Data analytics and prediction of cardiovascular disease with machine learning models: a systematic literature review

Ravipa Sonthana[1], Sakchai Tangprasert[1], Yuenyong Nilsiam[2], Nalinpat Bhumpenpein[3], Siranee Nuchitprasitchai[3]

[1]Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[2]Department of Electrical and Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[3]Department of Information Technology, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Article Info

Article history:

Received Apr 4, 2025
Revised Dec 18, 2025
Accepted Jan 16, 2026

Keywords:

Cardiovascular disease
Data analytics
Machine learning
Prediction
Systematics literature review

ABSTRACT

Cardiovascular disease (CVD) remains one of the leading causes of death globally, underscoring the need for effective early risk prediction. This systematic literature review analyzes research published between 2013 and 2023 on the application of machine learning (ML) in CVD risk prediction. Key areas examined include feature selection, data preprocessing, algorithm choice, and model evaluation. Studies were selected from ACM Digital Library, IEEE Xplore, ScienceDirect, and Scopus based on predefined research questions. Common challenges include limited or low-quality datasets, inconsistent preprocessing methods, and the need for clinically interpretable models. Widely used algorithms include random forest (RF), support vector machine (SVM), decision tree (DT), logistic regression (LR), naïve Bayes (NB), k-nearest neighbor (K-NN), and extreme gradient boosting (XGBoost). The review highlights that robust preprocessing, optimal feature selection, and thorough model validation significantly improve predictive accuracy. It also emphasizes the importance of balancing performance with interpretability for clinical adoption. Finally, the study proposes a structured framework to guide future research and practical implementation, including the integration of genetic and behavioral data to support more personalized and effective cardiovascular care.

Corresponding Author:

Ravipa Sonthana
Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand
Email: ravipa.sgn@gmail.com

1. INTRODUCTION

Each year, statistics from the centers for disease control and prevention (CDC) highlight the growing global burden of cardiovascular disease (CVD), a leading cause of death worldwide [1]. Early detection is crucial [2], particularly for individuals with chronic conditions such as kidney disease [3], diabetes [4], and dyslipidemia [5], who are at higher risk. Consequently, researchers increasingly adopt machine learning (ML) for large-scale data analysis, pattern recognition, and personalized prediction.

Recent advances in ML, including random forest (RF), support vector machine (SVM), extreme gradient boosting (XGBoosting), and artificial neural networks (ANN), outperform traditional methods such as logistic regression (LR) and naïve Bayes (NB). Challenges remain, including heterogeneous datasets, inconsistent preprocessing, variable feature selection, and limited interpretability [6], which hinder clinical adoption. Despite numerous studies on ML-based CVD prediction, no comprehensive synthesis evaluates the combined effects of preprocessing, feature selection, algorithm choice, and interpretability on model performance. This systematic literature review (SLR) of studies from 2013–2023 aims to address this gap, providing a holistic understanding of methodological trends, challenges, and opportunities for improving predictive accuracy and clinical applicability.

This systematic review investigates ML applications in CVD prediction to develop an effective risk assessment framework, employing a SLR to define scope, identify and evaluate relevant studies, and synthesize outcomes both qualitatively and quantitatively. The review specifically addresses the following research questions:

RQ1: What obstacles do researchers face when implementing machine learning for cardiovascular prediction, and how have successful studies overcome these challenges?

RQ2: Which physiological, behavioral, and demographic factors prove most significant in machine learning-based cardiovascular risk assessment?

RQ3: What preprocessing and data processing methods best improve prediction accuracy in ML-based CVD prediction?

RQ4: Which machine learning techniques demonstrate superior performance in cardiovascular prediction tasks?

RQ5: What methods best evaluate the clinical reliability of machine learning predictions?

The predictive success of ML in CVD risk assessment depends not only on algorithmic sophistication but also on appropriate preprocessing, feature engineering, and interpretability. These interconnected elements collectively determine whether ML models can deliver clinically reliable, explainable, and actionable predictions, facilitating their translation into real-world healthcare settings.

The remainder of this article is organized as follows. Section 2 details the methodology of the review. Section 3 presents result and discuss the findings, highlighting methodological trends, challenges, and opportunities compared to previous studies. Section 4 concludes with key implications, limitations, and directions for future research to advance ML-based CVD risk prediction.

2. RESEARCH METHOD

This research began with the formulation of research questions and the delineation of the study scope, which informed the selection of keywords for literature retrieval. Four major databases: ACM Digital Library, IEEE Xplore, ScienceDirect, and Scopus, were systematically searched. Retrieved studies were subjected to a rigorous four-stage screening process, including keyword filtering, title assessment, content evaluation, and reliability verification, to ensure the inclusion of studies most pertinent to the research objectives.

A systematic literature review was conducted to examine machine learning applications in cardiovascular disease risk prediction. Following established protocols, the review highlighted key patterns, data interpretation challenges, and insights on risk factors, modelling approaches, and clinical validation [7], emphasizing AI's role in preprocessing and deep learning for detecting subtle risk indicators.

2.1. Search strategy

A systematic search was performed across ACM Digital Library, IEEE Xplore, ScienceDirect, and Scopus, guided by research questions (RQ1–RQ5). Search strings were constructed using predefined keywords, synonyms, and Boolean operators.

a. Medical context: CVD-related studies were identified using keywords such as "cardiovascular disease," "CVD," "prediction," "forecasting," combined as:

X = {(Cardiovascular Disease OR CVD) AND (Prediction OR Predicting OR Predictive OR Forecast OR Forecasting)}.

b. Technical context: ML studies were targeted using "machine learning" and "deep learning":

Y = {Machine Learning OR Deep Learning}

For each *RQ*, *X*, and *Y* were combined with additional terms (*e.g.*, problem, factors, process, algorithm, accuracy). Challenges such as inconsistent terminology, missing ML references, and broadly categorized studies were resolved using synonym lists, iterative searches, and manual screening, consistent with systematic review protocols.

