EVALUATING THE PERFORMANCE OF MACHINE LEARNING MODELS IN CARDIOVASCULAR RISK PREDICTION

ĐÁNH GIÁ HIỆU SUẤT CÁC MÔ HÌNH HỌC MÁY TRONG DỰ ĐOÁN NGUY CƠ TIM MẠCH

Duong Thi Hang*, Pham Duy Phong

Electric Power University

Ngày nhận bài: 28/5/2025; Ngày chấp nhận đăng: 25/7/2025

Abstract:

Cardiovascular diseases (CVD) are a leading cause of illness and death worldwide, making it crucial to predict cardiovascular risk accurately for effective prevention and treatment. This study aims to evaluate the performance of several supervised machine learning algorithms in predicting cardiovascular risk using a dataset of clinical and demographic features. Six commonly used models—Random Forest, XGBoost, Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree—are tested based on their ability to predict risk and other important metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). The data is preprocessed using normalization and transformation techniques, such as Quantile Transformation and Standard Scaling, to ensure the best model performance. The results provide a detailed comparison of the models' performance, showing their strengths and weaknesses in predicting cardiovascular risk. The findings highlight the best-performing models for identifying high-risk individuals, which could help healthcare professionals prioritize early interventions. The study also discusses the broader role of machine learning in healthcare, especially in disease prediction and prevention.

Keywords:

Cardiovascular Disease; Heart Disease Detection; Machine Learning; Model Comparison; XGBoost; Predictive Modeling.

Tóm tắt:

Các bệnh tim mạch (CVD) là một trong những nguyên nhân hàng đầu gây ra bệnh tật và tử vong trên toàn thế giới, do đó việc dự đoán chính xác nguy cơ mắc bệnh tim mạch là rất quan trọng cho công tác phòng ngừa và điều trị hiệu quả. Nghiên cứu này nhằm đánh giá hiệu quả của một số thuật toán học máy có giám sát trong việc dự đoán nguy cơ mắc bệnh tim mạch dựa trên bộ dữ liệu gồm các đặc trưng lâm sàng và nhân khẩu học. Sáu mô hình phổ biến được sử dụng — Random Forest, XGBoost, Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), và Decision Tree — được kiểm tra dựa trên khả năng dự đoán rủi ro và các chỉ số đánh giá quan trọng như độ chính xác, độ chính xác truy xuất (precision), độ bao phủ (recall), điểm F1 và đường cong ROC (AUC). Dữ liệu được xử lý trước bằng các kỹ thuật chuẩn hóa và biến đổi, chẳng hạn như chuyển đổi phân vị (Quantile Transformation) và chuẩn hóa chuẩn (Standard Scaling), nhằm đảm bảo hiệu suất tối ưu cho mô hình. Kết quả đưa ra so sánh chi tiết hiệu suất của các mô hình, qua đó thể hiện điểm mạnh và hạn chế của từng mô hình trong dự đoán nguy cơ mắc bệnh tim mạch. Phát hiện của nghiên cứu nhấn mạnh các mô hình có hiệu suất tốt nhất trong việc xác định các cá nhân có nguy cơ cao, từ đó hỗ trợ các chuyên gia y tế ưu tiên can thiệp sớm. Nghiên cứu cũng thảo luận về vai trò rộng lớn hơn của học máy trong y tế, đặc biệt là trong dự đoán và phòng ngừa bệnh tật.

Từ khóa: Bệnh tim mạch; Phát hiện bệnh tim; Học máy; So sánh mô hình; XGBoost; Dự đoán y học

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for approximately 18 million deaths annually and placing immense pressure on healthcare systems [1]. Accurate and early prediction is crucial for improving clinical decisions and reducing costs [2]. Machine learning (ML) has shown strong potential in areas such as indoor localization [3, 4], wireless communication [5], and especially in healthcare [6–9]. Supervised ML algorithms—Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Neighbors (KNN), and XGBoost—have proven effective in analyzing structured clinical data for heart disease prediction [6–8]. Despite the rise of deep learning and hybrid approaches [9], traditional ML models remain popular due to their interpretability and efficiency on moderate-sized datasets [10].

This study evaluates six widely-used supervised ML algorithms—RF, XGBoost, LR, SVM, KNN, and Decision Tree—on the cardio.csv dataset. The main contribution of this paper is to provide a fair comparison of these models under consistent conditions, helping healthcare professionals and data scientists choose suitable methods for early CVD risk prediction.

The paper is organized as follows: Section 2 reviews related work; Section 3 outlines the methodology; Section 4 presents results and discussion; Section 5 concludes the study and suggests future work.

