol, age, body weight, smoking status, diabetes status, family history of CVD, physical activity, systolic blood pressure, and fasting blood glucose were the most influential predictors driving model decisions. The alignment between SHAP-derived explanations and established cardiovascular risk factors supports the clinical credibility of the proposed framework. Overall, the findings demonstrate that explainable ensemble machine learning provides a robust and clinically interpretable approach for multiclass cardiovascular risk prediction, offering considerable potential for integration into digital decision-support systems and preventive cardiovascular healthcare.

Keywords: Cardiovascular disease, Machine learning, XGBoost, Explainable artificial intelligence, Risk prediction

1. Introduction

Despite all the progress achieved in diagnosis, prevention and treatment, cardiovascular diseases (CVDs) remain the leading cause of mortality and morbidity, creating major health problems for the world. Hypertension, obesity, diabetes mellitus, dyslipidemia and unhealthy lifestyles have been responsible for a sustained rise in cardiovascular morbidity and mortality in both developed and developing countries. Global burden of disease (GBD) estimates indicated that cardiovascular diseases are the primary contributor to premature mortality, resulting in significant premature mortality and disability-adjusted life years (DALYs) and are projected to continue to account for a higher proportion of global mortality and DALYs in the future due to population ageing and rising exposure to cardiovascular disease modifiable risk factors (Roth et al., 2020). Recent estimates of the burden of cardiovascular disease at the global level also highlight that over the last 30 years this burden has grown significantly, reinforcing the importance of more effective preventive measures and identification of high-risk individuals (Lindstrom et al., 2022). In addition to its clinical implications, cardiovascular disease also has significant economic costs, affecting the health sector through the costs of diagnosis and treatment, the work sector through the decrease in workforce productivity, as well as the long-term disability that can result from the disease (Marinho, 2024).

The early cardiovascular risk assessment has a very important role to play in preventing cardiovascular events through timely lifestyle modification and appropriate clinical interventions. Traditional cardiovascular risk prediction tools, such as the World Health Organization (WHO) cardiovascular risk charts, or clinical scoring systems, rely on demographic information and traditional clinical measurements (such as age, blood pressure, cholesterol level, smoking, and diabetes) to estimate the future risk of cardiovascular events (Kaptoge et al., 2019). These tools have been incorporated into primary prevention guidelines due to their ability to aid in the evidence-based clinical decision-making and the individual treatment plan (Lloyd-Jones et al, 2019). Typically, however, the traditional statistical models make assumptions of linear relationships between the predictor variables, and poorly describe the complex interactions involved in the development of cardiovascular disease. As a result, they might not be as accurate in predicting outcomes in diverse groups of people and in those who have more than one complex risk factor.

Cardiovascular disease (CVD) prediction has seen significant advances due to the evolution of artificial intelligence (AI), especially machine learning (ML), which have provided valuable computational techniques to enhance prediction. Machine learning algorithms are able to detect nonlinear relationships, multidimensional interactions between features, and hidden patterns in health care data, resulting in better predictive performance than many traditional statistical methods. Multiple machine learning algorithms have been shown to perform better in terms of classification accuracy, robustness and generalizability in various prediction tasks through comparative studies (Sheth et al., 2022). Similarly, adequate tests for advanced machine learning models have demonstrated their ability to handle complex clinical data, improve predictive accuracy, and become more useful in the context of intelligent clinical decision support systems (Sekeroglu et al., 2022). The

potential applications in the healthcare sector are continually improving with the integration of advancements in machine learning techniques, enabling more accurate and data-driven predictions of disease and personalised patient care (Hatta et al., 2024).