2.2. Filtering process

Practical challenges, such as inconsistent terminology and labeling of ML studies, were addressed through synonym lists, iterative searches, and manual review. Articles were sequentially filtered through four stages: keyword-based search, title relevance, abstract evaluation, and full-text assessment.

a. Initial keyword-based search phase: The initial search across ACM, IEEE, ScienceDirect, and Scopus databases identified 6,265 articles using predefined keywords aligned with the research objectives.
b. Title relevance filtration phase: Titles were screened for explicit relevance and clarity in addressing cardiovascular disease prediction with machine learning. This step reduced the pool to 621 articles.
c. Abstract analysis evaluation phase: Abstracts of the shortlisted studies were examined to confirm substantive methodological alignment and document availability, narrowing the selection to 256 articles.
d. Full-text content evaluation phase: Comprehensive review of full-text manuscripts ensured contextual validity, methodological soundness, and complete alignment with the research questions, leading to the final inclusion of 91 articles in the systematic review.

2.3. Data extraction

Data extraction involves selecting relevant studies for analysis and documenting them according to review protocols, ensuring extracted data align with the targeted study categories [8]. This phase requires careful consideration of potential conflicts and data limitations, necessitating systematic collection practices aligned with research design and implementation parameters [7], enabling efficient extraction and aggregation of relevant information [9].

2.4. Data analytic

The analytical process incorporates statistical and scientific methodologies to synthesize individual study effects, generating comprehensive results from aggregated study data [7]. The synthesis may encompass both qualitative and quantitative data from verified sources, enhancing the reliability of research outcomes [10]. The breadth of information integrated within the systematic literature review correlates positively with the confidence level in analytical conclusions.

2.5. Data synthesis process

The analytical framework guided study categorization and reference management, with EndNote X9 and Google Spreadsheet supporting data storage and tracking. Some limitations remain, including incomplete database coverage, terminology variations, and the lack of standardized metrics for machine learning data evaluation.

3. RESULTS AND DISCUSSION

This systematic review examined 74 studies (2013–2023) on ML applications in CVD prediction. Guided by five research questions, the review highlights advance in ML-based risk assessment, ranging from early detection to acute event prediction, and identifies key patterns across cardiology research.

3.1. Obstacles in machine learning-based cardiovascular risk analysis and prediction

RQ1 examined technical challenges in ML-based cardiovascular prediction, identifying four key obstacles affecting accuracy [11]. These challenges include data quality, feature selection, model training, and validation methodologies [12]. Each represents a critical consideration for developing robust ML solutions for cardiovascular risk prediction [13], as detailed in Table 1.

Table 1. Table summarizing what problems or limitations are obstacles in analyzing and predicting cardiovascular disease risk by machine learning

No	Problems/Limitations	Solution	Research
1	Using small datasets and their reliability	Use big datasets and be more comprehensive by increasing the size of the data set and using data sets from reliable and accepted sources.	[11]–[17]
2	Examining datasets potential, missingness, and data asymmetries	Cleansing data, extracting data, and selection are performed to reduce vulnerabilities in the dataset that negatively impact analysis.	[11], [13]–[15], [18]–[21]
3	Important features in a dataset used for prediction	Select a dataset that has a relevant and comprehensive feature set.	[11], [14], [16], [17], [21]–[25]
4	Split data for training and testing the model	Divide the training and testing data set into multiple ratios and compare the performance of the model for each ratio.	[14], [15], [19]

From Table 1, the analysis identified four primary limitations in cardiovascular disease risk prediction:

a. Dataset size and reliability constraints: Small datasets with limited features and data points from unreliable sources frequently result in reduced analysis accuracy. Recent improvements have focused on utilizing comprehensive, reliable datasets to enhance prediction accuracy [12]. Muhammad *et al.* [16] highlighted that small datasets and unreliable sources raise concerns regarding result generalizability.
b. Data quality assessment: Evaluating dataset potential, missing values, and data asymmetries is crucial. Data cleaning processes are essential for ensuring analysis efficacy, as improper handling may negatively impact forecasting performance [18]. Ramesh and Pathinarupothi [21] determined that abnormal values, missing data, and incomplete entries significantly reduce model performance.
c. Feature selection complexity: The multifactorial nature of cardiovascular disease complicates the identification of key predictive features. Hossain *et al.* [11] proposed employing feature engineering to extract and transform prediction-relevant features, enhancing model performance. Nagaraju *et al.* [22] suggested the Relief method to filter data and select interconnected relevant features.
d. Training-testing data distribution: The absence of standardized approaches for dataset division presents ongoing challenges [14]. Therefore, Uddin and Halder [19] implemented the "train-test-split" methodology to evaluate multiple data proportions (80:20, 70:30, 60:40, 50:50) for optimal results, addressing training dataset imbalance issues.

3.2. Factors in cardiovascular risk analysis and prediction

RQ2 examined key physiological, behavioral, and demographic predictors of cardiovascular risk in ML applications [12]. Seven major CVD datasets were identified, predominantly hospital or public sources [26], with feature selection employed to optimize model training [27]. Core predictors included blood pressure, diabetes, cholesterol, pulse rate, exercise, smoking, age, and gender [28]. Despite these efforts, dataset diversity remains limited; further details are in Table 2.