2. RELATED WORK

Traditional machine learning (ML) models, including Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), and Decision Trees (DT), have been extensively employed for cardiovascular disease (CVD) prediction using structured clinical data [1, 2, 6]. In contrast, deep learning (DL) approaches such as Deep Neural Networks (DNNs) offer strong representational capacity, but their limited interpretability and substantial data requirements often hinder their applicability in real-world clinical environments [7, 8].

Feature selection plays a critical role in enhancing model performance. Techniques such as Pearson correlation and Extra Trees (ET) classifiers are widely adopted to identify relevant attributes and eliminate redundancy, thereby improving predictive accuracy [7, 8]. Moreover, hybrid strategies that combine ML with intelligent optimization algorithms—such as the integration of SVM with Quantum-behaved Particle Swarm Optimization (QPSO)—have demonstrated improved classification outcomes [9].

Nevertheless, many prior studies are constrained by limited dataset sizes and an overreliance on accuracy as the sole evaluation metric. In the context of imbalanced medical data, performance indicators such as precision, recall, F1-score, and confusion matrix offer a more

comprehensive and reliable assessment. Additionally, systematic comparisons of multiple ML algorithms under consistent experimental settings using publicly available datasets remain scarce. This gap motivates the present study, which aims to conduct a unified evaluation of various supervised models for early CVD risk prediction.

3. METHODOLOGY

3.1. Dataset description

This study uses the cardio.csv dataset from Kaggle, which contains 70,000 anonymized patient records (available at: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset). It includes 13 attributes covering demographic, anthropometric, clinical, and behavioral data, with a binary target variable indicating the presence (1) or absence (0) of cardiovascular disease.

Pearson correlation analysis revealed:

- BMI is strongly correlated with weight (r = 0.76), suggesting redundancy.
- Age, cholesterol, and weight show moderate correlations with the target (r = 0.24, 0.22, 0.18).
- Height and physical activity have weak correlations with the target (r = –0.01, –0.036).

As shown in Figure 1, these patterns help inform feature selection and model interpretation. Visualizations like pair plots and heatmaps further revealed patterns and redundancies in the data

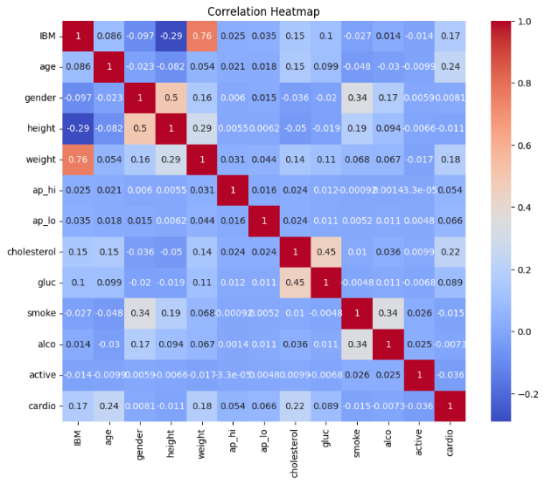

Figure 1. Pearson correlation heatmap

3.2. Data preprocessing

A preprocessing pipeline was applied to the cardio.csv dataset. The non-informative id column was removed, and the target variable cardio was separated from the features. To address skewed distributions, the Quantile Transformer was used to map non-Gaussian features to a uniform distribution, followed by standardization using z-score normalization via StandardScaler.

The dataset was split into training and testing subsets using a 70:30 ratio to simulate real-world deployment. No feature selection was performed, retaining all original features for a consistent evaluation across models.

3.3. Machine learning model training and evaluation process

To provide a comprehensive overview of the experimental workflow, Figure 2 illustrates the end-to-end modeling pipeline employed in this study.

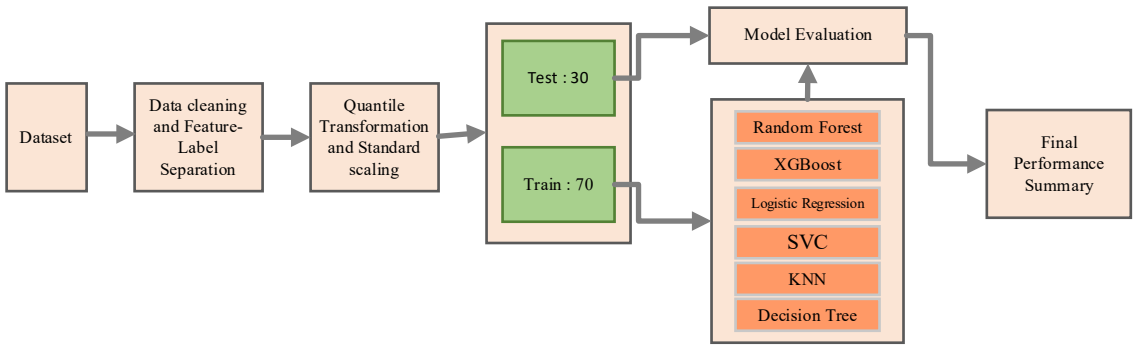

Figure 2. Workflow of Cardiovascular Disease Prediction Using Supervised Machine Learning Models