Apart from the use of high-quality algorithms, it is also important that the clinically relevant variables are well integrated into good CV risk prediction. Demographic characteristics, anthropometric measurements, metabolic biomarkers and lifestyle behaviours all play a role in the overall cardiovascular health of an individual that can impact cardiovascular disease. The traditional cardiovascular risk factors such as blood pressure, serum cholesterol, Fasting Blood Glucose and diabetes status remain important predictors of cardiovascular outcome. Moreover, anthropometric indicators, such as BMI, abdominal circumference and waist-to-height ratio have been shown to be good predictors of obesity-related metabolic complications and cardiovascular risk (Tabary et al., 2021). Recent studies have also suggested that WHtR could be a valuable index for the identification of high-risk individuals for cardiometabolic diseases and may be a better marker than other common obesity indices for clinical screening (Sartika et al., 2025). Comprehensive predictive models should also include lifestyle factors, such as smoking, physical activity and family history of cardiovascular disease, which also affect susceptibility to disease.

While there are numerous studies on machine learning applications in cardiovascular disease prediction, some important gaps still exist. Most current models are based on only a few clinical variables, consider a few algorithms, or have inadequate clinical validation or interpretability, which limits their use in clinical practice. Thus, validated machine learning frameworks that incorporate routine clinical, anthropometric and lifestyle data to provide accurate, reliable and clinically meaningful cardiovascular risk stratification are needed. In the present study, therefore, we aimed to design and test a complete machine learning system that can predict cardiovascular disease risk with the use of a variety of machine learning algorithms, with the routine clinical data of health care. The proposed framework aims to enhance the accuracy of prediction, to identify influential cardiovascular risk factors and to facilitate evidence-based clinical decision making for the early prevention of cardiovascular disease.

Research Objectives

1. To develop and validate machine learning models for accurate stratification of cardiovascular disease risk using clinical, anthropometric, and lifestyle factors.
2. To compare the predictive performance of multiple machine learning algorithms in identifying individuals with varying levels of cardiovascular disease risk.
3. To identify the most influential clinical, anthropometric, and lifestyle predictors contributing to cardiovascular disease risk using feature importance analysis.

2. Materials and Methods

2.1 Study Design

In this study, the researchers used the retrospective quantitative research design approach to develop and validate a machine learning framework for CVD risk stratification. In this study, a comparative

predictive modeling strategy was used to assess the classification of cardiovascular disease risk using different machine learning algorithms on health records. The process involved data pre-processing, development of models, validation of models, and evaluation of the performance of the models.

2.2 Dataset Description

For this analysis, an open-access cardiovascular disease dataset of 1,529 participants with 22 features was used (Zaki,2025). The features of this dataset included demographic features (age and gender), anthropometric features (height, weight, body mass index, abdominal girth, and waist-to-height ratio), clinical features (blood pressure, lipid levels, fasting blood glucose levels, and diabetes presence), behavioral factors (smoker or non-smoker, physical activity, and family history of cardiovascular disease), and cardiovascular disease risk groups. These were chosen as the variables due to their established relevance in cardiovascular risk assessment.

2.3 Data Preprocessing

Prior to the development of the model, the data set was analyzed for completeness, duplicate entries, and inconsistencies to maintain data quality. In this case, categorical variables were encoded as numeric variables, whereas the continuous variables were standardized to remove disparities in the measuring scale. Exploratory analysis was conducted to analyze the distribution of the variables and check for any outliers in the data set. Later, the data set was split into training (80%) and test sets (20%).

2.4 Machine Learning Model Development

Implementation of five machine learning algorithms that fall into the supervised learning category namely Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Extreme Gradient Boosting (XGBoost) for cardiovascular disease risk classification was done. Hyperparameter tuning of these models was carried out using grid search along with five folds cross-validation to ascertain the best model configuration. Further, the trained models were tested on the independent test set.

2.5 Model Performance Evaluation

Predictive accuracy of each machine learning algorithm was evaluated using classification metrics, such as accuracy, precision, recall, F1-score, and area under the curve of the receiver operating characteristic (AUC-ROC). Confusion matrices were constructed for the analysis of classification performance in various groups of cardiovascular risk stratification and possible misclassification. Combination of these additional evaluation metrics offered a complete evaluation of models' discriminatory power and accuracy.

2.6 Feature Importance Analysis

Analyzing the best machine learning algorithm to evaluate the impact of each independent variable in predicting the risk of developing cardiovascular diseases. This was achieved by measuring the impact of demographic, anthropometric, clinical, and lifestyle variables in the prediction made by the model. The feature importance rankings obtained helped in making the proposed framework more interpretable.

