

<https://doi.org/10.1038/s41514-025-00238-9>

STEMI-OP in-hospital mortality prediction algorithms: Frailty-integrated machine learning in older patients undergoing primary PCI

Check for updates

Tan Van Nguyen^{1,2}, Quyen The Nguyen³✉, Huong Quynh Nguyen⁴, Nghia Thuong Nguyen⁵, Khai Duc Luong^{1,2}, Lan Hoang Do Thi^{1,2}, Tu Cam Nguyen^{1,2}, Thuan Hoang Vo^{1,2}, Phan Huu Le^{1,2}, Phuc Thien Tran^{1,2} & Thanh Dinh Le⁶✉

Despite advances in medical care, older patients with ST-elevation myocardial infarction (STEMI) undergoing primary percutaneous coronary intervention (PCI) currently face high in-hospital mortality rates. Traditional prognostic models, primarily developed in Caucasian populations with fewer older participants and using classical statistical approaches, may not perform well in Southeast Asian settings. This study explores the need for artificial intelligence-based risk assessment models—the STEMI-OP algorithms—designed explicitly for STEMI patients aged 60 and older following primary PCI in Vietnam. Machine learning (ML) models were developed and validated using pre- and post-PCI features, with advanced feature selection techniques to identify key predictors. SHapley Additive exPlanations and Causal Random Forests were employed to improve interpretability and causal relationships between features and outcomes, highlighting the key factors, including the Killip classification, the Clinical Frailty Scale, glucose levels, and creatinine levels in predicting in-hospital mortality. The CatBoost model with ElasticNet regression for pre-PCI prediction and the Random Forest model with Ridge regression post-PCI prediction demonstrated significantly superior performance compared to traditional risk scores, achieving AUC values of 92.16% and 95.10%, respectively, outperforming the GRACE 2.0 score (83.48%) and the CADILLAC score (87.01%). By incorporating frailty and employing advanced ML techniques, the STEMI-OP algorithms produced more precise, personalized risk assessments that could enhance clinical decision-making and improve outcomes for older STEMI patients undergoing primary PCI.

Despite decades of application of primary percutaneous coronary intervention (PCI), in-hospital mortality among older ST-elevation myocardial infarction (STEMI) patients remains a critical concern. A 20-year registry in Germany revealed that in-hospital mortality in women aged ≥ 75 years showed little change, from 25.1% in 2000 to 23.6% in 2019, and remained significantly higher than in younger patients¹. Similar data from Singapore indicated in-hospital mortality

rates of 11.9% in older patients compared to 3.6% in younger cohorts following primary PCI².

Prognostic models predict patient outcomes, guide clinical decisions and improve care quality. However, significant gaps in the evidence underlying these models can reduce their accuracy and applicability in older patients following primary PCI. Regional differences in healthcare systems, patient demographics, and treatment approaches pose challenges for

¹Department of Geriatrics and Gerontology, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Vietnam. ²Thong Nhat Hospital, Ho Chi Minh City, Vietnam. ³Department of Geriatrics, Pham Ngoc Thach University of Medicine, Ho Chi Minh City, Vietnam. ⁴School of Biomedical Engineering, International University, Vietnam National University, Ho Chi Minh City, Vietnam. ⁵Department of Interventional Cardiology, Cho Ray Hospital, Ho Chi Minh City, Vietnam. ⁶Department of Internal Medicine, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam. ✉e-mail: quyendr0809@gmail.com; ldthanh@ntt.edu.vn

applying traditional models universally. Most widely recognized mortality risk scores such as GRACE 2.0, CADILLAC, TIMI, PAMI, and ALPHA were developed predominantly from Caucasian populations, with limited representation from Asian populations, necessitating local adaptation and validation, particularly in regions like Southeast Asia^{3–7}.

Older populations, often characterized by multiple comorbidities and higher risks of adverse outcomes, have been underrepresented in the clinical trials and registries used for model development. Consequently, traditional risk scores were developed based on cohorts primarily comprising younger and less clinically complex populations, which raised concerns about potential inaccuracies when these models are applied to older patient groups^{3–7}. Furthermore, a considerable number of younger STEMI patients were enrolled in the Southeast Asia validation cohorts used to develop these prognostic models. Notably, even the most recent machine learning (ML) models designed explicitly for STEMI patients undergoing primary PCI in Asia included a modest representation of older individuals within their cohorts^{8–14}. These limitations have further contributed to restricting the applicability of these models to older populations within this region.

Additionally, to our best-known knowledge, existing traditional and ML models lacked integration of geriatric-specific factors such as frailty, cognitive function, and functional status, which are critical in predicting outcomes in older patients^{3–14}. Frailty, commonly measured by the Clinical Frailty Scale (CFS)¹⁵, has been consistently associated with higher mortality in older patients following primary PCI^{16–19}. The exclusion of such factors from conventional models may lead to an incomplete risk assessment and underestimation of mortality in older populations. These gaps in evidence highlight the critical requirement of incorporating geriatric-specific factors to enhance the accuracy of mortality prediction models for older adults undergoing primary PCI in Southeast Asia.

Traditional models assume linear relationships between risk factors and mortality, which oversimplify the complexity of these interactions^{3–7}. ML algorithms offer a more nuanced approach, incorporating non-linear relationships and leveraging ensemble models that combine multiple algorithms to enhance prediction accuracy, reduce overfitting, and improve model robustness^{20–22}. To enhance interpretability, tree-based methods such as Random Forest offered feature importance scores, revealing the impact of each variable on the outcome²³. Yet, as noted by Lundberg et al., these metrics can sometimes be inconsistent²⁴. To overcome this limitation, the SHapley Additive exPlanation (SHAP) framework offers a more consistent and interpretable approach to Feature Importance by attributing each feature's contribution to the model predictions using Shapley values. In addition, Causal Random Forests (CRF) have emerged as a powerful tool for exploring causal relationships between features and outcomes²⁵. Unlike traditional methods, CRF estimates heterogeneous treatment effects, identifying how different variables influence outcomes across diverse subgroups. This approach improves the understanding of causal mechanisms and enables the development of more robust and generalizable models.

By integrating SHAP, CRF, and accounting for frailty as a geriatric-specific factor, this study aims to build a transparent and accurate ML risk calculator—the STEMI-OP (STEMI-Older Persons) algorithms—for older patients undergoing primary PCI in Vietnam. This comprehensive strategy overcomes the limitations of traditional and previous ML models, enhancing clinical decision-making and improving patient outcomes.

