A Comparative Analysis of the Machine Learning Models for Cardiovascular Risk Prediction

Author: Akshita Vivek

Affiliation: Livermore High School

Abstract

Cardiovascular diseases (CVDs) are among the leading causes of death worldwide and represent one of the significant public health concerns, highlighting the importance of early diagnosis for effective management and treatment. Traditional methods for predicting CVD risks have limitations and fail to capture the complex interactions among various risk factors that influence disease progression. In this study, we employed machine learning (ML) and artificial intelligence (AI) models, such as extreme gradient boosting (XGBoost), multi-layer perceptron (MLP), support vector machine (SVM), and linear discriminant analysis (LDA), on two publicly available CVD datasets to evaluate the accuracy of risk prediction for cardiovascular diseases. The results showed that the XGBoost and MLP models outperformed the SVM and LDA models in predicting CVD risks. Remarkably, with the heart disease dataset, we achieved 100% accuracy in predicting the CVD risk using XGBoost and MLP models. In this dataset, with both XGBoost and MLP models, the accuracy, precision, recall values, and AUC were 1.0. The cardiovascular disease dataset showed accuracy, precision, recall values of 0.99 and an AUC of 1.0 with the XGBoost model, while the MLP model had these values: 0.97, 0.95, 0.99, and 0.99, respectively. When comparing our results with other published research on machine learning approaches for predicting CVD risks across diverse CVD datasets, our results outperformed all other models, further emphasizing the significance of this study. These findings suggest that ML-driven CVD risk assessment could be integrated into clinical decision-support systems to help physicians in early intervention.

Introduction

Cardiovascular disease (CVD) remains the leading cause of death and illness globally, accounting for approximately 19.8 million deaths in 2022, representing approximately 32% of all global deaths [1]. This poses a significant public health challenge and immense socioeconomic burden, which manifests not only through direct healthcare expenses but also through indirect costs such as lost productivity and premature mortality [2], [3].

Among the most common forms of CVDs are coronary heart disease, stroke, and heart failure, all of which share modifiable risk factors such as hypertension (high blood pressure), hyperlipidemia (high cholesterol), smoking, tobacco use, diabetes, and obesity, physical inactivity [4]. These factors do not operate in isolation; instead, they interact in complex, often synergistic ways to accelerate disease progression. For instance, obesity is a major contributor to both high blood pressure and type 2 diabetes, which in turn significantly increases the risk of heart disease and stroke [5]. Hence, identifying and analyzing these risk factors is crucial for early diagnosis, effective management, and disease prevention. It is equally important to look at these risk factors within population diversity, as lifestyle, environment, and genetics which can significantly alter their impact.

Traditionally, heart disease risks were evaluated using prediction models, such as the Framingham Risk Score (FRS), which helped clinicians estimate a patient's 10-year risk of developing coronary heart disease based on factors including age, sex, blood pressure, cholesterol levels, and smoking history. These models, which were foundational in modern preventive cardiology, were developed using long-term health data from thousands of participants in the Framingham Heart Study. The FRS model primarily employed traditional statistical methods like logistic regression, which made it a significant advance in CVD risk prediction [6].

Despite its influence on CVD risk prediction, the FRS model has limitations. For instance, a significant limitation lies in the fact that these models assume a linear relationship between a variable (e.g., cholesterol level) and risk, and often oversimplify the complex, non-linear, synergistic interactions between risk factors, failing to capture the intricate interplay of multiple variables simultaneously influencing disease progression [7]. Furthermore, these models were derived and validated primarily on homogeneous populations, specifically a predominantly white population of European descent in a single geographic location. This has led to concerns about the generalizability and accuracy of the findings when applied to diverse racial, ethnic, or socioeconomic groups [8], [9].

Additionally, conventional risk models were biased towards gender-related disparities in their risk estimation. Because the models were trained on data where male patients cardiovascular events were historically more prominent, female patients with heart disease symptoms are more likely to be underdiagnosed or misclassified. The limitations of these conventional approaches highlight a critical need for more sophisticated, accurate, and personalized risk assessment tools that can overcome these biases and be inclusive of the vast amount of patient data now available through electronic health records (EHRs), genetic information, and wearable devices.