3. Results

3.1 Dataset Characteristics and Data Preparation

The cardiovascular disease (CVD) data set consisted of 1,529 subjects and 22 variables such as demographic, anthropometric, biochemical, lifestyle, clinical and cardiovascular risk factors. There were no duplicate observations found. There were 1022 missing values (4.19% to 5.36% of each variable). About 49.84% of the records were complete, and 50.16% were missing some data. Numerical variables were imputed using the median value, while for categorical variables the mode was used to maintain the sample size while minimizing the amount of potential bias.

Analysis of the data showed a high level of agreement between duplicated measurements for data consistency. Both the heights in meters and centimeters were consistent, and both pairs of blood pressure measurements were consistent. However, 461 records were recorded as inconsistencies between BMI reported and calculated, suggesting that BMI was not an attribute that was directly measured. Therefore, BMI was not included in the predictive modeling process as it was redundant. The same was done for the variables which have been calculated directly from the target or other predictors such as CVD Risk Score, Blood Pressure Category, Estimated LDL, Height (cm) and combined Blood Pressure. These 15 predictor variables were the final dataset used for modeling: 10 numerical and 5 categorical.

3.2 Data Partitioning and Preprocessing

The data set was split into 80% for training and 20% for independent testing, resulting in 1,223 training samples and 306 independent testing samples. The original class distribution remained unchanged in all subsets due to stratification. The HIGH-, INTERMEDIARY-, and LOW-risk classes made up 47.61%, 38.00%, and 14.39% of the entire data set, with almost the same proportions in both the training and testing partitions.

To avoid information leakage, a pipeline for pre-processing was only fitted on the training data. For numerical variables, median imputation was done followed by Z score standardization; for categorical variables, most frequent category imputation and one-hot encoding were performed. The preprocessing steps led to an increase in the number of predictors from 15 to 21. No missing value and no non-finite value were found in the datasets used for training and testing, with feature same structure in both sets for consistency during the model development.

3.3 Comparative Performance of Machine Learning Models

Five supervised machine learning algorithms were tested with five-fold cross validation on the training portion and tested on the independent test set. Algorithms evaluated were: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest (RF) and XGBoost. The results of the five ML models (predictive performance) are compared in Table 1. XGBoost outperformed all the other models in terms of generalization and thus was chosen for further study.

Table 1. Comparative Performance of Machine Learning Models

Model	CV Accuracy	Test Accuracy	Balanced Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.6509	0.6569	0.5397	0.5602	0.5397	0.5351	0.8095
Logistic Regression	0.6476	0.6405	0.5212	0.5940	0.5212	0.5153	0.7095
Support Vector Machine	0.6378	0.6340	0.4949	0.7519	0.4949	0.4651	0.7774
Random Forest	0.6313	0.6307	0.5073	0.5448	0.5073	0.4972	0.8016
K-Nearest Neighbors	0.5707	0.5784	0.4460	0.4205	0.4460	0.4261	0.6789

The algorithm that performed the best with the highest predictive performance was XGBoost, with a testing accuracy of 65.69%, a balanced accuracy of 53.97%, a macro-average F1 score of 53.51%, and a macro-ROC-AUC of 0.8095. Logistic Regression gave nearly the same overall accuracy, but a lower discrimination ability. Random Forest and SVM gave moderate predictive performance. KNN achieved the worst results in almost all the evaluation measures. In consideration of the results, XGBoost was chosen as the predictive model.

3.4 Performance of the Final XGBoost Model

The final XGBoost classifier showed satisfactory discrimination in the independent test data set with an overall accuracy of 65.69% and macro-ROC-AUC of 0.8095. The highest precision (70.89%), recall (76.71%) and F1 score (73.68%) were obtained for the model when identifying HIGH-risk individuals. The performance for the INTERMEDIARY-risk-group was also acceptable with an F1-score of 67.48%. The performance was much less good for the LOW-risk category with a recall of 13.64% and an F1-score of 19.35%, suggesting that there is significant overlap between LOW-risk cases and the other classes. Table 2 shows a summary of the class-wise predictive performance of the final XGBoost classifier, with better performance for HIGH- and INTERMEDIARY-risk patients compared to LOW-risk patients.