Results

The baseline characteristics of patients across the four centers included in the cohort are detailed in Table 1. The data demonstrates significant inter-centre variability in several key clinical parameters, which merit further attention.

The prevalence of cardiac arrest occurring before or at admission was markedly higher in Centers 3 and 4 (~3.0%) compared to Centers 1 and 2 (around 1.0%). Center 3 also exhibited significantly lower systolic blood pressure than the other centers. A history of diabetes mellitus was observed in approximately 17.0% of patients in Centers 1 and 3, whereas this proportion increased substantially to around 35.0% in Centers 2 and 4.

The duration from symptom onset to hospital admission was notably shorter at Center 4, with a median of 4 hours, compared to a significantly longer median of 7–8 hours at the other three centers. This disparity highlights the unique distribution of PCI-capable centres in developing countries, such as Vietnam, which are predominantly located in central urban areas. Consequently, most patients in this cohort experience prolonged travel times from their first medical contact to a PCI-capable center.

Anterior STEMI was the predominant type in Centers 2 and 4, comprising approximately 55.0%–60.0% of cases, in contrast to Centers 1 and 3, accounting for ~45.0%. The incidence of ventricular fibrillation or ventricular tachycardia was disproportionately higher in Center 4 (10.9%). Left ventricular ejection fraction (LVEF) was comparable across Centers 1, 3, and 4, ranging between 42.0% and 45.0%, but was notably higher in Center 2, at 51.9%.

The use of ticagrelor was relatively limited in Centers 3 (26.7%) and 4 (1.7%), where clopidogrel remained the antiplatelet agent of choice. By contrast, Centers 1 and 2 demonstrated a significantly higher adoption of ticagrelor (~64.0%). These differences can be attributed to the timeline of patient recruitment; Centers 3 and 4 primarily enrolled patients in 2017–2018, when clopidogrel was widely prescribed in Vietnam. Following subsequent updates to guidelines highlighting ticagrelor's superior efficacy over clopidogrel in the primary PCI setting, its use saw a significant increase in Centers 1 and 2.

Radial artery access was preferred in Centers 1 and 3 (over 90.0%), whereas femoral artery access was predominantly used in Centers 2 and 4 (93.0%–100.0%). Left main coronary artery involvement was observed in approximately 15.0% of cases in Centers 2 and 4, compared to around 5.0% in Centers 1 and 3.

These demographic and clinical variations among centers likely influenced in-hospital mortality rates, ranging from 10.0% to 12.0% across most centers but reaching a peak of 17.9% in Center 2.

Pre-PCI models

Feature selection. The performance metrics of various ML models using the complete set of features are outlined in Supplementary Table S1. The Catboost model demonstrated strong performance, with an accuracy of 90.67% and an AUC (Area Under the Curve) of 91.66%, indicating its effectiveness in differentiating diagnostic outcomes. In contrast, while the Random Forest model achieved an AUC of 91.88%, its lower sensitivity of 43.64% suggested challenges in identifying true positive cases. These results highlighted the need to improve model sensitivity without compromising accuracy or specificity. Feature selection methods help optimise models by reducing complexity and improving computational efficiency, making them more interpretable for clinical use.

The application of feature selection methods demonstrated that reducing the feature set to 10 still maintained robust model performance. Specifically, the CatBoost model, optimised with features selected through ElasticNet regression, achieved an improved AUC of 92.16%, an increased sensitivity of 59.74%, an enhanced G-Mean of 75.33%, and an F1 score of 90.21%. These models outperformed traditional risk-scoring systems, such as GRACE 2.0 (AUC of 83.48%) and CADILLAC (AUC of 87.01%), as shown in Table 2. DeLong's test was conducted using CatBoost (ElasticNet) as the reference model to assess the statistical significance of these differences. The analysis confirmed that the improvements in AUC over all traditional scores were statistically significant (all *p* values < 0.0001), further supporting the superior discriminative performance of the proposed ML approach. The complete performance comparisons of pre-PCI ML models with various feature selection methods are presented in Supplementary Table S2.

Key features selected by each method are detailed in Supplemental Table S3. Notably, the Killip class, the CFS, hemoglobin level, and heart rate were among the most chosen features, underscoring their importance across various predictive models.

Evaluation and interpretation of the top-performing model. The top-performing ML model underwent a comprehensive evaluation to ensure