Recent advancements in machine learning (ML) algorithms have opened up unprecedented possibilities for improving CVD risk prediction by analyzing large and complex CVD datasets more accurately [10]. Unlike traditional statistical models, which are limited in scope by assumptions of linearity and predefined variable interactions, ML algorithms can analyze massive, high-dimensional datasets to identify intricate, non-linear patterns and relationships that are often missed. Furthermore, explainable AI techniques now enable interpretation of how specific risk factors contribute to predictions, addressing concerns about model transparency. By integrating diverse data types, including clinical data from electronic health records (EHRs), detailed lab results, genetic information, and continuous lifestyle metrics from wearable devices, ML models can develop more robust and precise risk predictions for individuals [11]. Such integration could also facilitate personalized treatment planning, ensuring that interventions are patient-specific rather than being population-generic.

A growing number of research has demonstrated the potential of various ML algorithms to outperform conventional models in predicting critical cardiovascular outcomes. Studies have shown that algorithms such as Random Forest, Support Vector Machines, and especially deep learning models, achieve superior predictive performance for events like myocardial infarction, heart failure, and stroke [11] [12] [13]. This growing number of evidence affirms that AI-driven approaches would play a transformative role in cardiovascular risk assessment, paving the way for the formulation of more effective, personalized, and preventive strategies.

In this study, we evaluated different machine learning algorithms for CVD risk prediction. Our approach achieved better performance metrics than those reported in comparable recent research studies. By demonstrating higher accuracy, sensitivity, and overall predictive power, our work highlights the potential of ML models to improve CVD risk stratification and ultimately support more personalized preventive healthcare.

Materials and Methods

For this study, we used two CVD datasets collected from two different regions of the world. The cardiovascular disease dataset was obtained from the Mendeley Data Repository https://data.mendeley.com/datasets/dzz48mvjht/1. The dataset was acquired from a multispecialty hospital in India. The dataset comprises health record data from 1000 subjects, with 13 well-defined attributes. The patient attributes and attribute statistics are described in Table 1. Using this dataset ensured cultural and demographic variation, strengthening the external validity of our model.

The heart disease dataset was obtained from Kaggle https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data. This dataset, from 1988, consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It comprises health records of 1024 subjects. The original dataset contains 76 attributes; however, only a subset of 13 of these attributes have been used consistently in most published experiments. The patient characteristics and attribute statistics are described in Table 2. This dataset is one of the most widely benchmarked in cardiovascular ML research, enabling direct comparison with prior studies

Table 1: Cardiovascular disease dataset attribute description with statistical analysis

Total number of participants: 1000

S. No	Attribute	Assigned Code	Unit Range	Statistical value (Mean, +/- SD)
1	Patient Identification Number	patientid	Number	
2	Age	age	In Years	49.24, +/-17.86
3	Gender	gender	1,0(0= female, 1 = male)	F = 23.5%; M = 76.5%
4	Chest pain type	chestpain	0,1,2,3 (Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic)	0 = 42.0%; 1 = 22.4%; 2 = 31.2%; 3 = 44.0%
5	Resting blood pressure	restingBP	94-200 (in mm HG)	151.75, +/-0.95
6	Serum cholesterol	serumcholestrol	126-564 (in mg/dl)	328.88, +/- 113.01 ['0' values removed from calculations]
7	Fasting blood sugar	fastingbloodsugar	0,1 > 120 mg/dl (0 = false , 1 = true)	0 = 70.4%; 1 = 29.6%
8	Resting electrocardiogram results	restingrelectro	0,1,2 (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)	0 = 45.4%; 1 = 34.4%; 2 = 20.2%
9	Maximum heart rate achieved	maxheartrate	71-202	145.48, +/-34.17
10	Exercise induced angina	exerciseangia	0,1 (0 = no, 1 = yes)	0 = 50.2%; 1 = 49.8%
11	Oldpeak =ST	oldpeak	0-6.2	2.7, +/-1.71
12	Slope of the peak exercise ST segment	slope	1,2,3 (1-upsloping, 2-flat, 3-downsloping)	0 = 18.0%; 1 = 29.9%; 2 = 32.2%; 3 = 19.9%
13	Number of major vessels	noofmajorvessels	0,1,2,3	0 = 27.5%; 1 = 34.4%; 2 = 26.5%; 3 = 11.6%
14	Classification	target	0,1 (0= Absence of Heart Disease, 1= Presence of Heart Disease)	0 = 42.0%; 1 = 57.9%

