

Data-driven diagnosis of cardiovascular disease in patients with comorbid hypertension, dyslipidemia or diabetes using classification techniques

Supichaya Chumpong¹, Piyarat Promsuwan², Kittisak Chumpong^{2,3,4}

¹ Department of Medicine, Pak Phanang Hospital, Pak Phanang, Nakhon Si Thammarat, Thailand

² Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand

³ Research Center in Mathematics and Statistics with Applications, Prince of Songkla University, Songkhla, Thailand

⁴ Financial Mathematics, Data Science and Computational Innovations Research Unit (FDC), Department of Mathematics, Faculty of Science, Kasetsart University, Chatuchak, Bangkok, Thailand.

Corresponding Author: Kittisak Chumpong **Email:** kittisak.ch@psu.ac.th

Received: 20 June 2025 **Revised:** 29 September 2025 **Accepted:** 7 October 2025 **Available online:** May 2026

DOI: 10.55131/jphd/2026/240216

ABSTRACT

Cardiovascular disease (CVD) is a leading cause of death globally, especially among individuals with chronic comorbidities such as hypertension, diabetes, and dyslipidemia. Early identification of high-risk individuals is essential for timely intervention, yet conventional risk models often underperform in complex clinical populations. Using real-world clinical data, this study aimed to develop and evaluate machine learning (ML) models for predicting CVD with a focus on optimizing sensitivity and interpretability. We used hospital-based clinical data from patients with at least one of the following comorbidities: hypertension, diabetes or dyslipidemia. The applied ML algorithms included logistic regression, naive Bayes, k-nearest neighbors, decision tree (DT), random forest, support vector machine (SVM), and artificial neural networks. Repeated 10x10-fold cross-validation and downsampling were used to address class imbalance, and systematic hyperparameter tuning was performed. Model performance was evaluated using accuracy, precision, sensitivity, F1 score, and area under the curve (AUC). SHapley Additive Explanations (SHAP) was applied to interpret the final model and identify key features. The SVM model achieved the highest sensitivity, indicating suitability for use in CVD screening, while the DT also showed relatively strong sensitivity. Final SVM performances included a sensitivity of 89.29%, a precision of 57.94%, an F1-score of 69.97%, and an AUC of 75.65%. SHAP analysis revealed creatinine, systolic blood pressure, low-density lipoprotein, fasting blood sugar, and triglycerides as the top predictors of CVD, aligning with established cardiovascular risk factors. We present an interpretable ML model with high sensitivity for CVD prediction in patients with metabolic comorbidities. By integrating SHAP with robust validation techniques, the model supports personalized risk assessment and is suitable for screening applications aimed at early detection and preventive care in clinical settings.

Keywords:

cardiovascular diseases, machine learning, sensitivity, comorbidity, SHAP analysis

Citation:

Supichaya Chumpong, Piyarat Promsuwan, Kittisak Chumpong. Data-driven diagnosis of cardiovascular disease in patients with comorbid hypertension, dyslipidemia or diabetes using classification techniques. J Public Hlth Dev. 2026;24(2):234-249 (<https://doi.org/10.55131/jphd/2026/240216>)

INTRODUCTION

Cardiovascular disease (CVD) continues to be a major global health burden and one of the leading causes of death worldwide. It is closely linked to vascular dysfunction, metabolic abnormalities, and long-term lifestyle-related exposures, all of which contribute to significant morbidity and premature mortality. Early identification of individuals at elevated risk is essential for timely intervention and for reducing adverse cardiovascular outcomes.¹

Several clinical and biochemical markers, including creatinine, systolic blood pressure (SBP), low-density lipoprotein (LDL), triglycerides (Tg), and fasting blood sugar (FBS), have been consistently associated with increased cardiovascular risk across diverse populations.²⁻⁶ These parameters are routinely collected in clinical settings and serve as a practical foundation for predictive modeling. The presence of multiple comorbid conditions further amplifies this risk, and patients with overlapping metabolic disorders represent a high-priority population for targeted risk stratification.⁷⁻¹⁰

Machine learning (ML) has become an increasingly valuable approach in health prediction, due to its ability to detect complex, nonlinear patterns in large clinical datasets. ML has been applied in various chronic disease contexts, including cardiovascular risk estimation and diabetes

management.¹¹⁻¹⁹ Despite encouraging results, many existing ML-based models have been developed using general population samples without comorbid conditions and offer limited interpretability, which hinders their integration into clinical workflows.

This study seeks to address these limitations by developing and evaluating ML models for CVD prediction in patients with hypertension, diabetes, and dyslipidemia, using real-world hospital data. In this clinical context, where missing a high-risk case may delay preventive care, optimizing sensitivity is essential to ensure the model is suited for screening applications, especially in routine or resource-limited settings. Shapley Additive Explanations (SHAP) was applied to improve interpretability and enhance clinical transparency. By combining robust model validation with explainable artificial intelligence, this study aims to support personalized cardiovascular risk assessment and enable earlier detection in routine care.

METHODS

Overview of the proposed methodology

This study focuses on predicting the likelihood of CVD using ML techniques, offering valuable insights for medical professionals and patients. The methodology is broken down into several key steps, from data preparation through to final evaluation, as illustrated in Figure 1.