Table 1 | Patient characteristics

	Center 1 N = 749	Center 2 N = 285	Center 3 N = 251	Center 4 N = 175	P value
<i>Demographic features</i>					
Age	69.8 ± 7.6	69.7 ± 8.3	70.5 ± 7.8	72.5 ± 9.9	0.041
Weight (kg)	57.8 ± 9.3	57.5 ± 10.5	57.8 ± 9.6	57.9 ± 9.3	0.981
Height (cm)	160.6 ± 6.8	166.0 ± 107.2	159.6 ± 6.9	160.1 ± 7.5	0.329
Gender					0.003
Female	203 (27.1)	98 (34.4)	93 (37.1)	45 (25.7)	
Male	546 (72.9)	187 (65.6)	158 (62.9)	41 (74.3)	
Smoking	292 (39.0)	63 (22.1)	95 (37.8)	54 (30.9)	<0.001
<i>Clinical presentations</i>					
Time from onset to admission (hours)	8.0 (6.0–12.0)	7.0 (3.0–24.0)	8.0 (6.0–11.0)	4.0 (2.0–16.5)	0.019
Heart rate (bpm)	80.0 ± 19.4	81.5 ± 21.1	83.3 ± 21.5	77.5 ± 30.8	0.049
Systolic blood pressure (mmHg)	120.0 ± 26.4	127.4 ± 30.2	113.1 ± 28.7	126.6 ± 39.4	<0.001
Diastolic blood pressure (mmHg)	72.1 ± 15.8	77.3 ± 17.5	68.6 ± 17.8	74.0 ± 21.7	<0.001
Cardiac arrest before/at admission	9 (1.2)	2 (0.7)	7 (2.8)	6 (3.4)	0.050
Killip classification					0.039
I	579 (77.3)	211 (74.0)	199 (79.3)	130 (74.3)	
II	86 (11.5)	33 (11.6)	24 (9.5)	22 (12.6)	
III	17 (2.3)	20 (7.0)	11 (4.4)	13 (7.4)	
IV	67 (8.9)	21 (7.4)	17 (6.8)	10 (5.7)	
Clinical Frailty Scale					0.031
1	13 (1.7)	6 (2.1)	4 (1.6)	6 (3.4)	
2	172 (23.0)	56 (19.6)	66 (26.3)	51 (29.2)	
3	112 (15.0)	42 (14.7)	47 (18.7)	13 (7.4)	
4	296 (39.5)	108 (37.9)	81 (32.2)	80 (45.7)	
5	121 (16.2)	47 (16.5)	39 (15.6)	22 (12.6)	
6	16 (2.1)	19 (6.7)	10 (4.0)	0 (0.0)	
7	19 (2.5)	7 (2.5)	4 (1.6)	3 (1.7)	
<i>Comorbidities</i>					
Hypertension	539 (72.0)	223 (78.2)	181 (72.1)	143 (81.7)	0.085
Dyslipidemia	215 (28.7)	81 (28.4)	18 (7.2)	48 (27.3)	<0.001
Diabetes mellitus	127 (17.0)	94 (33.0)	46 (18.3)	21 (38.2)	<0.001
Chronic kidney disease	21 (2.8)	10 (3.5)	8 (3.2)	16 (9.1)	0.084
Stroke	39 (5.2)	7 (2.5)	13 (5.2)	13 (7.4)	0.151
Transient ischemic attack	12 (1.6)	4 (1.4)	5 (2.0)	13 (7.4)	0.063
Peripheral artery disease	3 (0.4)	0 (0.0)	0 (0.0)	3 (1.7)	0.122
Heart failure	45 (6.0)	11 (3.9)	1 (0.4)	10 (5.7)	<0.001
Chronic lung disease	28 (3.7)	5 (1.8)	7 (2.8)	3 (1.7)	0.388
Prior angina	88 (11.7)	10 (3.5)	24 (9.6)	16 (9.1)	<0.001
Prior myocardial infarction	30 (4)	11 (3.9)	11 (4.4)	16 (9.1)	0.335
Prior PCI	30 (4)	28 (9.8)	8 (3.2)	13 (7.4)	<0.001
Prior CABG	2 (0.3)	0 (0.0)	1 (0.4)	0 (0.0)	1.000
<i>Electrocardiographic features</i>					
Infarction wall					<0.001
Anterior	342 (45.7)	145 (50.9)	114 (45.4)	115 (65.7)	
Inferior	388 (51.8)	127 (44.6)	123 (49.0)	57 (32.6)	
Posterior	12 (1.6)	7 (2.5)	11 (4.4)	3 (1.7)	
Lateral	7 (0.9)	6 (2.1)	3 (1.2)	0 (0.0)	
Left bundle branch block	40 (5.3)	9 (3.2)	6 (2.4)	6 (3.4)	0.064
Atrioventricular block	113 (15.1)	32 (11.2)	26 (10.4)	38 (21.7)	0.029
Atrial flutter/fibrillation	22 (2.9)	6 (2.1)	8 (3.2)	0 (0.0)	0.534

Table 1 (continued) | Patient characteristics

	Center 1 N = 749	Center 2 N = 285	Center 3 N = 251	Center 4 N = 175	P value
Ventricular tachycardia/fibrillation	9 (1.2)	3 (1.1)	10 (4.0)	19 (10.9)	<0.001
<i>Echocardiographic features</i>					
Left ventricular ejection fraction (%)	42.8 ± 10.1	51.9 ± 13.3	45.2 ± 11.4	42.7 ± 14.1	<0.001
Pericardial effusion	20 (2.6)	6 (2.1)	6 (2.4)	10 (5.7)	0.769
Regional wall motion abnormality	710 (94.8)	243 (85.3)	174 (69.3)	130 (74.3)	<0.001
<i>Laboratory data</i>					
High-sensitivity Troponin T (pg/mL)	592.0 (74.0–1936.0)	352.7 (52–1472.0)	948.0 (71.0–2000.0)	297.0 (102.65–1717.5)	0.042
Hemoglobin (g/L)	129.8 ± 17.4	130.4 ± 17.4	131.3 ± 19.9	130.1 ± 19.0	0.618
White blood cell (K/ μ L)	12.6 ± 4.2	11.91 ± 4.5	11.9 ± 4.6	12.0 ± 5.4	0.027
Glucose (mg/dL)	155.7 ± 73.0	194.2 ± 118.9	157.5 ± 83.1	182.7 ± 69.9	<0.001
Creatinine (μ mol/L)	93.8 ± 49.6	101.7 ± 40.7	105.0 ± 36.2	108.0 ± 31.6	<0.001
<i>Medications at admission</i>					
Aspirin	749 (100.0)	285 (100.0)	371 (100.0)	55 (100.0)	.
P2Y12 inhibitor					<0.001
Ticagrelor	476 (63.6)	184 (64.6)	67 (26.7)	3 (1.7)	
Clopidogrel	273 (36.4)	101 (35.4)	184 (73.3)	172 (98.3)	
Statin	749 (100.0)	285 (100.0)	371 (100.0)	55 (100.0)	.
<i>PCI procedural data</i>					
Time from admission to wiring (hours)	12.3 (9.0–18.2)	9.7 (5.8–27.3)	12.0 (9.0–16.0)	7.0 (4.3–16.9)	0.002
Vascular access site					<0.001
Radial	716 (95.6)	20 (7.0)	231 (92.0)	0 (0.0)	
Femoral	33 (4.4)	265 (93.0)	20 (8.0)	175 (100.0)	
Number of diseased vessels					<0.001
1	291 (38.9)	64 (22.5)	103 (41.0)	29 (16.6)	
2	266 (35.5)	89 (31.2)	82 (32.7)	70 (40.0)	
3	192 (25.6)	132 (46.3)	66 (26.3)	76 (43.4)	
Left main coronary artery disease	54 (7.2)	46 (16.1)	9 (3.6)	22 (12.6)	<0.001
Pre-PCI TIMI blood flow					<0.001
0	413 (55.1)	176 (61.8)	127 (50.6)	111 (63.4)	
1	72 (9.6)	33 (11.6)	27 (10.8)	35 (20.0)	
2	143 (19.1)	60 (21.1)	97 (38.6)	29 (16.6)	
3	121 (16.2)	16 (5.6)	0 (0.0)	0 (0.0)	
Post-PCI TIMI blood flow					0.003
0	1 (0.1)	2 (0.7)	0 (0.0)	0 (0.0)	
1	3 (0.4)	5 (1.8)	2 (0.8)	0 (0.0)	
2	13 (1.7)	14 (4.9)	14 (5.6)	13 (7.4)	
3	732 (97.7)	264 (92.6)	235 (93.6)	162 (92.6)	
Balloon pre-dilation	559 (74.6)	267 (93.7)	147 (58.6)	165 (94.3)	<0.001
Thrombectomy	64 (8.5)	20 (7.0)	18 (7.2)	16 (9.1)	0.794
Balloon post-dilation	239 (31.9)	161 (56.5)	85 (33.9)	102 (58.3)	<0.001
Type of stent					0.005
No stent	19 (2.5)	5 (1.8)	0 (0.0)	0 (0.0)	
Bare metal stent	27 (3.6)	4 (1.4)	4 (1.6)	0 (0.0)	
Drug-eluting stent	703 (93.9)	276 (96.8)	247 (98.4)	175 (100.0)	
In-hospital mortality	80 (10.7)	51 (17.9)	26 (10.4)	21 (12.0)	0.012