Table 2. Classification Performance of the Final XGBoost Model

Risk Class	Precision	Recall	F1-score	Support
HIGH	0.7089	0.7671	0.7368	146

INTERMEDIARY	0.6385	0.7155	0.6748	116
LOW	0.3333	0.1364	0.1935	44

3.5 Confusion Matrix and ROC Analysis

The confusion matrix also validated the predictive performance of the XGBoost classifier. In 146 High-risk individuals, 112 high-risk patients were collected and 27 patients categorized in INTERMEDIARY were misclassified, and 7 patients categorized in LOW were misclassified. For the INTERMEDIARY class, 83 out of 116 cases were classified correctly while 28 were misclassified as HIGH risk. The misclassification rate was highest in the LOW-risk class, with 6 of 44 patients correctly classified. The distribution of correct and incorrect predictions made by the XGBoost classifier in the three cardiovascular risk categories is shown in figure 1.

The receiver operating characteristic (ROC) curve was used to validate good discriminative power for all three outcome classes. The receiver operating characteristic curves are shown for each cardiovascular risk group in Figure 2, which further demonstrates good discriminative ability. The class-specific AUC values are: 0.843 for the HIGH risk class, 0.809 for the INTERMEDIARY risk, and 0.777 for the LOW risk class; the macro-average ROC-AUC value is 0.8095.

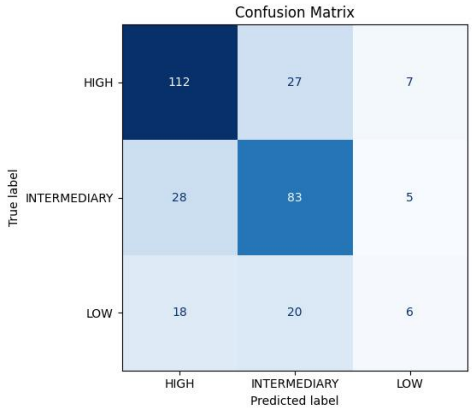

Figure 1. Confusion matrix of the final XGBoost classifier

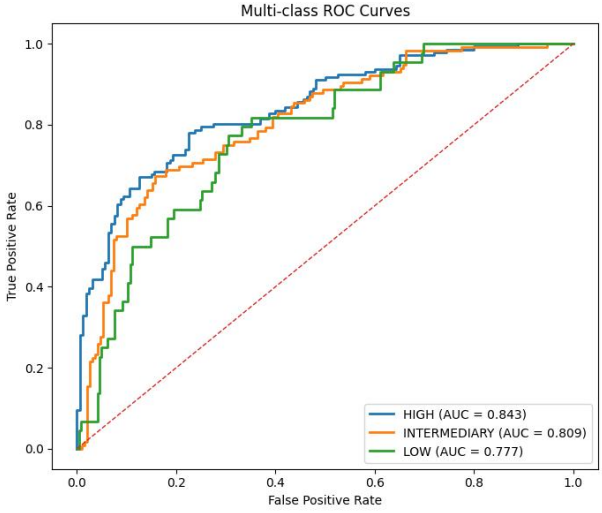

Figure 2. Multi-class ROC curves illustrating the discriminative performance of the XGBoost

model

3.6 Explainable Artificial Intelligence Analysis

The interpretability of the XGBoost classifier was enhanced using the SHAP analysis, which quantifies the importance of each predictor in the model's prediction. Seven factors were determined by the global SHAP analysis: Total Cholesterol, HDL, Weight, Age, Smoking Status, Family History of CVD, Diabetes Status, and Low Physical Activity were found to be the most influential factors affecting cardiovascular risk prediction.