Table 2 presents 13 key predictors for CVD risk assessment, integrating physiological, lifestyle, comorbid, and demographic factors. In addition to common attributes, kidney disease and diabetes were included as chronic risk factors. The framework comprises age, sex, height, weight, BMI, exercise-induced angina (Exang), physical activity, smoking, stroke history, alcohol intake, diabetes, kidney disease, and CVDs as the target variable, aiming to enhance ML predictive accuracy.

Table 2. Factors in cardiovascular risk analysis and prediction using machine learning

No	Data source	Attributes (n)	Attributes description	Research
1	UCI heart disease dataset	13	age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	[12], [20], [26], [27], [29]–[33] [34]–[41]
2	Ludwigshafen risk and cardiovascular health (LURIC) cohort	20	age, sex, weight, total cholesterol, HDL cholesterol, LDL cholesterol, cholesterol, triglycerides, LDL triglycerides, HDL triglycerides, type II diabetes, urea, uric acid, glycosylated hemoglobin, interleukin-6, oxidized LDL, history of arterial hypertension, heart rate, systolic blood pressure, diastolic blood pressure	[17]
3	Framingham	15	age, sex, education, current smoke, cigarettes per day, BP Meds, Prevalent Stroke, Prevalent Hyp, Diabetes, Tot Chol, Sys BP, Dia BP, BMI, Heart Rate, Glucose	[20], [40], [42], [43]
4	Kaggle machine learning repository	11	age, height, weight, gender, ap_hi, ap_lo, cholesterol, glucose, smoking, alcohol intake, physical activity	[13], [19], [24], [37], [44]–[47],
5	Kaggle machine learning repository, collected from labs, hospitals and friends' data	14	age, height, weight, gender, ap_hi, ap_lo, cholesterol, glucose, smoking, alcohol intake, physical activity	[28]
6	MONICA dataset	11	age, sex, Yronset, Premi, Smstat, diabetes, highbp, hicho, angina, stroke, hosp	[48]
7	Department of Computing of Goldsmiths University of London	14	age, blood pressure, cholesterol, maximum heart rate, peak, colored vessels, sex, chest pain type, resting ecg, slope, thal, Fasting blood sugar <120', 'angina'	[49]

3.3. Effective data processing methodology for enhanced prediction accuracy

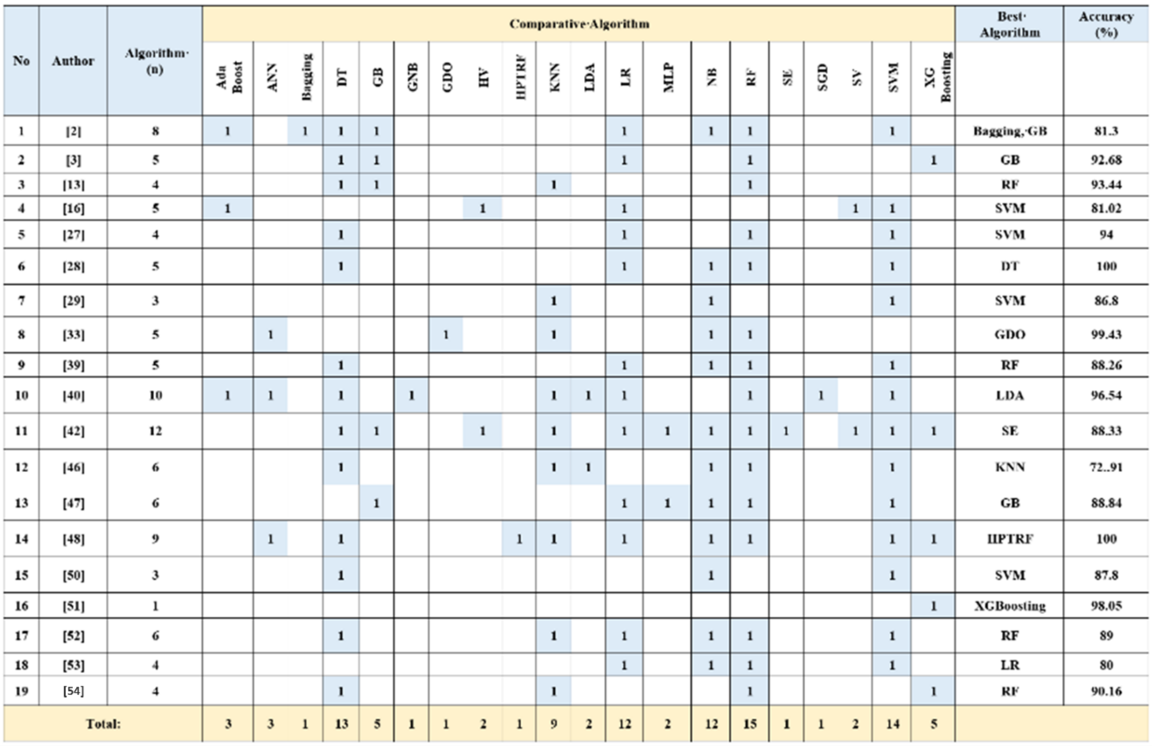
RQ3 examined advanced frameworks for cardiovascular risk prediction, comprising five stages.

a. Data acquisition and preprocessing ensure dataset validity through outlier removal and attribute optimization [19], to facilitate the acquisition of comprehensive quantitative and qualitative datasets aligned with cardiovascular risk assessment objectives [27], while ensuring methodological robustness.
b. Data analysis and transformation employ computational techniques [26], to optimize structure and standardize heterogeneous parameters.

c. Feature selection optimization enhances classification precision while mitigating sparsity and reducing computational complexity [34]. This approach facilitates the identification of critical cardiovascular risk predictors.
d. Model development enhancement integrates clinical parameters to improve predictive capabilities and learning efficiency [32].
e. Performance evaluation applies established metrics [33] to validate models and refine predictive accuracy.