Values are mean ± SD, n (%), or median (Q1–Q3).

bpm beats per minute, CABG coronary artery bypass grafting, PCI percutaneous coronary intervention, STEMI ST-segment elevation myocardial infarction, TIMI thrombolysis in myocardial infarction

Table 2 | Pre-PCI models

Models	Selection methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	G-mean (%)	F1 score (%)	AUC (95% CI) (%)	p value (DeLong's test)
CatBoost	ElasticNet	90.55	59.74	95.69	75.33	90.21	92.16 (91.93–92.38)	Reference
Logistic regression	RFE	86.24	81.38	87.20	84.17	87.34	90.10 (89.83–90.37)	0.9615
GRACE 2.0	-	80.38	78.38	82.38	80.01	57.30	83.48 (71.02–92.36)	<0.0001
TIMI	-	76.21	74.07	78.36	76.15	49.94	81.76 (70.49–89.05)	<0.0001
CADILLAC	-	81.79	85.36	78.21	81.34	58.27	87.01 (79.49–91.86)	<0.0001
PAMI	-	70.91	64.04	77.77	70.02	44.91	75.12 (62.25–84.99)	<0.0001
NCDR CathPCI v4	-	73.35	78.89	67.81	71.68	43.14	77.64 (66.62–85.52)	<0.0001
ALPHA	-	78.04	71.68	84.40	77.54	56.33	78.38 (63.65–89.29)	<0.0001
APEX AMI	-	79.76	71.70	87.81	79.15	60.20	82.02 (68.88–91.34)	<0.0001
Zwolle	-	81.13	78.29	83.97	80.87	57.61	85.16 (77.67–90.51)	<0.0001

Performance comparison of CatBoost model with ElasticNet feature selection, logistic regression model with RFE feature selection, and traditional models. AUC, area under the curve, CI Confidence Interval.

a thorough understanding of its predictive capabilities and the factors driving its performance. This evaluation included an analysis of feature importance metrics, SHAP values, calibration plots, and causal effect assessments to investigate the relationships between features and outcomes. The evaluation incorporated SHAP summaries, causal effect analyses, and dependence plots, providing detailed insights into how individual features influenced model predictions and highlighting the alignment between Feature Importance rankings and SHAP values. Additionally, the causal effect analyses elucidated the direct and indirect relationships between features and outcomes, reinforcing the robustness and interpretability of the model.

Feature importance. As illustrated in Fig. 1, the CatBoost model, refined through ElasticNet regression for feature selection, highlighted the CFS as the most pivotal predictor of patient outcomes. This was closely followed by glucose levels, the Killip class, systolic blood pressure, and the time interval from symptom onset to hospital admission, which emerged as other critical determinants influencing the model's predictive accuracy. Additional variables, including heart rate, creatinine levels, hemoglobin levels, patient age, and ventricular tachycardia/fibrillation at admission, also demonstrated substantial significance, albeit with a lesser impact.

Performance of the top-ranking model. The comparative analysis of the CatBoost model with ElasticNet feature selection, the Logistic Regression model enhanced with RFE feature selection, and traditional risk models provides valuable insights into the balance between model complexity and interpretability in predictive modeling. Advanced ML models like CatBoost excel at capturing intricate, non-linear patterns in data, yielding superior predictive accuracy. However, these benefits come at the cost of reduced interpretability and greater computational demands compared to simpler models such as Logistic Regression.

Logistic Regression, particularly when paired with RFE feature selection, strikes a compelling balance by offering enhanced interpretability, computational efficiency, and competitive performance, making it well-suited for specific clinical applications. Among the Logistic Regression approaches, the RFE-selected model stood out as the top performer. Nonetheless, the CatBoost model with ElasticNet feature selection demonstrated the highest AUC, outperforming all other models, including the RFE-enhanced Logistic Regression.

Performance differences were apparent in several evaluation metrics, including ROC curves, calibration plots, and precision-recall (PR) curves. The CatBoost model exhibited the most reliable calibration (Fig. 2b), aligning closely with the ideal diagonal across most probability ranges, with only minor overestimations at higher probabilities. In contrast, the Logistic Regression model with RFE displayed moderate calibration, exhibiting a tendency to overestimate risk across all levels of observed probabilities.

The CatBoost model consistently maintained superior precision across all recall levels in the PR domain (Fig. 2c), particularly excelling in high-recall scenarios. These findings underscore the CatBoost model's superior discrimination, calibration, and precision when applying ElasticNet feature selection, highlighting its potential for robust application in clinical settings.

SHapley additive exPlanations and causal random forests analysis. The CatBoost model, refined with features selected through ElasticNet regression, was evaluated using SHAP analysis (Figs. 3a and 4) and Causal Random Forests (Fig. 3b), highlighting their distinct yet complementary contributions to elucidating model predictions.

The SHAP summary plot (Fig. 3a) demonstrates individual features' relative importance and contributions to the model's predictions. Among these, the Killip class emerged as the most critical predictor, followed by the CFS, glucose, systolic blood pressure, and the time interval from symptom onset to hospital admission. Higher feature values (depicted in purple) were associated with an increased mortality risk, whereas lower values (shown in green) corresponded to a decreased risk. Other variables, including creatinine levels, heart rate, patient age, hemoglobin levels, and ventricular

Fig. 1 | Pre-PCI model. Feature importance of CatBoost model after applying ElasticNet regression method for feature selection.