Table 3. Top Ten Global SHAP Predictors

Rank	Predictor	Mean Absolute SHAP Value
1	Total Cholesterol (mg/dL)	0.337960
2	HDL (mg/dL)	0.283182
3	Weight (kg)	0.276106
4	Age	0.270806
5	Smoking Status	0.235053
6	Family History of CVD	0.234041
7	Diabetes Status	0.225248
8	Physical Activity Level (Low)	0.224035
9	Systolic BP	0.159972
10	Fasting Blood Sugar (mg/dL)	0.150329

When analysing the contribution for each class, the Total Cholesterol consistently had the highest contribution in each of the three cardiovascular risk categories. The other most influential factors – HDL, Weight, and Age – were also found to be significant in every class, and smoking status and diabetes status were found to be more influential in predicting HIGH-risk than in predicting LOW-risk. From these results, it can be seen that XGBoost model used clinically meaningful cardiovascular risk factors, and not demographic factors like sex, as the main feature.

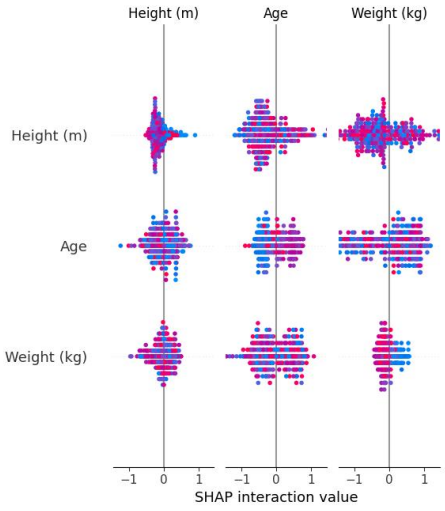

Figure 3. SHAP global feature importance illustrating the average contribution of each predictor across all cardiovascular risk classes

4. Discussion

The goal of the present study was to design and test a machine learning model for the prediction of cardiovascular disease (CVD) risk in a multi-class classification problem, based on demographic,

anthropometric, biochemical, lifestyle and clinical parameters. When it comes to the discriminative capability of the algorithms, XGBoost was the best with an accuracy of 65.69%, an F1-score of 53.51%, and an ROC-AUC of 0.8095 on the testing set, outperforming Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbors. The comparative evaluation shows that ensemble boosting methods are more successful in modelling nonlinear relationships and complex interactions of cardiovascular risk factors. In line with this, Kim et al. (2022) showed that machine learning models that combined anthropometric, lifestyle, and biochemical parameters significantly outperformed traditional models in predicting metabolic disorders. Similarly, comparative studies by Lane et al. (2020) demonstrated that, due to the capacity of ensemble learning algorithms to capture complex feature interactions with good generalization performance, they are consistently superior to traditional classifiers for a variety of biomedical prediction problems.

The key insight from this research is that the developed XGBoost model shows good predictive capability for the HIGH and INTERMEDIARY-risk classes and relatively poor performance for the LOW-risk class. The confusion matrix showed that most of the LOW-risk group were grouped into the adjacent classes resulting in a low recall for this minority group. This is most likely due to the relatively low number of LOW-risk participants, and the high clinical similarities between neighboring cardiovascular risk groups. Addressing similar issues, multiclass clinical prediction studies are often conducted with imbalanced data sets, with the sensitivity for minority classes being frequently reduced when model discrimination is acceptable (Lane et al., 2020). In future investigations, synthetic oversampling and cost sensitive learning approaches as well as larger multicenter datasets should be explored to increase the recognition rates for minority classes without degrading the overall predictive performance.

The measurement of the explainable artificial intelligence analysis also bolstered the validity of the proposed framework by showing that the model was mainly based on clinically relevant cardiovascular predictors. SHAP values showed that total cholesterol, HDL cholesterol, body weight, age, smoking history, family history of cardiovascular disease, diabetes, physical activity level, systolic blood pressure, and fasting blood sugar are the most important parameters to influence the model prediction. These results are like well-established cardiovascular risk factors known in preventive medicine and public health. High levels of lipids, poor regulation of glucose, obesity, high blood pressure, smoking and lack of physical activity combined with genetic susceptibility all play a role in the development of endothelial dysfunction and the progression of atherosclerosis, thus increasing a person's long term cardiovascular risk (Ford-Gilboe et al., 2018; Kim et al., 2022). The model building of the agreement of the SHAP is needed to make the developed model more interpretable and to increase confidence in the model's applicability to a clinical decision support system.