3.4. Machine learning techniques for cardiovascular disease prediction

RQ4 investigated the performance of machine learning algorithms for cardiovascular disease prediction. The review identified twenty distinct techniques, each exhibiting varying levels of effectiveness and implementation challenges, as illustrated in Figure 1. Based on this analysis, eight representative methods were selected, with the addition of ANN to provide a more comprehensive comparative evaluation.



No	Author	Algorithm (n)	Comparative Algorithm																				Best Algorithm	Accuracy (%)
			Ada Boost	ANN	Bagging	DT	GB	GNB	GDO	HV	HPTRF	KNN	LDA	LR	MLP	NB	RF	SE	SGD	SV	SVM	XG Boosting		
1	[2]	8	1		1	1	1							1		1	1				1		Bagging, GB	81.3
2	[3]	5				1	1							1			1					1	GB	92.68
3	[13]	4				1	1					1					1						RF	93.44
4	[16]	5	1							1				1						1	1		SVM	81.02
5	[27]	4				1								1			1				1		SVM	94
6	[28]	5				1								1		1	1				1		DT	100
7	[29]	3										1				1					1		SVM	86.8
8	[33]	5		1					1			1				1	1						GDO	99.43
9	[39]	5				1								1		1	1				1		RF	88.26
10	[40]	10	1	1		1		1				1	1	1			1		1		1		LDA	96.54
11	[42]	12				1	1			1		1		1	1	1	1	1		1	1	1	SE	88.33
12	[46]	6				1						1	1			1	1				1		KNN	72..91
13	[47]	6					1							1	1	1	1				1		GB	88.84
14	[48]	9		1		1					1	1		1		1	1				1	1	HPTRF	100
15	[50]	3				1										1					1		SVM	87.8
16	[51]	1																				1	XGBoosting	98.05
17	[52]	6				1						1		1		1	1				1		RF	89
18	[53]	4												1		1	1				1		LR	80
19	[54]	4				1						1					1					1	RF	90.16
Total:			3	3	1	13	5	1	1	2	1	9	2	12	2	12	15	1	1	2	14	5		

Figure 1. Machine learning techniques for cardiovascular disease prediction

Based on Figure 1, the research methodology incorporated eight analytical and forecasting techniques, including an additional ANN approach, representing a novel comprehensive comparison. The techniques can be categorized as follows:

a. Tree-based methods:
   - Random forest (RF): Excels in both regression and classification tasks, particularly with large-scale nonlinear datasets. Its feature selection capability enhances predictor identification, optimizing efficiency and accuracy.
   - Decision tree (DT): Offers straightforward implementation for regression and classification, effectively summarizing complex decisions. Widely adopted in medical applications.
   - Extreme gradient boosting (XGBoosting): An enhanced iteration of gradient boosting, utilizing sequential decision trees for model training. Each iteration learns from predecessor error values, maximizing predictive accuracy while minimizing computational resources.

b. Statistical learning methods:
   - Support vector machine (SVM): Specializes in binary classification for complex, high-dimensional datasets. Demonstrates particular efficacy in ambiguous data classification with moderate sample sizes.

- Logistic regression (LR) : Primarily applied in binary classification for decision-making and risk assessment, with extensive implementation across medical research.
- Naïve Bayes (NB): Employs Bayesian probability theory for classification tasks, requiring labeled data for supervised learning. Specializes in predictive analysis based on historical probability patterns.

c. Instance-based learning:

K-nearest neighbor (K-NN): Versatile in both classification and regression, utilizing proximity principles for class assignment. Particularly suited for numerical data and multi-class classification through feature-distance measurement.

d. Neural network approach:

Artificial neural network (ANN): Demonstrates versatility in regression and classification tasks, employing brain-inspired learning mechanisms for complex problem-solving. Exhibits particular strength in predictive modelling through experiential learning.

Frequency analysis from Figure 1 indicated that RF (15 implementations), SVM (14), DT (13), LR (12), NB (12), K-NN (9), and XGBoosting (5) were the most commonly applied algorithms in cardiovascular prediction tasks, with RF being the most frequently adopted.

3.5. Model validation strategies for clinical reliability

RQ5 investigated methodologies for validating cardiovascular prediction models. Nine evaluation approaches were identified, grouped into three categories:

- Primary performance metrics: Accuracy, ROC–AUC curve, and confusion matrix, providing fundamental assessment of overall predictive capability and classification effectiveness.
- Advanced performance indicators: F1-Score, sensitivity/recall, and specificity, enabling detailed performance analysis for individual classes and measurement of true negative rates.
- Statistical validation methods: Precision, macro average, and weighted average, offering class-specific accuracy evaluation and averaging methods suitable for datasets with unequal class distribution.

Table 3 summarizes the application frequency of these metrics across the reviewed studies, highlighting accuracy, sensitivity/recall, precision, F1-Score, ROC–AUC, confusion matrix, and specificity as the most commonly utilized. Based on implementation needs and research objectives, five key metrics—accuracy, sensitivity/recall, precision, F1-Score, and confusion matrix—were selected for robust model validation.