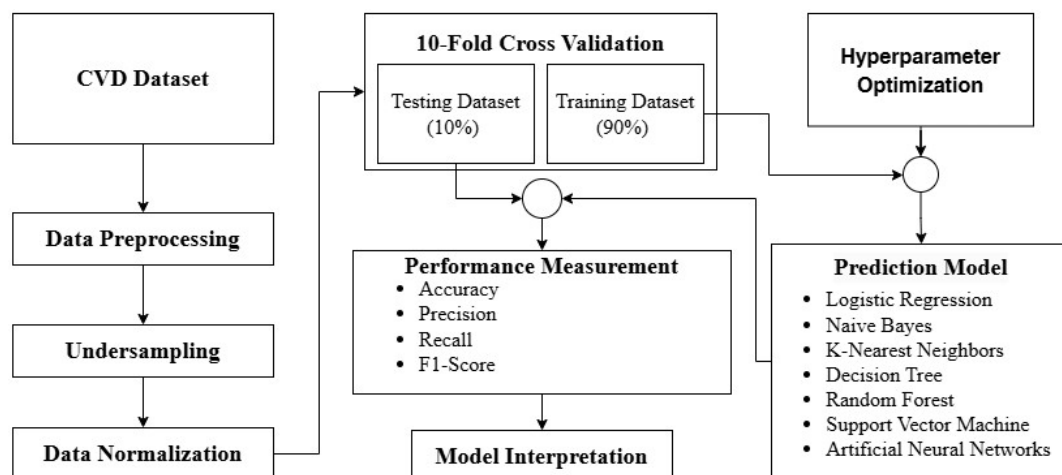


Figure 1. Flowchart of the proposed methodology.

The process began with the collection of a dataset for classification. During data preprocessing, the dataset underwent cleaning, transformation, outlier removal, and normalization. Following these operations, 10-fold cross-validation was used to split the data into training and test sets. This process was repeated 10 times to provide an unbiased estimate of the performance of each model. ML models were trained on the training set. Then, hyperparameters were tuned to improve the performance. Finally, the performance metrics were evaluated on the test set and the results of each model were compared. Model interpretation was performed using SHAP to identify the most influential clinical features contributing to predictions.

Dataset

The CVD dataset was collected from Pak Phanang Hospital in Nakhon Si Thammarat Province, Thailand, between January 1 and December 31, 2023. A total of 6,739 records with 15 variables were collected. The dataset included CVD cases with at least one of the following comorbidities: hypertension, diabetes, or dyslipidemia. The study was approved by the PSU Human Research Ethics Committee, Prince of Songkla University

(PSU-HREC-2024-069-1-1; approved on 22 November 2024), meeting the criteria for an Exempt Research Determination.

Data preprocessing

The dataset underwent several preprocessing steps to improve data quality, ensure consistency, and prepare the variables for ML analysis. The initial dataset comprised 6,739 patient records. To focus on the adult population at risk of CVD, individuals aged under 35 years were excluded, leaving 6,699 records. Duplicate entries from the same individuals were then removed to ensure data integrity, leaving 6,683 unique cases. Next, records with missing values in key clinical variables were excluded, reducing the dataset to 5,792 complete cases. Lastly, records containing zero or other physiologically inconsistent values were removed, yielding a final dataset of 5,652 records.

Categorical variables, including gender, hypertension, dyslipidemia, diabetes, and CVD status, were then converted into numerical format to facilitate the development of the model. The set of predictor variables and the binary outcome definitions are summarized in Table 1.

Table 1. Dataset specifications

No.	Feature	Description	Variable type
1	Age	Patient age in years	Numerical
2	Gender	Gender of the patient (Male=0/Female=1)	Categorical
3	Weight	Patient weight in kilograms	Numerical
4	Height	Patient height in centimeters	Numerical
5	SBP	Systolic Blood Pressure (mmHg)	Numerical
6	DBP	Diastolic Blood Pressure (mmHg)	Numerical
7	FBS	Fasting Blood Sugar (mg/dL)	Numerical
8	Creatinine	Creatinine (mg/dL)	Numerical
9	eGFR	Estimated Glomerular Filtration Rate (mL/min/1.73 m ²)	Numerical
10	LDL	Low-Density Lipoprotein Cholesterol (mg/dL)	Numerical
11	Tg	Triglycerides (mg/dL)	Numerical
12	Diabetes status	Indicates if the patient has diabetes (No=0/Yes=1)	Categorical
13	Dyslipidemia status	Indicates if the patient has dyslipidemia (No=0/Yes=1)	Categorical
14	Hypertension status	Indicates if the patient has hypertension (No=0/Yes=1)	Categorical
15	CVD status	Indicates if the patient has cardiovascular disease (No=0/Yes=1)	Categorical

Data balancing, normalization and cross-validation

To address the class imbalance between CVD and non-CVD cases, random undersampling was applied to equalize the class distribution. In each round, the majority class was randomly downsampled to match the number of minority class instances. This process was repeated 10 times using different subsets to enhance robustness and reduce sampling bias.

Each balanced dataset was normalized using robust scaling²⁰ and evaluated using 10-fold cross-validation²¹. Model performances were averaged across all folds and iterations to compare classifiers. For further interpretation and statistical evaluation, the fold that had a sensitivity value closest to the overall mean was selected in advance from the repeated cross-validation. This operation ensured that downstream analyses were based on representative and unbiased validation results.

Prediction models and Performance measurement

To predict CVD in patients with metabolic comorbidities, seven supervised machine learning algorithms were applied: logistic regression (LR), naive Bayes (NB), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), support vector machine (SVM), and artificial neural networks (ANN). These models were selected for their established use in clinical prediction and optimized via systematic hyperparameter tuning²²⁻²⁹.