This investigation also follows a very strict data processing procedure. Potential leakage variables like CVD Risk Score and other clinical measurements created were cleaned up beforehand from the analysis prior to model building. In addition, the same features preprocessing methods – e.g., missing

values imputation, standardization and categorical encoding – were performed on the training set and then on the testing set. This approach enabled the model to be fairly evaluated and would minimize the risk of the model being underestimated. These types of processing methods are commonly suggested to create trustworthy machine learning systems and enhance the external validity and the reproduceability (Dinghofer & Hartung, 2020; Gavrishchaka et al., 2018). Various algorithms are compared under identical preprocessing conditions and provide a more objective evaluation of the performance of the classifiers, and a selection of the most appropriate predictive model.

In a clinical context the presented framework shows great potential of helping in cardiovascular risk identification in contemporary digital healthcare settings. The ability to integrate explainable machine learning models into EHR systems has the potential to aid healthcare providers in identifying high-risk patients, prioritizing preventive care, and resource allocation for population health management. Past studies have cited the importance of EHRs for predictive analytics and the use of AI to enhance healthcare quality via timely clinical decision support (Atasoy et al., 2019; Kruse et al., 2018). In addition, scalable machine learning frameworks with cloud-based computational power have also demonstrated great potential in improving healthcare delivery and real-time decision-making (Wang et al., 2018). However, there are a few points that should be noted. The study used a single dataset which had moderate amount of missing data and imbalanced class distribution, which influenced the prediction of the minority classes. Furthermore, the external validation with geographically distinct populations is required prior to clinical routine use. In conclusion, the current results show that explainable XGBoost-based learning is an effective and clinically interpretable method to multiclass cardiovascular risk prediction, laying the groundwork for future implementation of intelligent decision-support tools in preventive cardiovascular medicine, despite its limitations.

5.Conclusion

In this study, the authors showed that machine learning methods could effectively predict the risk of cardiovascular disease (CVD) across multiple classes based on typical demographic, anthropometric, biochemical, lifestyle, and clinical factors collected in clinical practice. The five algorithms evaluated, XGBoost had the highest predictive accuracy and discriminative ability, outperforming Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbors. Additionally, the explainable artificial intelligence analysis confirmed that the model primarily used clinically meaningful predictors, such as total cholesterol, HDL cholesterol, age, body weight, smoking status, diabetes, family history of CVD, physical activity, blood pressure, fasting blood glucose, and further improved the transparency and clinical interpretability of the predictive framework. The model was found to be reliable in identifying HIGH- and INTERMEDIARY-risk individuals, offering promises for preventing cardiovascular disease by identifying at-risk individuals for early clinical care. A thorough preprocessing technique, target leakage rejection and complete comparative analysis enhance the robustness and repeatability of the proposed framework. Future studies should incorporate external validation in larger multicenter cohorts, longitudinal clinical information, and

optimize the prediction of the minority class to increase generalizability. The results demonstrate that predictive models based on XGBoost with explainability hold great potential as a decision-support tool for precision cardiovascular risk assessment and preventive health care.