To evaluate the performance of the machine learning models, the following commonly used metrics were computed based on the confusion matrix:

The diagram encapsulates the sequence of processes, beginning with data cleaning and feature-label separation, followed by quantile transformation and standard scaling. The dataset is then partitioned into training and test sets (with a 70:30 split), after which six supervised machine learning models are trained and evaluated. The final stage consolidates the evaluation results to produce the overall performance summary. This structured pipeline ensures a fair and consistent comparison across all models.

4. RESULTS AND DISCUSSION

The performance of six traditional supervised learning models was evaluated on the cardiovascular disease dataset using a 70:30 train-test split. As shown in Figures 3 to 7, which compare Accuracy, Precision, Recall, F1-score, and AUC, both Support Vector Classifier (SVC) and XGBoost outperformed the other models across most evaluation metrics.

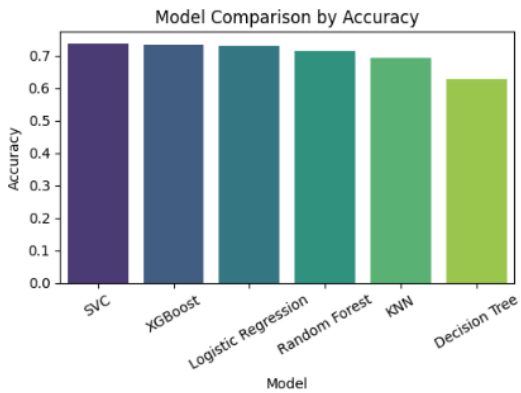

Figure 3. Accuracy Comparison of Models

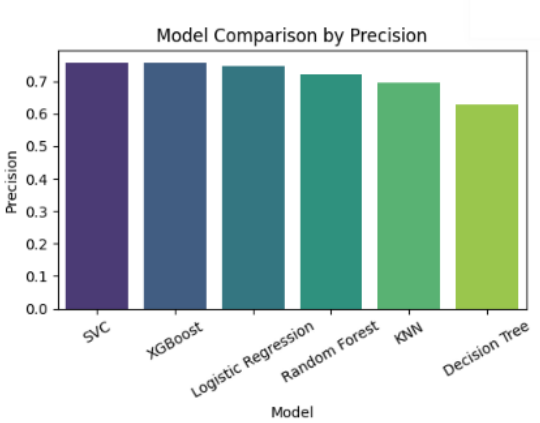

Figure 4. Precision Comparison of Models

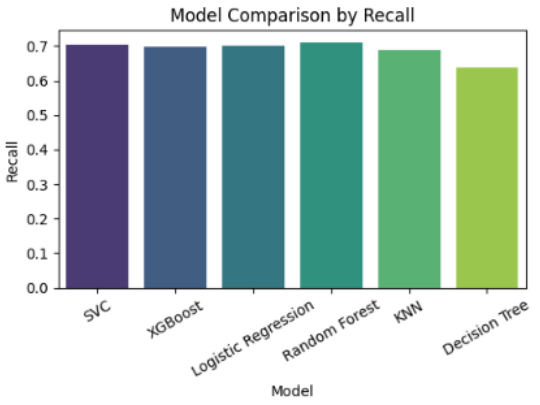

Figure 5. Recall Comparison of Models

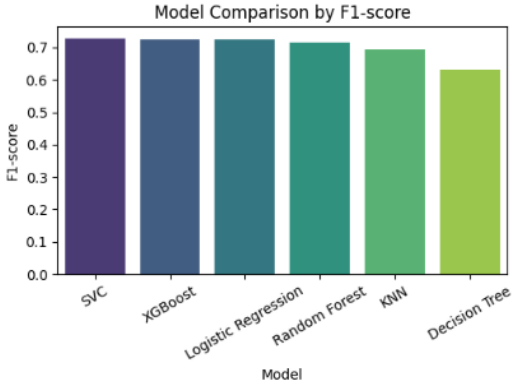

Figure 6. F1-score Comparison of Models

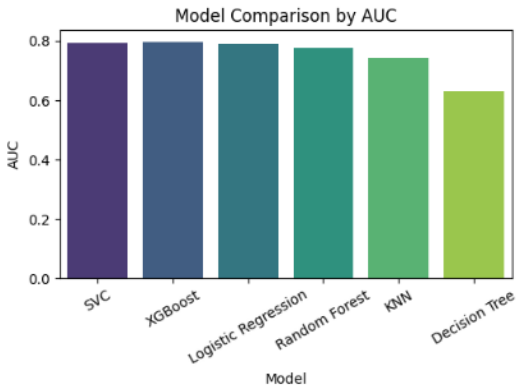

Figure 7. AUC Comparison of Models

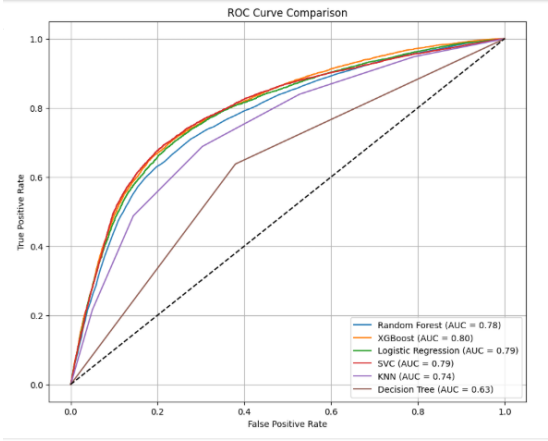

Figure 8. ROC curves of the Models

Table 1. Summary of model results

Model	Accuracy	Precision	Recall	F1-score	AUC
SVC	0.737048	0.755422	0.703957	0.728782	0.793966
XGBoost	0.735429	0.756037	0.698074	0.725900	0.797883
Random Forest	0.716429	0.720852	0.709840	0.715303	0.775151
KNN	0.692429	0.695552	0.688490	0.692003	0.742351
Logistic Regression	0.731095	0.747271	0.701395	0.723606	0.790308
Decision Tree	0.629048	0.628542	0.637727	0.633101	0.628963

As shown in the table 1, SVC achieved the best overall performance with the highest accuracy (73.7%), precision (75.5%), and F1-score (72.9%). XGBoost followed

closely with slightly lower accuracy but the highest AUC (0.798), highlighting its strong classification capability. Logistic Regression provided stable and competitive results, while Random Forest and KNN showed moderate performance. The Decision Tree model consistently ranked lowest across all metrics.

The performance of six traditional supervised learning models—Random Forest (RF), XGBoost, Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree (DT)—was comprehensively evaluated using a 70:30 train-test split on the cardiovascular disease dataset. The models were assessed based on several evaluation metrics: Accuracy, Precision, Recall, F1-score, and AUC.