Table 2: Heart disease dataset attribute description with statistical analysis

Total number of participants: 1025

Attributes	Assigned Code	Statistical Value (Mean, +/- SD)
Age	age	54.43, +/-9.07
Sex	sex	Female = 30.44%; Male = 69.56%
Chest pain type (4 values, 0-3)	cp	0 = 48.49%; 1 = 16.29%; 2 = 27.71%; 3 = 7.51%
Resting blood pressure	trestbps	131.61, +/- 17.51
Serum cholesterol in mg/dl	chol	246, +/- 51.56
Fasting blood sugar > 120 mg/dl	fbs	0 = 85.07%; 1 = 14.93 %
Resting electrocardiographic results (values 0,1,2)	restecg	0 = 48.49%; 1 = 50.05%; 2 = 1.46%
Maximum heart rate achieved	thalach	149.11, +/-22.99
Exercise induced angina	exang	0 = 66.28%; 1 = 33.62%
Oldpeak = ST depression induced by exercise relative to rest	oldpeak	1.07, +/-1.17
The slope of the peak exercise ST segment (values 0,1,2)	slope	0 = 7.2%; 1 = 47.02%; 2 = 45.76%
Number of major vessels (0-3) colored by flourosopy	ca	0 = 56.39%; 1 = 22.05%; 2 = 13.07%; 3 = 6.73%; 4 = 1.76%
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect	thalach	0 = 0.68%; 1 = 6.24%; 2 = 53.07%; 3 = 40%
Classification: 0 = No disease; 1 = Disease	target	0 = 48.68%; 1 = 51.32%

Data analysis

The data analysis in this study involved applying and comparing multiple machine learning algorithms to predict the risk of cardiovascular disease (CVD). Four distinct methods were used: XGBoost, a gradient boosting ensemble model that excels at capturing nonlinear relationships and feature interactions; multilayer perceptron (MLP), a neural network capable of modeling complex decision boundaries through layers of interconnected neurons; support vector machine (SVM), which finds an optimal separating hyperplane between classes in a high-dimensional space; and linear discriminant analysis (LDA), a traditional statistical technique that projects data onto a lower-dimensional space to maximize class separability.

For each method, the dataset was split into training and testing subsets. The training-to-test ratio was 80% for training and 20% for testing. The models were trained on the training dataset before being evaluated on the test dataset. Performance was measured using a range of standard classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, specificity, and confusion matrix components such as true positives, true negatives, false positives, and false negatives). These metrics provided a comprehensive evaluation of how effectively each model identified individuals at risk versus those not at risk. The workflow of the ML-based CVD risk prediction is explained in Figure 1.