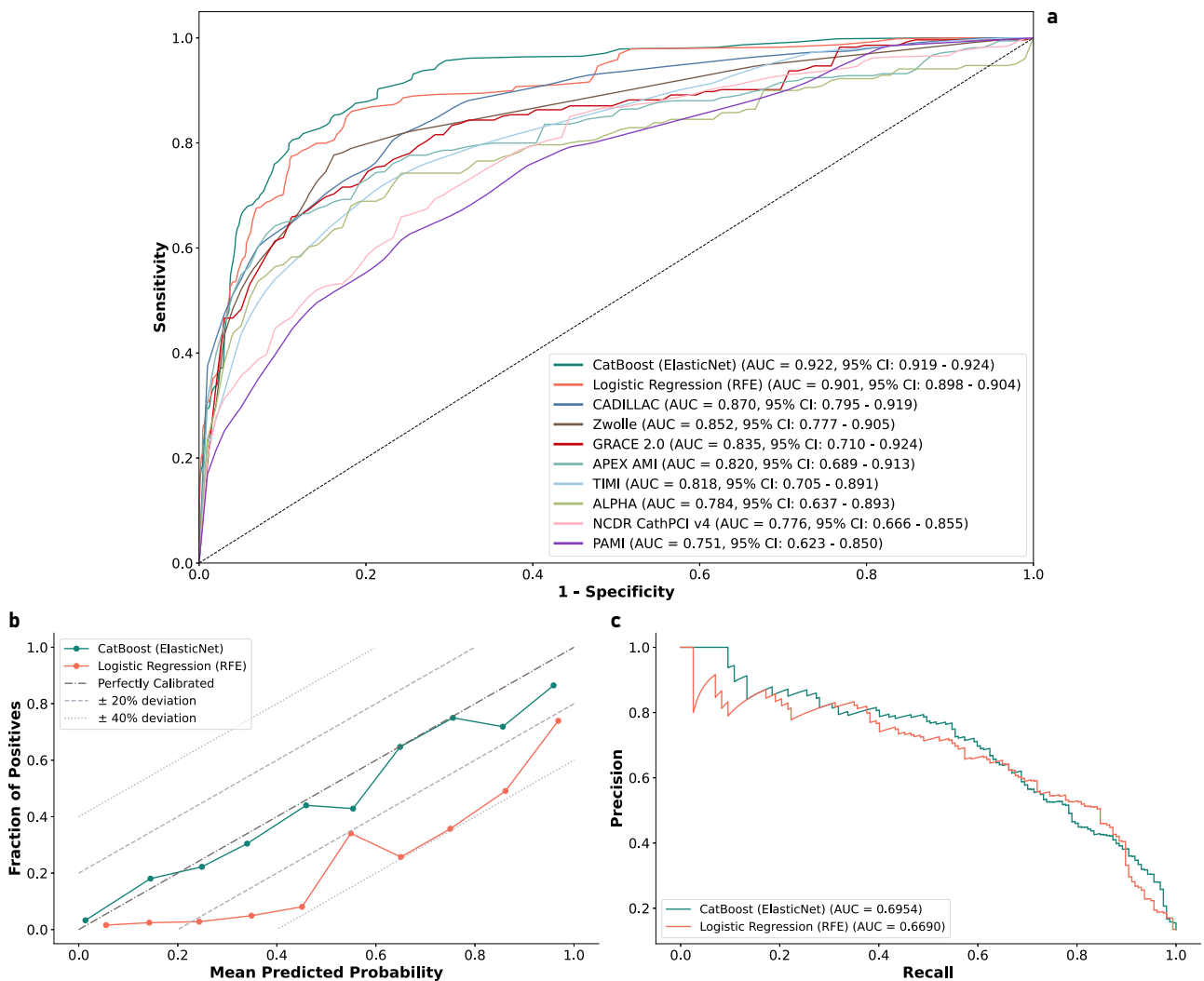
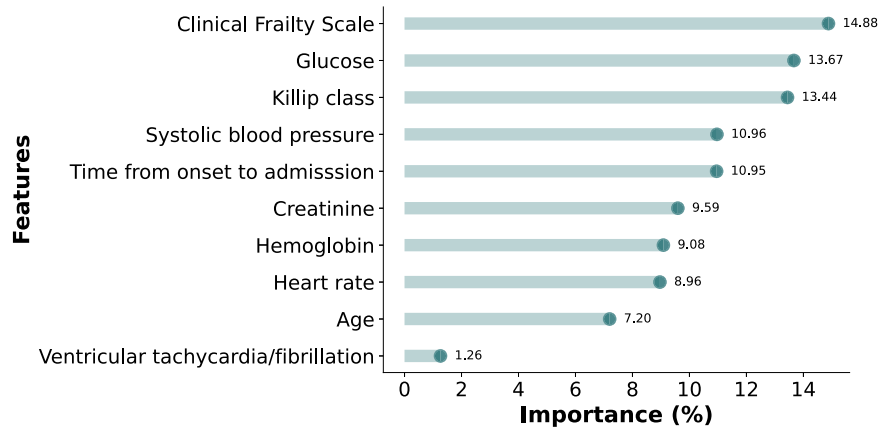