SVC and XGBoost demonstrated superior performance across most metrics. Specifically, SVC achieved the highest accuracy (73.7%), precision (75.5%), and F1-score (72.9%), indicating a strong capability to balance both true positive classification and the reduction of false positives. XGBoost exhibited comparable results, particularly excelling in terms of AUC (0.7979), suggesting its strong discriminative ability between cardiovascular disease cases and non-cases.

Logistic Regression delivered stable results but slightly underperformed relative to SVC and XGBoost in most metrics. Nevertheless, it still maintained competitive scores, demonstrating its robustness in relatively simpler classification tasks.

In contrast, Random Forest and KNN displayed moderate performance. These models, while effective in certain contexts, lacked the discriminative power and precision of SVC and XGBoost. Finally, Decision Tree consistently showed the weakest performance across all metrics, with the lowest AUC (0.6289), indicating limited generalization capacity and predictive power for this particular dataset.

These findings underscore the importance of selecting advanced models such as SVC and XGBoost for complex medical prediction tasks, particularly when high precision and recall are essential for minimizing false negatives in healthcare applications. The performance disparities among the models highlight the trade-offs between accuracy, interpretability, and the ability to generalize across unseen data.

5. CONCLUSIONS

This study presented a comparative evaluation of six supervised machine learning models for cardiovascular disease prediction using a clinical dataset comprising 13 features. Among the evaluated algorithms, the Support Vector Classifier (SVC) outperformed all other models across key metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). XGBoost and Logistic Regression followed closely behind, demonstrating competitive performance, especially in terms of AUC and precision.

The results of this study underscore the potential of machine learning techniques, particularly SVC, in enhancing early detection and risk assessment of cardiovascular diseases, contributing to more efficient preventive healthcare practices. By effectively balancing accuracy and recall, SVC offers a promising model for identifying high-risk patients and minimizing false negatives in clinical settings.