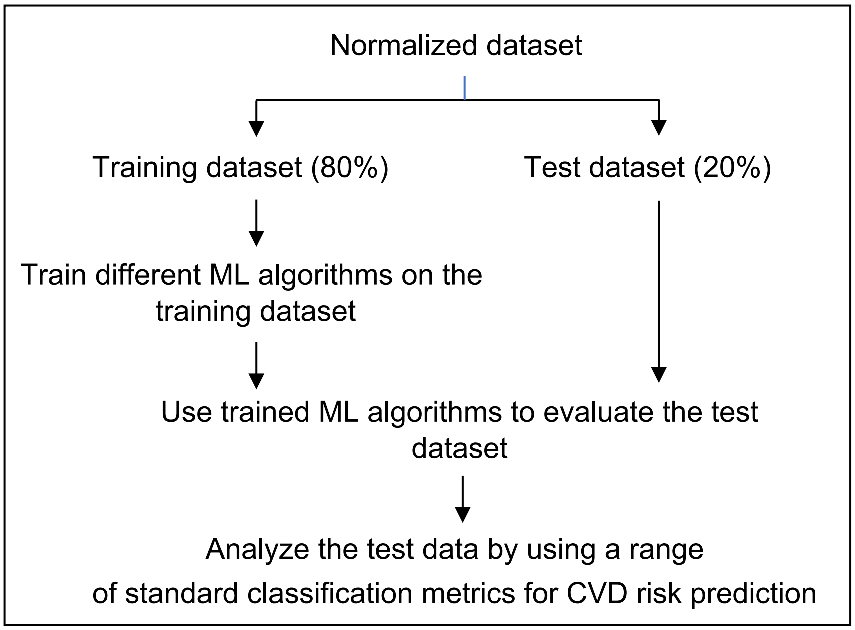


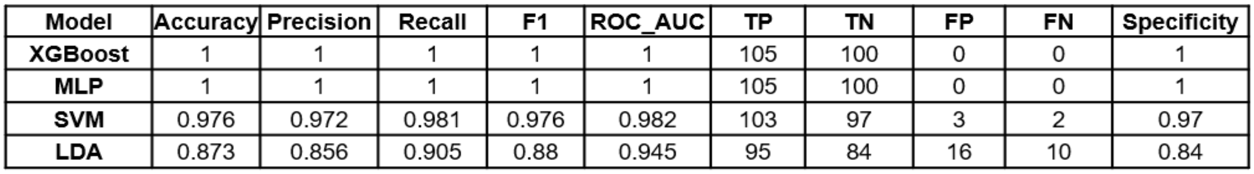
Figure 1. Workflow of the ML model-based CVD risk prediction

Result

In this study, we compared four different machine learning models for CVD risk prediction. These models were selected to represent a wide range of machine learning algorithms, ranging from traditional methods like LDA and SVM to more advanced, highly flexible models like MLP and XGBoost. This analysis approach highlights both the advantages of complex nonlinear models and the relative limitations of simpler linear approaches in the context of CVD risk assessment.

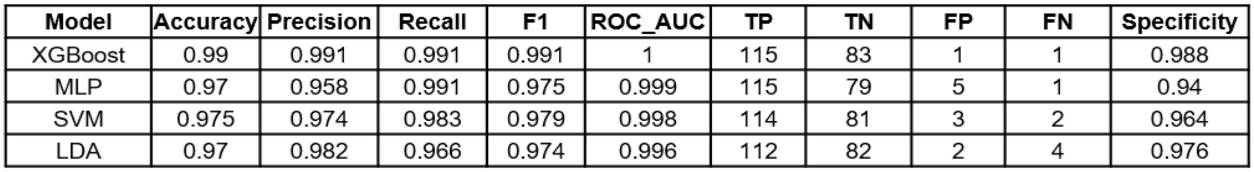
Linear Discriminant Analysis (LDA) is a supervised machine learning model primarily used for classification and dimensionality reduction. It assumes that a linear combination of features can separate data classes. It is computationally efficient and highly interpretable, making it a good starting point for comparison. We applied the LDA model to the heart disease dataset, achieving the lowest accuracy and precision among all four models. The accuracy, precision, recall, and AUC values were 0.873, 0.856, 0.905, and 0.945, respectively. In contrast, on the cardiovascular dataset, the model performed slightly better with the accuracy, precision, recall, and AUC values of 0.97, 0.982, 0.966, and 0.996, respectively (Tables 3 and 4).

Table 3. Heart disease dataset



Model	Accuracy	Precision	Recall	F1	ROC_AUC	TP	TN	FP	FN	Specificity
XGBoost	1	1	1	1	1	105	100	0	0	1
MLP	1	1	1	1	1	105	100	0	0	1
SVM	0.976	0.972	0.981	0.976	0.982	103	97	3	2	0.97
LDA	0.873	0.856	0.905	0.88	0.945	95	84	16	10	0.84

Table 4. Cardiovascular disease dataset



Model	Accuracy	Precision	Recall	F1	ROC_AUC	TP	TN	FP	FN	Specificity
XGBoost	0.99	0.991	0.991	0.991	1	115	83	1	1	0.988
MLP	0.97	0.958	0.991	0.975	0.999	115	79	5	1	0.94
SVM	0.975	0.974	0.983	0.979	0.998	114	81	3	2	0.964
LDA	0.97	0.982	0.966	0.974	0.996	112	82	2	4	0.976

Tables 3 and 4 present the accuracy, precision, recall, F1 score, and area under the curve (AUC) results from the individual machine learning models for the heart disease dataset and the cardiovascular disease dataset, respectively. ROC curve = plot of True Positive Rate (Recall) vs False Positive Rate (1 – Specificity) at different thresholds; TP = True positive; TN = True negative; FP = False positive; FN = False negative.