Model performance was evaluated using standard classification metrics: accuracy, precision, sensitivity, F1-score, and area under the receiver operating characteristic curve (AUC). Sensitivity was prioritized to minimize false negatives in screening contexts. The F1-score provided a balanced view of sensitivity and precision, while AUC assessed discriminative ability across all thresholds³⁰⁻³³.

Model interpretability

To improve model interpretability, SHAP (Shapley Additive Explanations) was used to show how each feature influenced the predictions. Important features were identified at the overall level by averaging their contributions across patients (global explanation), and individual predictions were examined to understand how specific features affected the decision of the model (local explanation)³⁴. Analysis was performed on one representative validation fold.

Statistical confidence estimation

To estimate uncertainty, 95% confidence intervals were computed using bootstrap resampling with 1,000 iterations. Sampling was performed with replacement from the predicted outcomes in a fixed test set, obtained from the selected validation fold. Performance metrics were recalculated in each iteration to generate confidence intervals³⁵.

RESULTS

This section presents the results of data preparation and predictive modeling, beginning with patient characteristics, followed by model evaluation, optimization outcomes, and interpretability analysis.

Patient characteristics and model preparation

Following the removal of missing and duplicate records, a total of 6,739 patient records remained. After excluding incomplete cases, the final dataset consisted of 5,652 records. Group differences between patients with and without CVD were assessed using Welch's t-tests for numerical variables and chi-square tests for categorical variables. As shown in Table 2, significant differences ($p < 0.05$) were observed in age, height, SBP, FBS, creatinine, eGFR, LDL, and gender distribution. However, all features were included in the models regardless of their univariate significance.

Table 2. Characteristics of patients with and without CVD

Feature	Patients without CVD (n=5,373)	Patients with CVD (n=279)	P-value
Age	65.87 ± 12.34	69.90 ± 11.73	<0.0001
Gender			<0.0001
Male	1,947 (36.24%)	170 (60.93%)	
Female	3,426 (63.76%)	109 (39.07%)	
Weight	64.10 ± 17.79	63.94 ± 12.78	0.8432
Height	159.10 ± 8.81	161.13 ± 9.82	0.0008
SBP	129.64 ± 14.01	127.16 ± 14.84	0.0069
DBP	76.75 ± 10.59	75.56 ± 10.89	0.0736
FBS	126.47 ± 42.31	120.58 ± 31.75	0.0032
Creatinine	0.89 ± 0.37	1.08 ± 0.52	<0.0001
eGFR	81.10 ± 21.12	71.13 ± 22.09	<0.0001
LDL	116.23 ± 37.75	103.36 ± 34.22	<0.0001
Tg	144.20 ± 83.00	146.80 ± 66.27	0.5290
Diabetes status			0.6125
No	2,979 (55.44%)	159 (56.99%)	
Yes	2,394 (44.56%)	120 (43.01%)	
Dyslipidemia status			0.2852
No	1,610 (29.96%)	92 (32.97%)	
Yes	3,763 (70.04%)	187 (67.03%)	

Feature	Patients without CVD (n=5,373)	Patients with CVD (n=279)	P-value
Hypertension status			0.2264
No	1,348 (25.09%)	79 (28.32%)	
Yes	4,025 (74.91%)	200 (71.68%)	

(SBP) systolic blood pressure; (DBP) diastolic blood pressure; (FBS) fasting blood sugar; (eGFR) estimated glomerular filtration rate; (LDL) low-density lipoprotein cholesterol; (Tg) triglycerides.

To further refine the dataset, downsampling was performed 10 times by randomly undersampling the majority class to match the number of minority class instances. Each downsampled dataset was scaled using robust scaling to reduce the influence of outliers. For each iteration, 10-fold cross-validation was conducted, and hyperparameter tuning was applied with the objective of maximizing sensitivity, which is critical in minimizing errors in clinical diagnosis. The best parameter set was chosen from the iteration with a sensitivity closest to the mean across all tuning rounds, ensuring representative and stable performances. The selected hyperparameters for each model are summarized in Table 3.

Table 3. Best hyperparameters for each prediction model after tuning

Model	Hyperparameter	Parameter spaces	Best hyperparameters
LR	C	0.001, 0.01, 0.1, 1, 10, 100	0.1
	penalty	l1, l2, elasticnet, none	l2
	solver	liblinear, saga, lbfgs, newton-cg	newton-cg
	class_weight	None, balanced	balanced
	max_iter	100, 500, 1000	500
NB	model	GaussianNB, BernoulliNB	BernoulliNB
	var_smoothing	$10^{-9} - 1$	-
	alpha	0.1, 0.5, 1.0	0.5
	fit_prior	True, False	True
	binarize	0.0, 0.5, 1.0	1.0
KNN	class_prior	None	None
	n_neighbors	3, 5, 7, 9, 11	11
	weights	uniform, distance	uniform
	algorithm	auto, ball_tree, kd_tree, brute	auto
	metric	euclidean, manhattan, minkowski	euclidean
DT	p	1,2	1
	criterion	gini, entropy, log_loss	log_loss
	splitter	best, random	best
	max_depth	None, 5, 10, 20	10
	min_samples_split	2, 5, 10	5
	min_samples_leaf	1, 2, 4	4
	max_features	None, sqrt, log2	sqrt
	class_weight	None, balanced	None
	max_leaf_nodes	None, 10, 20, 50	10
	min_impurity_decrease	0.0, 0.01, 0.1	0.01