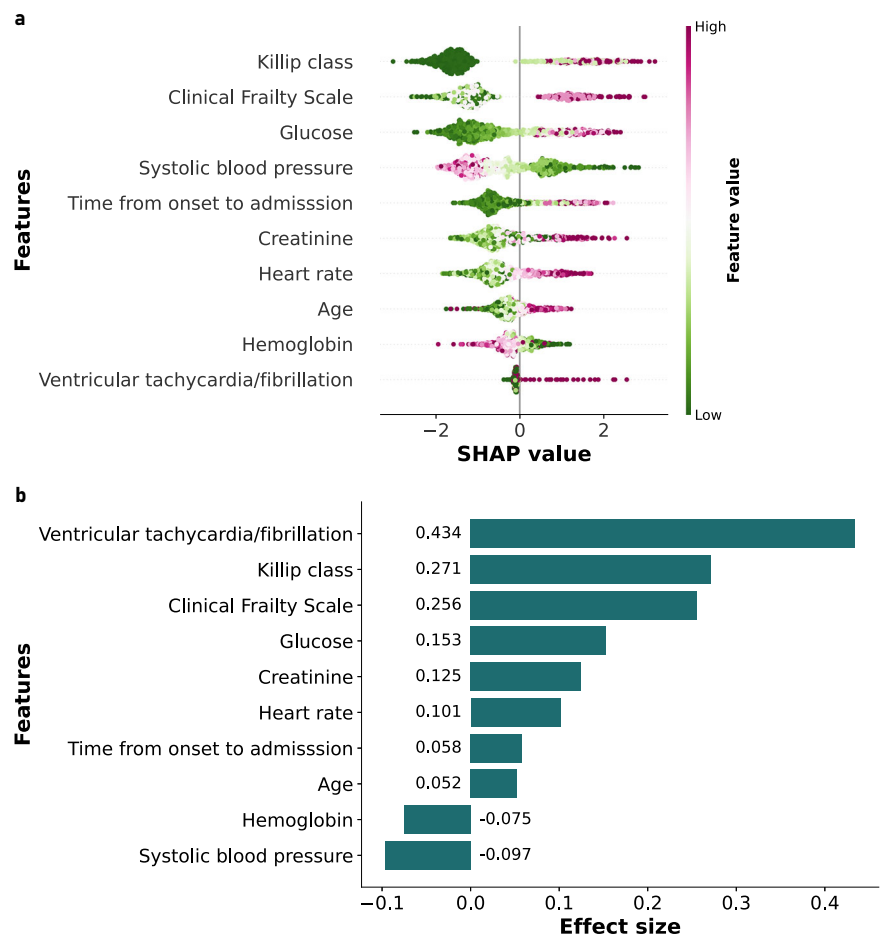
Fig. 2 | Pre-PCI model. CatBoost model with ElasticNet feature selection versus Logistic Regression model with RFE feature selection versus traditional models. **a** Receiver operating characteristic (ROC) curves, **b** calibration plots, **c** precision-recall curves. AUC area under the curve, CI confidence interval, PR precision-recall.

tachycardia/fibrillation at admission, also played significant roles. This visualisation encapsulates the magnitude and direction of each feature’s impact while capturing the complex, non-linear, and interactive effects that influence the model’s predictions.

SHAP dependence plots (Fig. 4) offer deeper insights into the relationships between individual features and their contributions to the model.

These plots illustrate several critical patterns: a strong positive association of the Killip class and the CFS with mortality risk, non-linear glucose and creatinine level effects, and an inverted U-shaped impact of systolic blood pressure and heart rate. Furthermore, these plots underscore the critical role of timely admission, revealing a strong correlation between delayed admission and increased mortality risk.

Fig. 3 | Pre-PCI model. SHAP summary plot and causal effect of CatBoost model with ElasticNet feature selection. **a** SHAP summary plot, **b** causal effect.



The Causal Random Forests analysis (Fig. 3b) complements these findings by quantifying the direct impact of each feature on patient outcomes, thereby providing a causal perspective. The presence of ventricular tachycardia/fibrillation at admission exhibited the most substantial causal influence, followed by the Killip class, corroborating its prominent role in the SHAP analysis. Similarly, the CFS, glucose, and creatinine levels also displayed significant causal effects. This alignment reinforces the conclusions drawn from the SHAP analysis, affirming that the most influential features identified by SHAP also hold substantial causal relevance.

SHAP analysis explains the magnitude and direction of feature contributions to individual predictions, accounting for both linear and non-linear interactions. In contrast, causal effect analysis focuses on direct relationships between features and outcomes. The strong agreement between SHAP-identified influential features—such as the Killip class, the CFS, glucose, and creatinine levels—and their corresponding causal effects highlight these variables’ robustness and clinical significance. This synergy enhances the model’s interpretability, reliability, and clinical relevance, providing a comprehensive understanding of patient mortality risk.

Post-PCI models

Feature selection. In the post-PCI assessment, ML models with a complete set of features exhibited marked improvements in performance metrics compared to pre-PCI models (Supplementary Table S4). These improvements were achieved without feature selection, most likely attributable to the inclusion of additional post-PCI features. The consistently improved metrics across models highlight their capacity to more accurately differentiate clinical outcomes, paving the way for further exploration of feature selection methods to refine these results while balancing computational efficiency.

As shown in Table 3, the Random Forest with Ridge feature selection achieved the highest AUC of 95.10%, outperforming its pre-PCI performance, where the same method yielded an AUC of 91.36%. Additionally, compared to traditional risk-scoring systems like GRACE 2.0 and CADILLAC, the observed improvements in accuracy and AUC highlighted the superior predictive capabilities of post-PCI ML models, especially when combined with optimal feature selection methods. These improvements were further supported by statistical comparison using DeLong’s test, which confirmed that the observed differences in AUC were highly significant (all *p* values < 0.0001). The complete performance comparisons of post-PCI ML models with various feature selection methods are presented in Supplementary Table S5.

Key features selected by each method are detailed in Supplemental Table S6. Once again, the Killip class and the CFS were the top two most frequently chosen features across various types of models.

Evaluation and interpretation of the top-performing model

Feature importance. A comparison of Feature Importance between the pre-PCI top-performing model (Fig. 1) and the post-PCI top-performing model (Fig. 5) reveals distinct shifts in the key predictors of patient outcomes. In the pre-PCI model, primary drivers of mortality risk included the CFS, glucose levels, the Killip class, and systolic blood pressure, underscoring the predictive value of baseline clinical and laboratory parameters before intervention. In contrast, the post-PCI model retained the Killip class, glucose levels, the CFS, and creatinine levels as pivotal predictors but introduced the significance of post-procedural factors, such as LVEF, post-PCI TIMI blood flow, and the interval from symptom onset to wire crossing. This shift highlights the evolving role of procedural and recovery-related variables in shaping outcomes.