Support Vector Machine (SVM): SVM is a supervised learning model used for classification and regression analysis. It can handle both linear and non-linear data for analysis. It works by finding the optimal hyperplane that separates different classes of data with the largest possible margin. This method has been widely used for CVD risk prediction [10] [14] [15]. Using various kernel functions, SVM can effectively capture nonlinear relationships. On both heart disease and cardiovascular disease datasets, the SVM model was comparatively less effective compared to the XGBoost and MLP models, with accuracy, precision, recall, and AUC values being 0.976, 0.972, 0.981, and 0.982, respectively, for the heart disease dataset. While on the cardiovascular dataset, the accuracy, precision, recall, and AUC values were 0.975, 0.974, 0.983, and 0.998, respectively (Tables 3 and 4).

Multilayer Perceptron (MLP): The MLP is a supervised learning model and a foundational type of artificial neural network model. The MLP algorithm is designed to capture complex, non-linear relationships between different variables. By using multiple layers of interconnected "neurons," MLP can learn intricate patterns that are beyond the scope of traditional statistical models. Our analysis showed that the MLP model outperformed both the SVM and LDA models on both heart disease and cardiovascular disease datasets, with accuracy, precision, recall, and AUC values of 1.0 on the heart disease dataset and 0.97, 0.958, 0.991, and 0.975, respectively, for the cardiovascular disease dataset (Tables 3 and 4).

Extreme Gradient Boosting (XGBoost): This is an advanced gradient boosting algorithm that belongs to the ensemble learning family. XGBoost builds a sequence of decision trees, where each new tree corrects the errors of the previous ones. It is renowned for its exceptional performance, speed, and efficiency. It also includes regularization techniques to prevent overfitting and provides feature importance scores, offering valuable insights into which variables contribute most to the prediction. The results indicated that the XG Boost model outperformed all three other models in both datasets. The accuracy, precision, recall, and AUC values were 1.0 for the heart disease dataset, while these values were 0.99, 0.991, 0.991, and 1.0, respectively, for the cardiovascular dataset (Tables 3 and 4).

Finally, we compared our results with published research on machine learning in cardiovascular risk prediction across various datasets. A review of some of the published results is summarized in Table 5. Among all the published findings reported here, our results outperformed those of other studies, highlighting the significance of this work.

Table 5: A review of published research on using ML models to predict CDV risks.

Author	Journal	Datasets	ML model	Results
Pal M, et al.[16]	Open Medicine (2022)	UCI Heart Disease dataset (303 samples)	Multi-Layer Perceptron (MLP)	Accuracy (82.47%); Precision (79.0% with no CVD and 85.0% with CVD)
			Nearest Neighbor (K-NN)	Accuracy (73.77%); Precision (0.68% with no CVD and 0.71% with CVD)
Mathur P, et al.[14]	Clinical Medicine Insights: Cardiology (2020)	EKG + echocardiogram (CNN, ~97,000 patients), MESA & FLEMENGHO cohorts (SVM), congenital heart disease records (>44,000 records)	CNN	Accuracy (85.7%); Sensitivity (86.3%); Specificity (85.7%)
			Deep Learning	Accuracy (91.1%–97.0%)
			Support Vector Machine (SVM)	Better than ACC/AHA models
Dorraki M, et al.[17]	JACC: Advances (2024)	UK Biobank (Number of subjects = 375,145)	Decision Tree	Accuracy [68.0% (81.0% with psych factors)]; Precision [71.0% (74.0% with psych factors)]
			Random Forest	Accuracy [70.0% (84.0% with psych factors)]; Precision [76.0% (81.0% with psych factors)]
			XGBoost	Accuracy [69.0% (82.0% with psych factors)]; Precision [72.0% (77.0% with psych factors)]
			SVM	Accuracy [67.0% (80.0% with psych factors)]; Precision [70.0% (75.0% with psych factors)]
			DNN	Accuracy [70.0% (84.0% with psych factors)]; Precision [75.0% (80.0% with psych factors)]
			Ensemble (Voting)	Accuracy [71.3% (85.1% with psych factors)]; Precision [74.9% (79.6% with psych factors)]
Hossain S, et	BMC	391 CVD patients + 260	Logistic Regression	Accuracy (95.42%); Precision (93.67%)