Model	Hyperparameter	Parameter spaces	Best hyperparameters
RF	n_estimators	50, 100, 200	100
	max_depth	None, 10, 20, 30	30
	min_samples_split	2, 5, 10	10
	min_samples_leaf	1, 2, 4	2
	max_features	sqrt, log2, None	sqrt
	bootstrap	True, False	True
	class_weight	None, balanced	balanced
SVM	C	0.1, 1, 10, 100	1
	kernel	linear, poly, rbf, sigmoid	poly
	degree	2, 3, 4	4
	gamma	scale, auto	scale
ANN	class_weight	None, balanced	None
	hidden_layer_sizes	(50,), (100,), (50,50), (100,50), (50,100), (100,100)	(50,)
	activation	tanh, relu	relu
	solver	sgd, adam	gd
	alpha	0.0001, 0.001, 0.01	0.001
	learning_rate	constant, adaptive	constant
	max_iter	200, 400, 800	500
	learning_rate_init	0.001, 0.01, 0.1	0.1
	early_stopping	True, False	True
	batch_size	auto, 64, 128	Auto
	momentum	0.9, 0.95	0.95

(LR) logistic regression; (NB) naive Bayes; (KNN) k-nearest neighbors; (DT) decision tree; (RF) random forest; (SVM) support vector machine; (ANN) artificial neural networks.

Performance comparison of classifiers

As shown in Figure 2, and based on a fixed probability threshold of 0.5, the SVM model achieved the highest sensitivity (88.28%), indicating its superior ability to identify true CVD cases. Although its accuracy (53.84%) and precision (52.57%) were lower than those of other models, its strong sensitivity makes it particularly suitable for clinical applications, where minimizing FNs is critical. The DT model had the highest F1-score (67.79%) and strong sensitivity (80.04%), offering a more balanced trade-off between precision and sensitivity. The

RF model also demonstrated a solid overall performance, with an F1-score of 66.46% and an accuracy of 65.11%, and an AUC of 0.6927, but a lower sensitivity (69.18%) compared to SVM and DT. Among the remaining models, NB achieved a relatively high sensitivity (69.75%) but had low precision, while LR showed moderate sensitivity (66.99%) but the highest AUC (0.7077), reflecting robust overall discrimination. ANN yielded balanced results with moderate sensitivity (66.52%) and an AUC of 0.6828, while KNN performed similarly but with lower sensitivity.

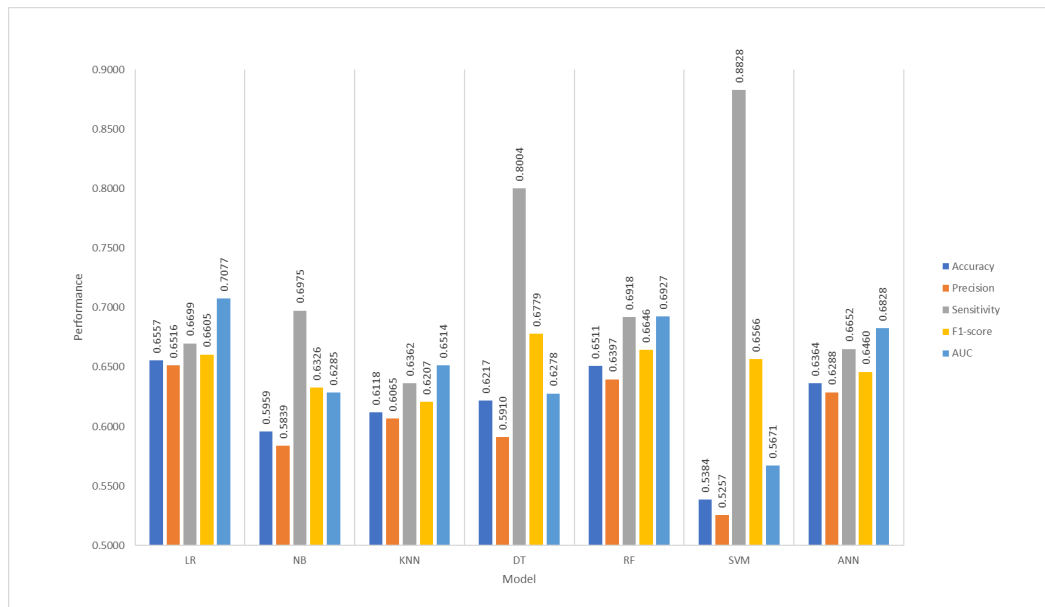


Figure 2. Performance comparison of the optimized studied models. (LR) logistic regression; (NB) naive Bayes; (KNN) k-nearest neighbors; (DT) decision tree; (RF) random forest; SVM, support vector machine; (ANN) artificial neural networks.

Given the clinical imperative to minimize FNs, sensitivity was prioritized as the key criterion for model selection. Despite its lower accuracy and precision, the superior sensitivity of the SVM model justified its selection for further evaluation. Subsequent analyses explored how varying the probability threshold impacted sensitivity and precision, to further evaluate the utility of SVM in screening contexts.

Threshold selection and diagnostic behavior of the SVM model

To further explore the classification behavior of the SVM model, we examined

the relationship between sensitivity and precision across probability thresholds ranging from 0.1 to 0.9. As shown in Figure 3, the model maintained extremely high sensitivity at lower thresholds, exceeding 99% between 0.1 and 0.4, while precision remained relatively stable at approximately 50%. At the default threshold of 0.5, a more favorable trade-off was observed, where sensitivity decreased to 88.28% and precision improved to 52.57%. This balance reflects a practical operating point, particularly in screening settings, where minimizing FNs is critical.