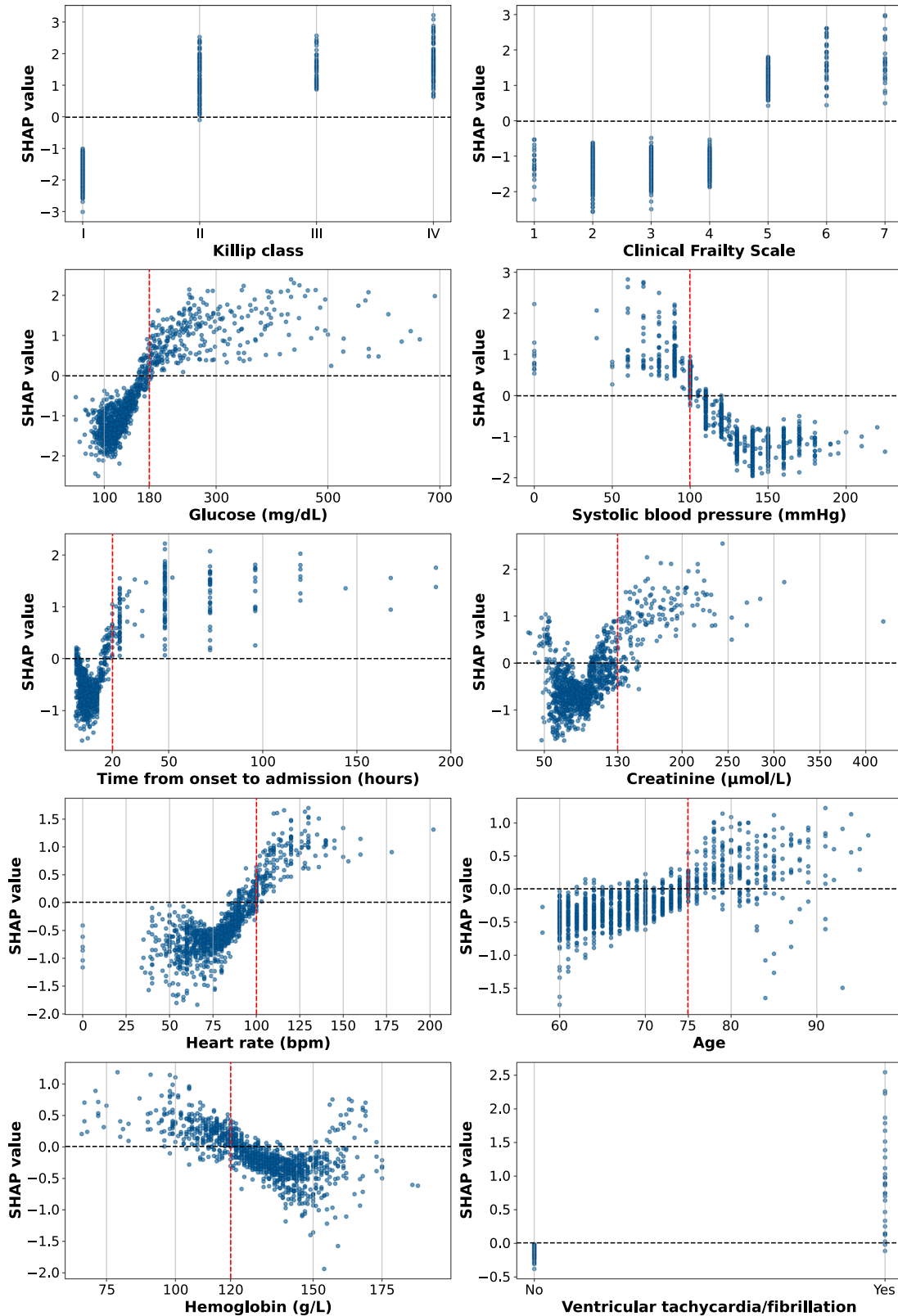


Fig. 4 | Pre-PCI model. SHAP dependence plot of the 10 features of the CatBoost model with ElasticNet feature selection. bpm, beat per minute.

Performance of the top-ranking model. Consistent with the pre-PCI analysis, the Random Forest model with Ridge feature selection emerged as the top performer, achieving the highest AUC and outperforming the Logistic Regression model with Lasso feature selection and established risk models like

GRACE 2.0 and TIMI. However, the performance gap between the Random Forest and Logistic Regression models narrowed slightly in the post-PCI context, suggesting that incorporating procedural and outcome-related features enhances predictive capability across complex and simpler models.

Table 3 | Post-PCI model

Models	Selection methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	G-mean (%)	F1 score (%)	AUC (95% CI) (%)	p value (DeLong's test)
Random Forest	Ridge	92.12	63.44	97.10	77.81	91.62	95.10 (94.94–95.26)	Reference
Logistic regression	Bayesian	87.98	83.99	88.61	86.17	88.88	94.39 (94.20–94.59)	0.8414
GRACE 2.0	-	80.38	78.38	82.38	80.01	57.30	83.48 (71.02–92.36)	<0.0001
TIMI	-	76.21	74.07	78.36	76.15	49.94	81.76 (70.49–89.05)	<0.0001
CADILLAC	-	81.79	85.36	78.21	81.34	58.27	87.01 (79.49–91.86)	<0.0001
PAMI	-	70.91	64.04	77.77	70.02	44.91	75.12 (62.25–84.99)	<0.0001
NCDR CathPCI v4	-	73.35	78.89	67.81	71.68	43.14	77.64 (66.62–85.52)	<0.0001
ALPHA	-	78.04	71.68	84.40	77.54	56.33	78.38 (63.65–89.29)	<0.0001
APEX AMI	-	79.76	71.70	87.81	79.15	60.20	82.02 (68.88–91.34)	<0.0001
Zwolle	-	81.13	78.29	83.97	80.87	57.61	85.16 (77.67–90.51)	<0.0001

Performance comparison of Random Forest model with Ridge feature selection, Logistic regression model with Bayesian feature selection, and traditional models. AUC area under the curve, CI confidence interval.

Regarding calibration (Fig. 6b), the Random Forest model exhibited robust reliability, aligning closely with the ideal diagonal line over most probability ranges and demonstrating minimal underestimation at higher probability levels. While this pattern was consistent with pre-PCI findings, the Logistic Regression model's calibration in the post-PCI scenario revealed more significant deviations, overestimating risk across all levels of observed probabilities.

The PR curves (Fig. 6c) further underscored the Random Forest model's advantage, maintaining superior precision across all levels of recall, particularly in high-recall scenarios. Although this trend echoed the pre-PCI results, the PR performance gap between the models was slightly less pronounced post-PCI.

In summary, the Random Forest model with Ridge feature selection consistently performed better in predicting post-PCI risk. However, the reduced AUC performance gap relative to the Logistic Regression model indicates that the enriched post-PCI feature set enhances predictive performance across a broader range of model complexities.