References

1. Ağbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews, 135, 110114.
2. Atasoy, H., Greenwood, B. N., & McCullough, J. S. (2019). The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. Annual review of public health, 40(1), 487-500.
3. Dinghofer, K., & Hartung, F. (2020, February). Analysis of criteria for the selection of machine learning frameworks. In 2020 International Conference on Computing, Networking and Communications (ICNC) (pp. 373-377). IEEE.
4. Ford-Gilboe, M., Wathen, C. N., Varcoe, C., Herbert, C., Jackson, B. E., Lavoie, J. G., ... & Browne, A. J. (2018). How equity-oriented health care affects health: key mechanisms and implications for primary health care practice and policy.
5. Gavrishchaka, V., Yang, Z., Miao, R., & Senyukova, O. (2018). Advantages of hybrid deep learning frameworks in applications with limited data. Int. J. Mach. Learn. Comput, 8(6), 549.
6. Hatta, M., Wahid, W. N., Yusuf, F., Hidayat, F., Santoso, N. A., & Aini, Q. (2024). Enhancing predictive models in system development using machine learning algorithms. International Journal of Cyber and IT Service Management (IJCITSM), 4(2), 80-87.
7. Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M. T., Kavousi, M., Stevens, G., ... & Di Angelantonio, E. (2019). World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. The Lancet global health, 7(10), e1332-e1345.
8. Kim, J., Mun, S., Lee, S., Jeong, K., & Baek, Y. (2022). Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea. BMC Public Health, 22(1), 664.
9. Kruse, C. S., Stein, A., Thomas, H., & Kaur, H. (2018). The use of electronic health records to support population health: a systematic review of the literature. Journal of medical systems, 42(11), 214.
10. Lane, T. R., Foil, D. H., Minerali, E., Urbina, F., Zorn, K. M., & Ekins, S. (2020). Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. Molecular pharmaceutics, 18(1), 403-415.
11. Lindstrom, M., DeCleene, N., Dorsey, H., Fuster, V., Johnson, C. O., LeGrand, K. E., ... & Roth, G. A. (2022). Global burden of cardiovascular diseases and risks collaboration, 1990-2021. Journal of the American College of Cardiology, 80(25), 2372-2425.
12. Lloyd-Jones, D. M., Braun, L. T., Ndumele, C. E., Smith Jr, S. C., Sperling, L. S., Virani, S. S., & Blumenthal, R. S. (2019). Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the American Heart Association and American College of Cardiology. Circulation, 139(25), e1162-e1177.
13. Marinho, F. (2024). The Impact of Cardiovascular Disease on Economic Loss. Arquivos Brasileiros de Cardiologia, 121(3), e20240175-e20240175.
14. Mensah, G. A., Roth, G. A., & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond. Journal of the American college of cardiology, 74(20), 2529-2532.
15. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., ... & Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990-2019:

update from the GBD 2019 study. Journal of the American college of cardiology, 76(25), 2982-3021.

16. Sartika, R. A. D., Sigit, F. S., Shukri, N. H. M., Purwanto, E., Yunita, J., & Lubis, P. N. (2025). Redefining Obesity in the Indonesian Population: The Critical Role of Waist-to-Height Ratio in Screening for Diabetes Mellitus and Hypertension. Journal of Nutrition and Metabolism, 2025(1), 5815261.
17. Sekeroglu, B., Ever, Y. K., Dimililer, K., & Al-Turjman, F. (2022). Comparative evaluation and comprehensive analysis of machine learning models for regression problems. Data intelligence, 4(3), 620-652.
18. Sheth, V., Tripathi, U., & Sharma, A. (2022). A comparative analysis of machine learning algorithms for classification purpose. Procedia Computer Science, 215, 422-431.
19. Tabary, M., Cheraghian, B., Mohammadi, Z., Rahimi, Z., Naderian, M. R., Danehchin, L., ... & Poustchi, H. (2021). Association of anthropometric indices with cardiovascular disease risk factors among adults: a study in Iran. European journal of cardiovascular nursing, 20(4), 358-366.
20. US Preventive Services Task Force, Curry, S. J., Krist, A. H., Owens, D. K., Barry, M. J., Caughey, A. B., ... & Wong, J. B. (2018). Risk assessment for cardiovascular disease with nontraditional risk factors: US preventive services task force recommendation statement. Jama, 320(3), 272-280.
21. Wang, J. B., Wang, J., Wu, Y., Wang, J. Y., Zhu, H., Lin, M., & Wang, J. (2018). A machine learning framework for resource allocation assisted by cloud computing. IEEE Network, 32(2), 144-151.
22. Zaki, A. M. (2025). Cardiovascular Disease Risk Assessment Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/ahmeduzaki/cardiovascular-disease-risk-assessment-dataset