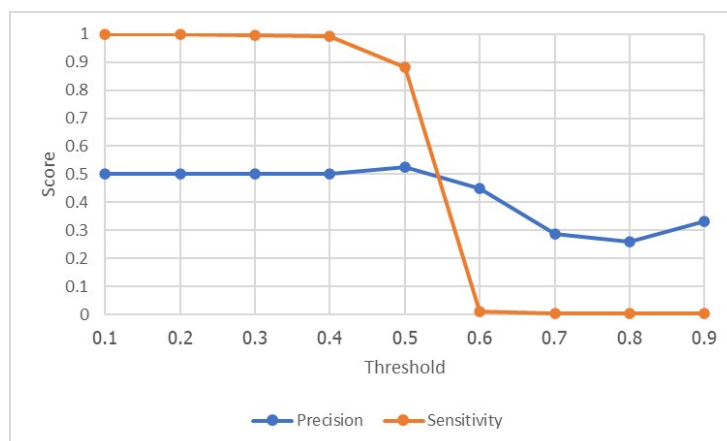


Figure 3. Sensitivity and precision of the SVM model across probability thresholds.

Beyond the threshold of 0.5, sensitivity declined sharply, reaching 1.15% at 0.6 and 0.22% at 0.9. Precision showed only minor variation—between 26% and 45%. This pattern supports the use of the default threshold of 0.5 as the most appropriate balance between detecting true CVD cases and minimizing FNs.

Interpretability of the final SVM model

SHAP was applied to the final SVM model from the fold with sensitivity closest to the mean. The top 10 features

contributing to the model predictions were ranked by their mean absolute SHAP values (Figure 4). Creatinine was the most influential feature (mean SHAP = 0.0166), followed by systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL), and fasting blood sugar (FBS). These findings are consistent with established cardiovascular risk factors. In contrast, variables such as weight, height, and diastolic blood pressure (DBP) had relatively limited influence.

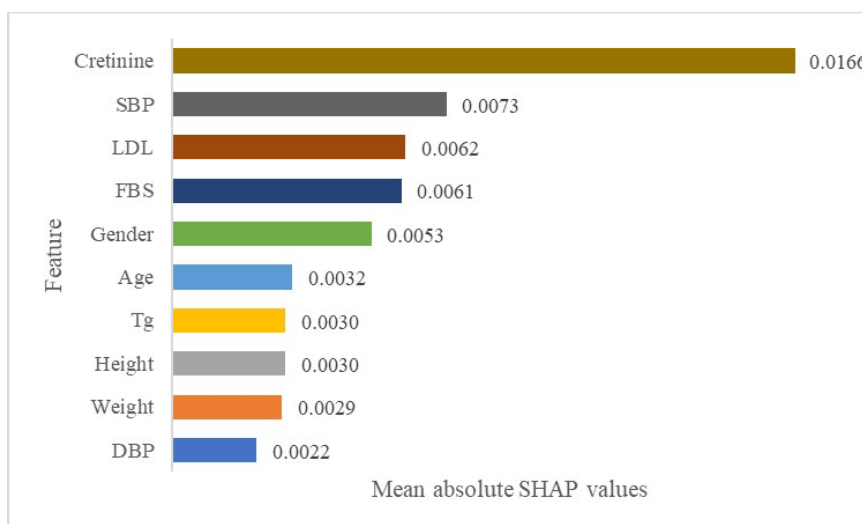


Figure 4. Top 10 features ranked by mean absolute SHAP values. (SBP) systolic blood pressure; (DBP) diastolic blood pressure; (FBS) fasting blood sugar; (LDL) low-density lipoprotein cholesterol; (Tg) triglycerides; (SHAP) Shapley additive explanations.

At the patient level, a SHAP waterfall plot for an individual classified with high confidence (predicted probability

= 1.00) demonstrated that elevated creatinine was the primary contributor (+0.47), followed by FBS and triglycerides

(Tg) (Figure 5). Other variables, including SBP, dyslipidemia, and eGFR, had minimal or negative impacts. This prediction aligns with known clinical reasoning. Elevated creatinine indicates renal dysfunction, a known cardiovascular risk, while high FBS reflects poor glycemic control. Together,

they increase the metabolic burden, and their identification as predictors supports the confidence of the classification produced by the model, which is consistent with risk profiles in diabetic nephropathy or cardiorenal syndrome.