Table 3. Model validation strategies for clinical reliability

No	Model Evaluation	Research
1	Accuracy	[12], [26], [28], [29], [33], [42], [46], [47], [50]–[54]
2	ROC – AUC Curve	[12], [15], [21], [28], [33], [50], [51]
3	Confusion Matrix	[12], [21], [33], [46], [50]
4	F1-Score	[21], [29], [42], [46], [47], [50], [51], [53], [54]
5	Sensitivity/Recall	[21], [26], [29], [33], [42], [46], [47], [50], [51], [53], [54]
6	Specificity	[26], [29], [33], [50]
7	Precision	[21], [26], [29], [42], [46], [47], [50], [51], [53], [54]
8	Macro Average	[53]
9	Weight Average	[53]

The systematic review of existing cardiovascular disease prediction research facilitated the development of an optimized analytical framework. This framework, illustrated in Figure 2, integrates established methodologies to enhance computational efficiency and mitigate analytical challenges in machine learning-based cardiovascular risk prediction. The proposed structure streamlines data processing while maintaining robust predictive capabilities.

As illustrated in Figure 2, the analytical framework for cardiovascular disease risk prediction integrates systematic machine learning methodologies across three distinct operational steps. Based on extensive literature review, this framework optimizes computational processes while ensuring prediction accuracy.

Step 1. Data preparation:

This stage encompasses data collection, cleansing, feature selection, and data splitting. Data were obtained from the behavioral risk factor surveillance system (BRFSS), ensuring reliable large-scale health information. Cleansing involved removing duplicates, correcting inconsistencies, filtering outliers, and managing missing values to enhance data integrity. Feature selection extracted salient predictors through systematic screening, while data splitting divided records into training and testing sets for unbiased evaluation.

Step 2. Model implementation:

Eight classification algorithms—RF, SVM, DT, LR, NB, K-NN, XGBoosting, and ANN—were applied. Each was trained on the prepared dataset and validated to compare predictive performance.

Step 3. Performance assessment:

Model evaluation involved calculating accuracy, sensitivity, precision, F1-Score, and confusion matrix metrics, enabling comparative analysis of predictive capability and computational efficiency to identify optimal techniques.

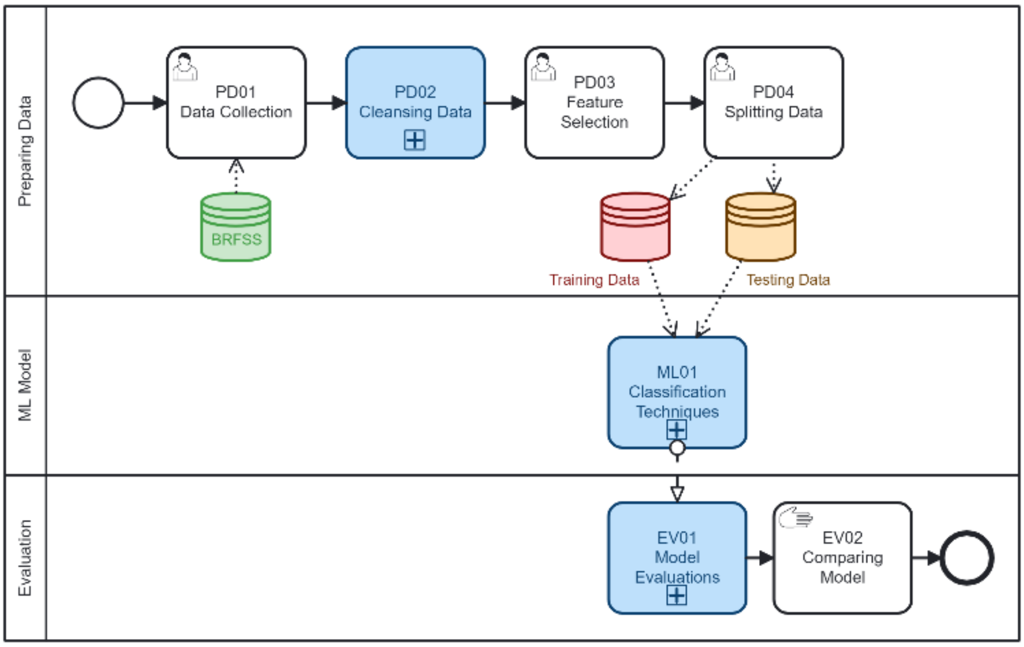


Figure 2. Machine learning framework for cardiovascular disease risk analytics and prediction

4. CONCLUSION

This systematic literature review synthesized a decade of research on the application of machine learning for cardiovascular disease prediction. The evidence demonstrates that predictive performance depends not only on algorithm selection but also on preprocessing, feature selection, and interpretability. Advanced models such as random forest, support vector machines, and XGBoosting consistently outperformed traditional linear methods, while deep learning approaches showed strong predictive potential in handling complex and large-scale datasets. However, challenges of interpretability, heterogeneous data quality, and limited clinical adoption remain significant barriers.

The thesis of this paper is that the success of ML in CVD risk assessment lies in the combined strength of algorithmic sophistication, data preprocessing, feature engineering, and explainability was supported by the reviewed evidence. By integrating these factors, ML-based models can move closer to achieving clinically reliable and meaningful outcomes.

Future research should focus on the development of interpretable ML frameworks, the integration of multimodal health data, and the design of privacy-preserving approaches such as federated learning. Addressing these issues will accelerate the transition of ML-based CVD prediction from research to clinical practice, ultimately improving patient outcomes and contributing to more personalized, data-driven healthcare systems.