al.[18]	Cardiovascular Disorders (2024)	controls (Bangladesh, 2022–2023)		
			Naïve Bayes	Accuracy (~96.0%); Precision (94.87%)
			Decision Tree	Accuracy (~97.0%); Precision (96.10%)
			AdaBoost	Accuracy (~96.0%); Precision (94.94%)
			Random Forest	Accuracy (98.04%); Precision (96.15)
			Bagging Tree	Accuracy (~96.0%); Precision (94.8.0%)
Salah H, et al.[19]	Scientific Reports (2025)	National Longitudinal Study of Adolescent to Adult Health (Add Health), 14,083 adolescents	XGBoost	Accuracy (84.5%); Precision (96.90%)
Matheson M, et al.[20]	Circulation Reports (2022)	Japanese cohort, employer-mandated health checkups, 155,108 adults ≥40 years	Random Survival Forests (RSF)	Accuracy >84% (coronary artery disease), >82% (atherosclerotic CVD)
Shah P, et al.[15]	Scientific Reports (2025)	IEEE Dataport CVD, Cleveland Heart Disease, Hungarian dataset (~70,000 instances, 12 features)	Gradient Boosting	Accuracy (77.5%); Precision (76.50%)
			CatBoost	Accuracy (78.5%); Precision (77.2%)
			LightGBM	Accuracy (79.5%); Precision (79.0%)
			Random Forest	Accuracy (73.2%); Precision (72.0%)
			Logistic Regression	Accuracy (76.2%); Precision (75.0%)
			Support Vector Machine	Accuracy (78.2%); Precision (77.0%)
			Neural Network	Accuracy (79.0%); Precision (78.0%)
			XGBoost	Accuracy (79.0%); Precision (78.5%)
			Hybrid Attention	Accuracy (82.0%); Precision (81.0%)

Discussion

In this comparative analysis of different ML algorithms on two CVD datasets, the models achieved nearly perfect accuracy in predicting cardiovascular disease (CVD) risk because the dataset likely contains features that are highly predictive of the outcome, such as age, gender, blood pressure, chest pain, cholesterol, blood sugar, ECG results, etc., allowing complex algorithms like XGBoost and multilayer perceptrons (MLPs) to capture the underlying patterns with ease. Both models are capable of representing complex nonlinear relationships and interactions among variables, which makes them particularly effective in medical datasets where risk factors such as cholesterol, blood pressure, blood glucose level, and age often align strongly with disease outcomes.

Another factor that would have influenced the outcome is the possibility of data leakage, which occurs if the model inadvertently has access to information that is too closely tied to the target labels. This can inflate performance metrics, so it is crucial to ensure that preprocessing and splitting of the data were carefully done. The evaluation setup also plays a role in determining the outcome. If only a single train-test split were used instead of more robust methods like stratified k-fold cross-validation. In that case, the reported results may reflect an overly optimistic estimate of performance. That said, the fact that not only accuracy but also precision, recall, specificity, and ROC-AUC are all at their maximum suggests that the dataset itself is highly separable with respect to CVD risk. In other words, the clinical variables in this case may provide nearly deterministic information about whether a patient falls into the high-risk or low-risk group, and the expressive capacity of XGBoost and MLP was sufficient to capture those relationships fully. Nevertheless, future studies should validate these models on prospective, real-world clinical cohorts to confirm generalizability.

Our results outperformed published research using machine learning techniques on diverse CVD datasets (summarized in Table 5). The closest results are those of Hossain et al., who achieved an accuracy of approximately 96-98% with ML algorithms like Naïve Bayes, Decision Tree, AdaBoost, Random Forest, and Bagging Tree (Table 5). However, comparing these results is challenging because the studies used different datasets, which are influenced by factors such as the number of study participants, subject characteristics, data quality, experimental design, data preprocessing, geographical distribution of the subjects, and other variables.

Conclusions

Our machine learning approach produced better results than those reported in other studies, which could be attributed to several key methodological and data-related factors. A clean, well-defined dataset with fewer missing values, a strong correlation between attributes and the outcome, an evaluation setup, and the use of advanced machine learning algorithms like XGBoost and MLP models that can capture complex and synergistic patterns among the variables have all contributed to these improved results. This work highlights the importance of utilizing advanced machine learning techniques on high-quality healthcare datasets, which can facilitate risk prediction, effective management, and treatment of cardiovascular diseases. In the long run, integrating such ML models into electronic health systems could enable proactive interventions and significantly reduce CVD burden globally.