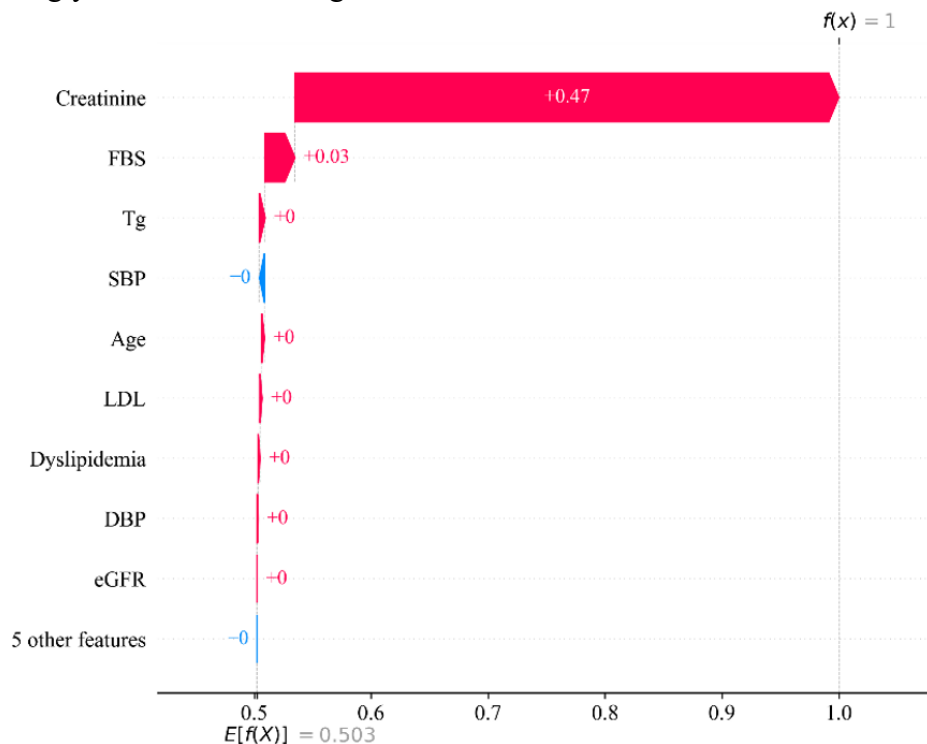


Figure 5. SHAP waterfall plot showing feature contributions for a patient confidently predicted to have CVD ($f(x) = 1.00$). (SBP) systolic blood pressure; (DBP) diastolic blood pressure; (FBS) fasting blood sugar; (eGFR) estimated glomerular filtration rate; (LDL) low-density lipoprotein cholesterol; (Tg) triglycerides; (SHAP) Shapley additive explanations.

Bootstrap confidence interval of the final SVM model

We computed 95% bootstrap confidence intervals on the same validation fold previously selected for SHAP analysis, using 1,000 iterations with sampling from the test set. The model achieved a mean sensitivity of 89.29% (95% CI: 76.00%–100.00%), a precision of 57.94% (95% CI: 43.18%–72.10%), and an F1-score of 69.97% (95% CI: 57.14%–81.01%), indicating a reasonable trade-off between sensitivity and specificity. The average accuracy was 61.64% (95% CI: 49.09%–74.55%), and the AUC was 75.65% (95%

CI: 61.76%–87.67%), reflecting good overall discriminative ability.

DISCUSSION

Principal findings

This study evaluated multiple ML models for predicting CVD using routinely collected clinical data from patients with comorbidities. Among the models, SVM achieved the highest sensitivity (88.28%) and a strong AUC (75.65%), making it particularly suited for early-stage detection based on its success in identifying true CVD cases. The model configuration was

optimized through repeated cross-validation, which supported its stable performance across variations in the data. Although the overall accuracy and precision of the model were moderate, its ability to consistently capture true positives highlights its potential clinical value in initial risk assessment. SHAP analysis confirmed that creatinine, SBP, LDL, and FBS were the most influential predictors of CVD, aligning with established cardiovascular risk factors. These findings support the integration of SVM into clinical screening workflows to enhance early identification and referral of high-risk individuals.

Model validation and clinical alignment with prior evidence

Several previous studies have applied ML to predict CVD using structured clinical or public datasets. Models such as DT, SVM, and RF were commonly used in prior studies^{12–15,17–19}. While many reported strong accuracy or sensitivity, few employed rigorous validation or optimization. Only a small number used cross-fold validation, and none performed repeated cross-validation or systematic hyperparameter tuning.

In contrast, our study applied repeated 10x10-fold cross-validation and comprehensive tuning, which improved performance stability and generalizability. Moreover, unlike prior works that focused on general populations, our study specifically targeted patients with hypertension, diabetes, and dyslipidemia. These comorbid conditions are known to elevate cardiovascular risk and are commonly encountered in real-world clinical settings. Importantly, while previous studies focused primarily on predictive performances, few incorporated explainability frameworks such as SHAP, and even fewer applied them in a clinically interpretable way. By integrating SHAP analysis, we enhanced model transparency and provided clinically meaningful

interpretations at both the population and individual levels.

SHAP analysis identified creatinine, SBP, LDL, FBS, and Tg as the top five contributors. Each of these variables has been previously established as a cardiovascular risk factor across diverse clinical populations. Elevated creatinine during hospitalization predicts post-discharge mortality in heart failure patients, independent of baseline renal function². SBP shows a log-linear relationship with cardiovascular risk beginning at levels as low as 120 mmHg³. LDL shows a U-shaped association with cardiovascular outcomes and mortality, with increased risk at both low and high levels⁴. Tg demonstrates a nonlinear relationship, with risk increasing up to a threshold and then leveling off⁵. FBS levels above 100 mg/dL are associated with higher risks of ischemic heart disease and stroke, even in non-diabetic individuals⁶. These findings reinforce the clinical validity of our SHAP-based feature importance. Beyond confirming known risk factors, SHAP provides case-specific explanations that enhance interpretability and support individualized decision-making, bridging data-driven modeling and clinical practice.

Clinical implications

The high sensitivity of the SVM model suggests strong potential for use in screening, where the priority is to avoid missing patients with true CVD. This is particularly relevant in primary care, where early identification and referral of high-risk individuals can improve clinical outcomes. The model supports decision-making based only on variables, such as blood pressure, creatinine levels, LDL and FBS, routinely collected during initial assessments without the need for advanced tests.