FUNDING INFORMATION

Authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. K. Debnath, S. Malik, G. Kaur, S. Bagchi, A. M. Soomro, and A. Naeem, “Prediction accuracy improvement for cardiovascular diseases using machine learning algorithm,” in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2023, pp. 1032–1037, doi: 10.1109/UPCON59197.2023.10434559.

[2] I. Dawar and S. Wadhwan, “Predicting cardiovascular disease using machine learning techniques,” in *Conference Proceedings - 2023 IEEE Silchar Subsection Conference, SILCON 2023*, 2023, pp. 1–6, doi: 10.1109/SILCON59133.2023.10405318.

[3] J. Jamaluddin, M. S. Mohamed-Yassin, S. N. Jamil, M. A. Mohamed Kamel, and M. Y. akob Yusof, “Frequency and predictors of inappropriate medication dosages for cardiovascular disease prevention in chronic kidney disease patients: A retrospective cross-sectional study in a Malaysian primary care clinic,” *Heliyon*, vol. 9, no. 4, p. e14998, 2023, doi: 10.1016/j.heliyon.2023.e14998.

[4] B. V Howard and M. F. Magee, “Diabetes and cardiovascular disease,” *Current Atherosclerosis Reports*, vol. 2, no. 6, pp. 476–481, 2000.

[5] M. Hedayatnia and others, “Dyslipidemia and cardiovascular disease risk among the MASHAD study population,” *Lipids in Health and Disease*, vol. 19, pp. 1–11, 2020.

[6] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?,” *PLoS ONE*, vol. 12, no. 4, p. e0174944, 2017, doi: 10.1371/journal.pone.0174944.

[7] A. Pollock and E. Berge, “How to do a systematic review,” *International Journal of Stroke*, vol. 13, no. 2, pp. 138–156, 2018.

[8] E. Aromataris and A. Pearson, “The systematic review: an overview,” *AJN The American Journal of Nursing*, vol. 114, no. 3, 2014.

[9] R. W. Wright, R. A. Brand, W. Dunn, and K. P. Spindler, “How to write a systematic review,” *Clinical Orthopaedics and Related Research*, vol. 455, 2007.

[10] J. Thompson Coon and others, “Developing methods for the overarching synthesis of quantitative and qualitative evidence,” *Research Synthesis Methods*, vol. 11, no. 4, pp. 507–521, 2020, doi: 10.1002/jrsm.1383.

[11] M. M. Hossain *et al.*, “Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI,” *Informatics in Medicine Unlocked*, vol. 42, p. 101370, 2023, doi: 10.1016/j.imu.2023.101370.

[12] H. Lyu, “A machine learning-based approach for cardiovascular diseases prediction,” in *ACM International Conference Proceeding Series*, 2022, pp. 59–66, doi: 10.1145/3529836.3529863.

[13] J. Kensarin, V. M. Arul Xavier, U. J. V. Sai, S. S. Srujan, and K. V. Prakash, “Prediction of cardiovascular disease risk using machine learning models,” in *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, 2023, pp. 977–982, doi: 10.1109/ICACCS57279.2023.10112763.

[14] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, “A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data,” *Medical Engineering and Physics*, vol. 105, p. 103825, 2022, doi: 10.1016/j.medengphy.2022.103825.

[15] H. Zheng, S. W. A. Sherazi, S. Arif, M. J. Lee, and J. Y. Lee, “A voting ensemble-based model to predict the risk of cardiovascular disease in ordinary people,” in *Proceedings of the 2023 7th International Conference on High Performance Compilation, Computing and Communications*, Jun. 2023, pp. 127–133, doi: 10.1145/3606043.3606061.

[16] G. Muhammad *et al.*, “Enhancing prognosis accuracy for ischemic cardiovascular disease using k nearest neighbor algorithm: a robust approach,” *IEEE Access*, vol. 11, pp. 97879–97895, 2023, doi: 10.1109/ACCESS.2023.3312046.

[17] K. Tsarapatsani *et al.*, “Machine learning models for cardiovascular disease events prediction,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2022, pp. 1066–1069, doi: 10.1109/EMBC48229.2022.9871121.

[18] R. Poonkuzhali, S. Pavithra, K. S. Kumar, and C. Nallusamy, “Heart disease prediction using machine learning techniques,” *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10725519.

[19] M. N. Uddin and R. K. Halder, “An ensemble method based multilayer dynamic system to predict cardiovascular disease,” *Informatics in Medicine Unlocked*, vol. 24, p. 100584, 2021, doi: 10.1016/j.imu.2021.100584.

[20] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, “An integrated machine learning framework for effective prediction of cardiovascular diseases,” *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[21] H. V. Ramesh and R. K. Pathinarupothi, “Performance analysis of machine learning algorithms to predict cardiovascular disease,” *2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023*, 2023, doi: 10.1109/I2CT57861.2023.10126428.

[22] M. E. Nagaraju, M. V. Amrutha, R. Harika, V. Manikanta, and P. S. Datta, “An early prediction of cardiovascular disease using random forest bagging method,” in *Proceedings of the 2nd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2023*, 2023, pp. 800–805, doi: 10.1109/ICAAIC56838.2023.10141319.

[23] H. B. Kibria and A. Matin, “The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach,” *Computational Biology and Chemistry*, vol. 98, 2022, doi: 10.1016/j.compbiolchem.2022.107672.

[24] M. Ananthi, A. S. Narayanan, T. P. Dhiraj Prasad, and R. Jai Vignesh, “Cardiovascular disease prediction using randelistic algorithm,” *Proceedings of the International Conference on Circuit Power and Computing Technologies, ICCPCT 2023*, pp. 20–25, 2023, doi: 10.1109/ICCPCT58313.2023.10245957.