Future research should focus on expanding the feature set to include additional clinical variables, genetic data, and lifestyle factors, which could further improve predictive performance. Additionally, exploring deep learning models and ensemble methods may lead to even greater improvements in model robustness and generalization. These advances could pave the way for the development of intelligent, data-driven decision support systems to aid healthcare professionals in diagnosing and managing cardiovascular conditions.

REFERENCES

[1] A. I. Alhassan et al., "Machine learning prediction in cardiovascular diseases: A meta-analysis," Scientific Reports, vol. 10, no. 1, pp. 1–11, 2020. doi: 10.1038/s41598-020-72685-1

[2] R. Li et al., "Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques," in Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021), pp. 294–300, 2021. doi: 10.5220/0011088300003170.

[3] Duong, T. H., Trinh, A. V., & Hoang, M. K. (2024). Efficient and Accurate Indoor Positioning System: A Hybrid Approach Integrating PCA, WKNN, and Linear Regression. J. Commun, 19, 37-43

[4] Hang, D. T., Manh, K. H., Vu, T. A., Quynh, T. P. T., & Viet, T. N. (2023). Dimensionality Reduction with Truncated Singular Value Decomposition and K-Nearest Neighbors Regression for Indoor Localization. International Journal of Advanced Computer Science and Applications, 14(10).

[5] Trang, P. T. Q., Hang, D. T., Son, H. X., Duong, D. T., & Vu, T. A. (2023, November). Millimeter Wave Path Loss Modeling for UAV Communications Using Deep Learning. In International Conference on Ad Hoc Networks (pp. 125-134). Cham: Springer Nature Switzerland.

[6] M. M. Deo, "Heart Disease Detection Using Machine Learning Models," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, pp. 1–7, 2020. doi: 10.14569/IJACSA.2020.0110977

[7] A. Madani et al., "Applications of Machine Learning in Cardiology," Nature Reviews Cardiology, vol. 19, pp. 1–17, 2022. doi: 10.1038/s41569-022-00698-1

[8] R. Dey et al., "Artificial Intelligence in Cardiology," Journal of the American College of Cardiology, vol. 77, no. 13, pp. 1565–1579, 2021. doi: 10.1016/j.jacc.2021.02.012

[9] M. R. U. Rehman et al., "A Robust Heart Disease Prediction System Using Hybrid Deep Neural

Networks," Computers in Biology and Medicine, vol. 136, p. 104672, 2021. doi: 10.1016/j.compbiomed.2021.104672

[10] R. Naik et al., "A Review of Machine Learning for Cardiology," Computer Methods and Programs in Biomedicine, vol. 208, p. 106221, 2021. doi: 10.1016/j.cmpb.2021.106221.

[11] M. D. Kumar et al., "Early Heart Disease Prediction Using Feature Engineering and Machine Learning Algorithms," Materials Today: Proceedings, vol. 45, pp. 5179–5184, 2021. doi: 10.1016/j.matpr.2021.01.116.

[12] A. K. Singh and S. Srivastava, "New Cardiovascular Disease Prediction Approach Using Support Vector Machine and Quantum-Behaved Particle Swarm Optimization," Multimedia Tools and Applications, vol. 82, pp. 3599–3623, 2023. doi: 10.1007/s11042-023-16194-z.

Authors Biography:

Duong Thi Hang, was born in Bac Giang, Vietnam. She received her Bachelor's degree in 2000, her Master's degree in 2006, and completed her Ph.D. in 2025, all from the University of Engineering and Technology, Vietnam National University, Hanoi. She is currently a lecturer at Electric Power University (EPU), Hanoi, Vietnam. Her main research interests include indoor localization using machine learning, optimization algorithms, and UAV-based communication systems. Her work focuses on developing efficient, interpretable, and scalable solutions for real-time positioning challenges, with an emphasis on combining machine learning and wireless technologies to enable smarter environments and autonomous systems.

Duy Phong Pham is the Dean of the Faculty of Electronics and Telecommunications at the Electric Power University, Hanoi, Vietnam. He received the B.E degree in Telecommunications Engineering from University of Communications and Transport, Hanoi, in 2000 and the Master degree from Hanoi University of Science and Technology, Hanoi, Vietnam in 2007. He received the Ph.D degree in theTelecommunications Engineering at Vietnam Research Institute of Electronics, Informatics and Automation, Hanoi, Vietnam in 2013. He was a researcher in Research Institute of Posts and Telecommunications (2000-2005). His main research interests include wireless communications, antenna design for wireless communications, underwater acoustic communications, electromagnetic interference on telecommunication systems due to power systems, earthing and lightning protection.