Sensitivity was prioritized during model selection to reflect the clinical importance of minimizing missed diagnoses. Although accuracy, precision, and F1-score were also considered,

sensitivity was prioritized because false negatives carry greater clinical risk than false positives. This rationale justified selecting SVM as the final model despite its lower performance in other metrics. Once the final model was chosen, AUC was reported, providing a broader view of the ability of the final model to distinguish between CVD and non-CVD cases across thresholds. This assessment more accurately identified the potential of the final model for integration into flexible decision-support systems. Given its simplicity and reliance on widely available data, the model could also be deployed as a web-based tool to assist screening and referral decisions, particularly in community hospitals or primary care units. Such tools may also help support early triage, streamline referral pathways, and strengthen preventive care strategies.

Explainability and transparency

In clinical settings, predictive accuracy alone is not enough to justify the adoption of ML models. Clinicians also need to understand how each prediction is made, especially in high-stakes decisions such as CVD screening. In this study, SHAP was applied to help interpret the predictions of the SVM model at both the individual patient and overall system levels.

The ability to visualize how the model makes its decisions can help build trust and promote more responsible use of AI in healthcare. In real-world practice, particularly in primary care, this type of explanation is especially useful in borderline or unexpected cases, where clinicians may need to review and verify model-generated suggestions with confidence. In addition to ranking overall feature importance, SHAP waterfall plots allow clinicians to understand how each variable has contributed to a specific prediction. This individualized explanation

supports real-time clinical decision-making, enabling providers to cross-check AI outputs with patient-specific findings, and improving trust in model recommendations.

Model deployment potential

The model developed in this study is well suited for integration into clinical workflows, as it relies solely on routinely collected variables and requires no specialized infrastructure. Its simplicity supports deployment in primary care settings or community hospitals, including integration with electronic health record systems or decision-support tools. Such implementation can facilitate routine screening, streamline referrals, reduce delays, and enhance the efficiency of preventive care services.

Strengths and limitations

This study has several strengths. We used real-world clinical data and evaluated multiple ML models using a rigorous repeated ten-fold cross-validation framework. Sensitivity was prioritized to align with clinical screening needs, and SHAP analysis was applied to improve model interpretability at both global and individual levels. The model was developed using routinely available clinical variables, making it suitable for use in resource-limited settings and feasible for integration into existing workflows.

However, the study also has limitations. The data were obtained from a single center, which may limit the generalizability of the findings. External validation was not performed, and the retrospective nature of the dataset may introduce biases related to data collection and documentation. Additionally, the model did not incorporate social or behavioral variables, which may also influence cardiovascular risk.

RECOMMENDATIONS

This study demonstrates that an SVM model trained on routine clinical data can effectively predict cardiovascular disease in high-risk patients with high sensitivity and clinical interpretability. Key predictors such as creatinine, SBP, LDL, FBS, and Tg align with known risk factors, supporting the model's integration into primary care for early detection and improved patient outcomes.

Future work should focus on external validation across different populations to assess generalizability, and prospective studies in clinical settings to evaluate real-world performances and impacts on clinical decision-making. Additionally, user-centered testing is needed to ensure the smooth integration of the model into existing clinical workflows and support safe, effective deployment in routine care. Future iterations could incorporate social or behavioral data to enhance predictive power, while broader implementation studies may help define the role and success of the proposed model in improving care pathways and supporting population-level cardiovascular screening.

AUTHOR CONTRIBUTION

SC: Conceptualization, Formal analysis, Investigation, Data curation, Writing – original draft and Writing – review & editing. KC: Methodology, Formal analysis, Investigation, Data curation, Writing – original draft and Writing – review & editing. PP: Investigation, Data curation, Writing – original draft and Writing – review & editing.

ETHICAL APPROVAL

This study was approved by the PSU Human Research Ethics Committee, Prince of Songkla University (Protocol No.

PSU-HREC-2024-069-1-1; approved on 22 November 2024).

ACKNOWLEDGEMENTS

We gratefully acknowledge the Medical Records Department at Pak Phanang Hospital for providing the cardiovascular disease dataset. We also extend our gratitude to Thomas Duncan Coyne for his assistance in improving the English language of this manuscript. All remaining errors are the responsibility of the authors. Our heartfelt thanks go to Oab and Chul Chumpong for their unwavering encouragement, which has brought happiness and inspiration to the first and third authors, enriching both this research and our family life.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. [cited 2024 Nov 27]. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
2. Shlipak MG, Chertow GC, Massie BM. Beware the rising creatinine level. *J Card Fail*. 2003;9:26-8. doi: 10.1054/jcaf.2003.10.
3. McCarthy CP, Natarajan P. Systolic blood pressure and cardiovascular risk: straightening the evidence. *Hypertension*. 2023;80:577-9. doi: 10.1161/HYPERTENSIONAHA.123.20788.
4. Scudeler TL, da Costa LM, Nunes RA, Schneidewind RO, Brito TM, Pereira DC, et al. Association between low-density lipoprotein cholesterol levels and all-cause mortality in patients with coronary artery disease: a real-world