[25] D. M. R, S. Kuwelkar, and R. Sivakumar, “An hybrid technique for optimized clustering of EHR using binary particle swarm and constrained optimization for better performance in prediction of cardiovascular diseases,” *Measurement: Sensors*, vol. 25, 2023, doi: 10.1016/j.measen.2022.100577.

[26] B. Anishfathima, R. Vikram, S. R. T, M. Sri Vishnu, and C. Venumadhav, “A comparative analysis on classification models to predict cardio-vascular disease using machine learning algorithms,” in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Feb. 2022, pp. 259–264, doi: 10.1109/ICAIS53314.2022.9741831.

[27] P. Saraswathi, P. J G, M. Prabha, and J. N A, “Machine learning based cardiovascular disease prediction,” in *2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE)*, Nov. 2023, pp. 507–511, doi: 10.1109/AECE59614.2023.10428257.

[28] D. R. Krithika and K. Rohini, "Ensemble based prediction of cardiovascular disease using bigdata analytics," in *Proceedings - 2021 International Conference on Computing Sciences, ICCS 2021*, 2021, pp. 42–46, doi: 10.1109/ICCS54944.2021.00017.

[29] N. Louridi, M. Amar, and B. El Ouahidi, "Identification of cardiovascular diseases using machine learning," in *2019 7th Mediterranean Congress of Telecommunications (CMT)*, Oct. 2019, pp. 1–6, doi: 10.1109/CMT.2019.8931411.

[30] R. Haque *et al.*, "Evaluating the efficacy of feature selection methods in cardiovascular disease prediction with machine learning," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, Dec. 2023, pp. 1–6, doi: 10.1109/EICT61409.2023.10427833.

[31] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[32] P. Srinivas and R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomedical Signal Processing and Control*, vol. 73, p. 103456, Mar. 2022, doi: 10.1016/j.bspc.2021.103456.

[33] M. S. Nawaz, B. Shoaib, and M. A. Ashraf, "Intelligent cardiovascular disease prediction empowered with gradient descent optimization," *Heliyon*, vol. 7, no. 5, p. e06948, May 2021, doi: 10.1016/j.heliyon.2021.e06948.

[34] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data and Knowledge Engineering*, vol. 140, p. 102042, 2022, doi: 10.1016/j.datak.2022.102042.

[35] A. A. Romalt and R. M. S. Kumar, "Prediction of cardio vascular disease by deep learning and machine learning-a combined data science approach," in *2022 International Conference on Computer, Power and Communications (ICCPC)*, 2022, pp. 83–85, doi: 10.1109/ICCPC55978.2022.10072141.

[36] D. C. Yadav, S. Pal, R. K. Yadav, and H. Pant, "Analyzing risk elements in cardiovascular diseases prediction using neural networks algorithm," in *Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023*, 2023, pp. 262–266.

[37] K. D. P., P. Rao, K. V B., N. S. P., and P. R. Kamath, "Detection and analysis of cardiovascular diseases using machine learning techniques," in *2023 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 2023, pp. 258–262, doi: 10.1109/DISCOVER58830.2023.10316703.

[38] S. Tyagi, P. Sirohi, and P. Maheshwari, "Predicting cardiovascular disease in patients with machine learning and feature engineering techniques," in *2022 5th International Conference on Signal Processing and Information Security (ICSPIS)*, 2022, pp. 107–112, doi: 10.1109/ICSPIS57063.2022.10002692.

[39] Mahaveer, Puneet, and Deepika, "Cardiovascular disease prediction analysis using classification techniques," in *2022 IEEE Delhi Section Conference, DELCON 2022*, 2022, pp. 1–6, doi: 10.1109/DELCON54057.2022.9753356.

[40] S. Kusuma and K. R. Jothi, "Cardiovascular disease prediction and comparative analysis of varied classifier techniques," in *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*, 2021, pp. 1–7, doi: 10.1109/GCAT52182.2021.9587734.

[41] R. Li, S. Yang, and W. Xie, "Cardiovascular disease prediction model based on logistic regression and euclidean distance," in *Proceedings - 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering, AEMCSE 2021*, 2021, pp. 711–715, doi: 10.1109/AEMCSE51986.2021.00147.

[42] Z. Rustamov, J. Rustamov, M. S. Sultana, J. Ywei, V. Balakrishnan, and N. Zaki, "Cardiovascular disease prediction using ensemble learning techniques: a stacking approach," in *2023 19th IEEE International Colloquium on Signal Processing and Its Applications, CSPA 2023 - Conference Proceedings*, 2023, pp. 93–98, doi: 10.1109/CSPA57446.2023.10087730.

[43] V. R. Burugadda, V. Dutt, Mamta, and N. Vyas, "Personalized cardiovascular disease risk prediction using random forest: An optimized approach," in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, Jul. 2023, pp. 226–232, doi: 10.1109/AIC57670.2023.10263915.

[44] T. Y. Rashme, L. Islam, S. Jahan, and A. A. Prova, "Early prediction of cardiovascular diseases using feature selection and machine learning techniques," in *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*, 2021, pp. 1554–1559, doi: 10.1109/ICCES51350.2021.9489057.

[45] M. I. Uddin, A. A. A. Ismom, Z. H. Anik, and M. N. Arefin, "Cardio vascular disease prediction using machine learning with improved feature selection," in *12th International Conference on Electrical and Computer Engineering, ICECE 2022*, 2022, pp. 336–339, doi: 10.1109/ICECE57408.2022.10088591.