References:

[1] World Health Organization (WHO), "Cardiovascular diseases (CVDs)," 2025. Accessed: Jul. 30, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] D. S. Kazi *et al.*, "Forecasting the Economic Burden of Cardiovascular Disease and Stroke in the United States Through 2050: A Presidential Advisory From the American Heart Association," *Circulation*, vol. 150, no. 4, Jul. 2024, doi: 10.1161/CIR.0000000000001258.

[3] G. A. Mensah *et al.*, "Global Burden of Cardiovascular Diseases and Risks, 1990-2022," *J Am Coll Cardiol*, vol. 82, no. 25, pp. 2350–2473, Dec. 2023, doi: 10.1016/j.jacc.2023.11.007.

[4] D. Adhikary, S. Barman, R. Ranjan, and H. Stone, "A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern," *Cureus*, Oct. 2022, doi: 10.7759/cureus.30119.

[5] Center for Disease Control and Prevention (CDC), "Heart Disease Facts," Oct. 2024. Accessed: Oct. 23, 2024. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[6] C. W. Tsao and R. S. Vasan, "The Framingham Heart Study: past, present and future," *Int J Epidemiol*, vol. 44, no. 6, pp. 1763–1766, Dec. 2015, doi: 10.1093/ije/dyv336.

[7] I. Garcha and S. P. Phillips, "Social bias in artificial intelligence algorithms designed to improve cardiovascular risk assessment relative to the Framingham Risk Score: a protocol for a systematic review," *BMJ Open*, vol. 13, no. 5, p. e067638, May 2023, doi: 10.1136/bmjopen-2022-067638.

[8] I. Talha, N. Elkhoudri, and A. Hilali, "Major Limitations of Cardiovascular Risk Scores," *Cardiovasc Ther*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/4133365.

[9] J. A. A. G. Damen *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," *BMJ*, p. i2416, May 2016, doi: 10.1136/bmj.i2416.

[10] Dr. R. Deepa, V. B. Sadu, P. G. C, and Dr. A. Sivasamy, “Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records,” *AIP Adv*, vol. 14, no. 3, Mar. 2024, doi: 10.1063/5.0191990.

[11] I. Fereydooni *et al.*, “A Systematic Review of Machine-Learning Models for Cardiovascular Risk Prediction,” *InfoScience Trends*, vol. 2, no. 7, pp. 65–82, Jul. 2025, doi: 10.61882/ist.202502.07.06.

[12] C. Krittanawong *et al.*, “Machine learning prediction in cardiovascular diseases: a meta-analysis,” *Sci Rep*, vol. 10, no. 1, p. 16057, Sep. 2020, doi: 10.1038/s41598-020-72685-1.

[13] A. Lakhani *et al.*, “Different machine learning language models for cardiovascular disease risk prediction: a systematic review,” *Int J Res Med Sci*, vol. 13, no. 1, pp. 331–339, Dec. 2024, doi: 10.18203/2320-6012.ijrms20244132.

[14] P. Mathur, S. Srivastava, X. Xu, and J. L. Mehta, “Artificial Intelligence, Machine Learning, and Cardiovascular Disease,” *Clin Med Insights Cardiol*, vol. 14, p. 117954682092740, Jan. 2020, doi: 10.1177/1179546820927404.

[15] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, “Predicting cardiovascular risk with hybrid ensemble learning and explainable AI,” *Sci Rep*, vol. 15, no. 1, p. 17927, May 2025, doi: 10.1038/s41598-025-01650-7.

[16] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, “Risk prediction of cardiovascular disease using machine learning classifiers,” *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022, doi: 10.1515/med-2022-0508.

[17] M. Dorraki *et al.*, “Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data,” *JACC: Advances*, vol. 3, no. 9, p. 101180, Sep. 2024, doi: 10.1016/j.jacadv.2024.101180.

[18] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, R. Hossain, and K. Hossain, “Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023,” *BMC Cardiovasc Disord*, vol. 24, no. 1, p. 214, Apr. 2024, doi: 10.1186/s12872-024-03883-2.

[19] H. Salah and S. Srinivas, “Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents,” *Sci Rep*, vol. 12, no. 1, p. 21905, Dec. 2022, doi: 10.1038/s41598-022-25933-5.

[20] M. B. Matheson, Y. Kato, S. Baba, C. Cox, J. A. C. Lima, and B. Ambale-Venkatesh, “Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort,” *Circ Rep*, vol. 4, no. 12, p. CR-22-0101, Dec. 2022, doi: 10.1253/circrep.CR-22-0101.