- analysis using data from an international network. *Sci Rep.* 2024; 14:29201. doi: 10.1038/s41598-024-80578-w.
5. Abera T, Peterson ED, Pagidipati NJ, Mulder H, Wojdyla DM, Philip S, et al. The association between triglycerides and incident cardiovascular disease: what is “optimal”? *J Clin Lipidol.* 2020;14:438-47. doi: 10.1016/j.jacl.2020.04.009.
 6. Park C, Guallar E, Linton JA, Lee DC, Jang Y, Son DK, et al. Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases. *Diabetes care.* 2013;36:1988-93. doi: 10.2337/dc12-1577.
 7. Fuchs FD, Whelton PK. High blood pressure and cardiovascular disease. *Hypertension.* 2020;75:285-92. doi: 10.1161/HYPERTENSIONAHA.119.14240.
 8. Leon BM, Maddox TM. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. *World J Diabetes.* 2015;6:1246-58. doi: 10.4239/wjd.v6.i13.1246.
 9. Petrie JR, Guzik TJ, Touyz RM. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Can J Cardiol.* 2018;34:575-584. doi: 10.1016/j.cjca.2017.12.005.
 10. Washio M, Sasazuki S, Kodama H, Yoshimasu K, Liu Y, Tanaka K, et al. Role of hypertension, dyslipidemia and diabetes mellitus in the development of coronary atherosclerosis in Japan. *Jpn Circ J.* 2001;65:731-37. doi: 10.1253/jcj.65.731.
 11. Chapakiya I, Traisuwan A, Chumpong S, Chumpong K. Follow-up period classification of type 2 diabetes patients using data mining techniques. *J Health Sci Med Res.* 2025;43:20241083. doi: 10.31584/jhsmr.20241083.
 12. Dinesh KG, Arumugaraj K, Santhosh KD, Mareeswari V. Prediction of cardiovascular disease using machine learning algorithms. In: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT); 2018 Mar 1-3; Coimbatore, India. *IEEE*;2018:1-7.
 13. Drozd K, Nabrdalik K, Kwiendacz H, Hendel M, Olejarz A, Tomasik A et al. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. *Cardiovasc Diabetol.* 2022;21:240. doi: 10.1186/s12933-022-01672-9.
 14. Enriko IK. Comparative study of heart disease diagnosis using top ten data mining classification algorithms. In: 5th International Conference on Frontiers of Educational Technologies; 2019 Jun 1-3; Beijing, China. *ACM*; 2019:159-64.
 15. Hasan N, Bao Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* 2021;11:49-62. doi: 10.1007/s12553-020-00499-2.
 16. Khanarsa P, Suwanmanee S, Chumpong S, Chumpong K. Enhancing diabetes follow-up period prediction through classification algorithms with feature selection techniques. *J Public Health Emerg.* 2025;25:9. doi: 10.21037/jphe-24-119.
 17. Maiga J, Hungilo GG. Comparison of machine learning models in prediction of cardiovascular disease using health record data. In: 2019 International Conference on Informatics, Multimedia, Cyber and Information

- System (ICIMCIS); 2019 Oct 24-25; Jakarta, Indonesia. IEEE;2019: 45-8.
18. Saqlain M, Hussain W, Saqib NA, Khan MA. Identification of heart failure by using unstructured data of cardiac patients. In: 2016 45th International Conference on Parallel Processing Workshops (ICPPW); 2016 Aug 16-19; Philadelphia, United States. IEEE;2016:426-431.
19. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci. 2020;1:345. doi: 10.1007/s42979-020-00365-y.
20. Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC bioinformatics. 2016;17:359. doi: 10.1186/s12859-016-1236-x.
21. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6:10. doi: 10.1186/1758-2946-6-10.
22. Debnath R, Takahide N, Takahashi H. A decision based one-against-one method for multi-class support vector machine. Pattern Anal Appl. 2004;7:164-75. doi: 10.1007/s10044-004-0213-6.
23. Devetyarov D, Nouretdinov I. Prediction with confidence based on a random forest classifier. In: 6th IFIP WG 12.5 International Conference on Artificial Intelligence Applications and Innovations (AIAI); 2010 Oct 6-7; Larnaca, Cyprus. Springer; 2010;37-44.
24. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer. 1996;29:31-44. doi: 10.1109/2.485891.
25. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395-99. doi: 10.1161/CIRCULATIONAHA.106.682658.
26. Scholarpedia. K-nearest neighbor [Internet]. [cited 2024 Nov 27]. Available from: http://www.scholarpedia.org/article/K-nearest_neighbor.
27. Song YY, Ying LU. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-35. doi: 10.11919/j.issn.1002-0829.215044.
28. Yang FJ. An implementation of naive bayes classifier. In: 2018 International Conference on Computational Science and Computational Intelligence (CSCI); 2018 Dec 12-14; Las Vegas, United States. IEEE;2018:301-6.
29. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing. 2020;415:295-316. doi: 10.1016/j.neucom.2020.07.061.
30. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 2009;77:103-23. doi: 10.1007/s10994-009-5119-5.
31. Pope JH, Aufderheide TP, Ruthazer R, Woolard RH, Feldman JA, Beshansky JR, Griffith JL, Selker HP. Missed diagnoses of acute cardiac ischemia in the emergency department. N Engl J Med. 2000;342:1163-70. doi: 10.1056/NEJM200004203421603.
32. Vujović Ž. Classification model evaluation metrics. Int J Adv Comput Sci Appl. 2021;12:599-606. doi: 10.14569/IJACSA.2021.0120670.
33. Warraich HJ, Kaltenbach LA, Fonarow GC, Peterson ED, Wang TY. Adverse change in employment status after acute myocardial infarction: analysis from the TRANSLATE-ACS study. Circ Cardiovasc Qual Outcomes. 2018;11:

004528. doi: 10.1161/CIRC
OUTCOMES.117.004528.
34. Nohara Y, Matsumoto K, Soejima H,
Nakashima N. Explanation of machine
learning models using shapley additive
explanation and application for real
data in hospital. *Comput Methods
Programs Biomed.* 2022;214:106584.
doi: 10.1016/j.cmpb.2021.106584.
35. DiCiccio TJ, Efron B. Bootstrap
confidence intervals. *Statistical science.*
1996;11:189-228. doi: 10.1214/ss/
1032280214.