[46] I. A. Marbaniang, N. A. Choudhury, and S. Moulik, "Cardiovascular disease (CVD) prediction using machine learning algorithms," in *2020 IEEE 17th India Council International Conference, INDICON 2020*, 2020, pp. 1–6, doi: 10.1109/INDICON49873.2020.9342297.

[47] P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning," *Intelligent Systems with Applications*, vol. 16, p. 200121, 2022, doi: 10.1016/j.iswa.2022.200121.

[48] J. Wang and Y. Wan, "Study on the causal relationship of cardiovascular disease influencing factors based on Bayesian causal network," in *ACM International Conference Proceeding Series*, 2022, pp. 127–131, doi: 10.1145/3523286.3524529.

[49] S. A. Sabab, M. A. R. Munshi, A. I. Pritom, and Shihabuzzaman, "Cardiovascular disease prognosis using effective classification and feature selection technique," in *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, 2016, pp. 1–6, doi: 10.1109/MEDITEC.2016.7835374.

[50] G. B. Prasath, V. Jagadeesh, J. Kavitha, and P. S. Kumar, "Cardiovascular disease prediction using extreme gradient boosting algorithm," in *2023 2nd International Conference on Advances in Computational Intelligence and Communication, ICACIC 2023*, 2023, pp. 1–6, doi: 10.1109/ICACIC59454.2023.10435219.

[51] H. Vinzey, A. Tidke, P. Palsodkar, S. Kottawar, Y. Dubey, and P. Fulzele, "Predictive analysis of cardiovascular diseases," in *2022 International Conference on Emerging Trends in Engineering and Medical Sciences, ICETEMS 2022*, 2022, pp. 461–465, doi: 10.1109/ICETEMS56252.2022.10093374.

[52] I. U. Haq, A. H. Rather, and G. Kaur, "A comparative analysis of machine learning algorithms for the early prediction of cardiovascular disease," in *Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023*, 2023, pp. 987–993, doi: 10.1109/ICECAA58104.2023.10212214.

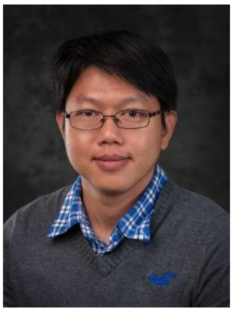
[53] Z. Ali, N. Naseer, and H. Nazeer, "Cardiovascular disease detection using multiple machine learning algorithms and their performance analysis," in *2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, Dec. 2022, pp. 1–7, doi: 10.1109/ETECTE55893.2022.10007319.

[54] R. Punugoti, V. Dutt, A. Kumar, and N. Bhati, "Boosting the accuracy of cardiovascular disease prediction through SMOTE," in *2023 International Conference on IoT, Communication and Automation Technology, ICICAT 2023*, 2023, pp. 1–6, doi: 10.1109/ICICAT57735.2023.10263703.

BIOGRAPHIES OF AUTHORS

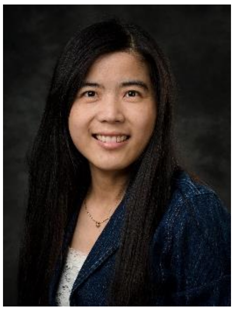
Ravipa Sonthana holds a Bachelor's degree in applied mathematics from King Mongkut's Institute of Technology Ladkrabang and is currently pursuing a Master's degree in mathematics with computer science at King Mongkut's University of Technology North Bangkok. She is currently working as a system and business analyst at Sino S-Tech Co., Ltd. Her research interests include cybersecurity, machine learning, data analytics, and data science. She can be contacted at email: ravipa.sgn@gmail.com.

Sakchai Tangprasert received his Ph.D. in information technology from King Mongkut's University of Technology North Bangkok. He currently serves as a lecturer in the Department of Mathematics at the Faculty of Applied Science, King Mongkut University of Technology North Bangkok. His research interests IT and cyber security, IT and project management, software development, and data science. He can be contacted at email: sakchai.t@sci.kmutnb.ac.th.



Yuenyong Nilsiam received his Ph.D. in computer engineering from Michigan Technological University. He currently serves as a lecturer in the Department of Electrical and Computer Engineering at the Faculty of Engineering, King Mongkut's University of Technology North Bangkok. His research interests focus on open-source technology and renewable energy, where he has made significant contributions to advancing accessible technological solutions for sustainable development. Dr. Nilsiam actively publishes in peer-reviewed journals and participates in international conferences in his fields of expertise. He can be contacted at email: yuenyong.n@eng.kmutnb.ac.th.

Nalinpat Bhumpenpein received her Doctoral degree in business informatics from University of Vienna, Austria. She is a lecturer in the Department of Information Technology at the Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok (KMUTNB). Her research interests encompass IT/digital strategic management, software development, and knowledge transfer. She can be contacted at email: nalinpat.b@itd.kmutnb.ac.th.



Siranee Nuchitprasitchai received both her Ph.D. and Master's degrees in computer engineering from Michigan Technological University, USA, and holds additional master's and bachelor's degrees from KMUTNB in information technology and applied mathematics, respectively. She is a lecturer in the Department of Information Technology at the Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok (KMUTNB). Her research interests encompass user experience design (UX), digital transformation in education, design thinking, human-computer interaction (HCI), technology-enhanced learning, IoT, image processing, and computer vision. Beyond her academic endeavors, she actively promotes digital literacy and the creative integration of technology among students, educators, and professionals. She serves as the Chair of the Bangkok ACM SIGCHI Chapter, contributing to the advancement of HCI and UXUI practices in Thailand. She can be contacted at email: siranee.n@itd.kmutnb.ac.th.