SHapley additive exPlanations and causal random forest analysis. These analyses unveiled critical predictors and their roles in understanding patient outcomes. Compared to the pre-PCI analysis, the post-PCI findings highlight the additional influence of procedural factors alongside baseline clinical characteristics.

The SHAP summary plot (Fig. 7a) demonstrates that the Killip class, the CFS, and glucose consistently ranked among the most influential predictors in pre- and post-PCI models. However, the post-PCI model further underscored the significance of post-procedural variables, including LVEF, post-PCI TIMI blood flow, and the time from symptom onset to wire crossing.

Dependence plots (Fig. 8) reveal nuanced relationships, such as the linear increase in risk associated with higher Killip class and CFS scores, the non-linear associations of glucose and creatinine levels with outcomes, and a U-shaped relationship for systolic blood pressure, where deviations from an optimal range elevated risk. Delays in intervention were strongly linked to poorer outcomes, while high post-PCI TIMI blood flow correlated with improved survival rates.

Causal effect analysis (Fig. 7b) reinforced these findings, with the Killip class exhibiting the most significant causal impact on mortality. Other variables, such as the CFS, glucose, and creatinine, also displayed notable causal contributions. Procedural metrics, particularly post-PCI TIMI blood flow, further highlighted the critical role of successful intervention in shaping patient outcomes.

Both pre- and post-PCI models consistently identified the Killip class, the CFS, and glucose as pivotal predictors, underscoring their central role in mortality prediction. However, the post-PCI model expanded the pre-PCI framework by integrating procedural features, offering a more nuanced and comprehensive understanding of post-intervention risk. This enhanced perspective bridges baseline clinical factors with the impact of procedural success, providing valuable insights into optimizing post-PCI patient care.

Discussion

ML models of the STEMI-OP project have shown significant promise in predicting in-hospital mortality among older patients undergoing primary PCI, consistently surpassing traditional risk-scoring systems across pre- and post-PCI phases. The STEMI-OP pre-PCI model demonstrated exceptional predictive accuracy using only clinical and basic laboratory data available at admission. These advancements are particularly significant in resource-constrained settings like Vietnam, where limited health insurance coverage often leaves patients facing substantial financial burdens for PCI procedures. The enhanced predictive accuracy of pre-PCI models supports more informed decision-making regarding treatment strategies and financial planning for patients and their families. The STEMI-OP post-PCI model demonstrated even greater predictive power by incorporating procedural and post-intervention data. These findings underscore the critical

Fig. 5 | Post-PCI model. Feature Importance of Random Forest model after applying Ridge regression method for feature selection. PCI percutaneous coronary intervention.

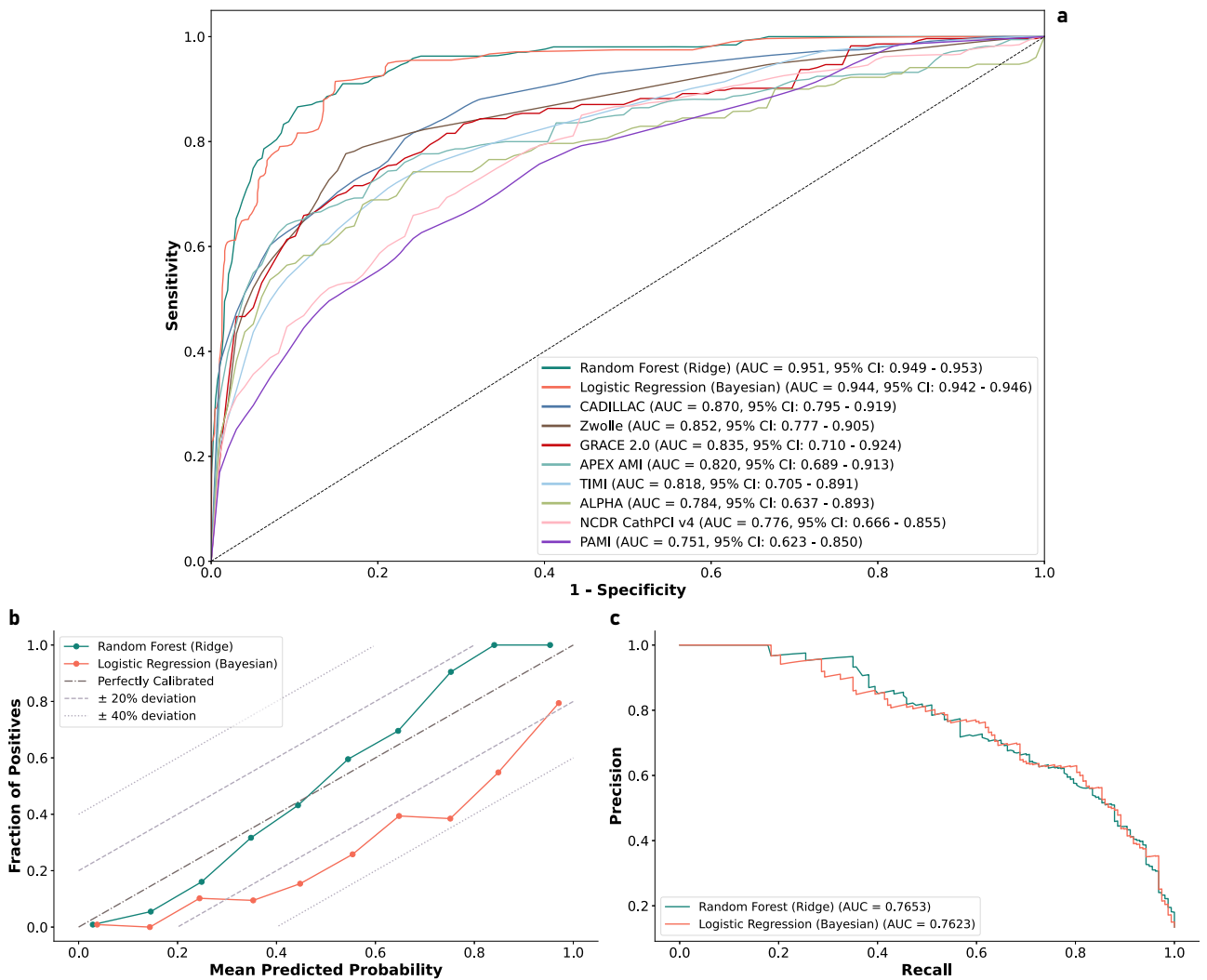
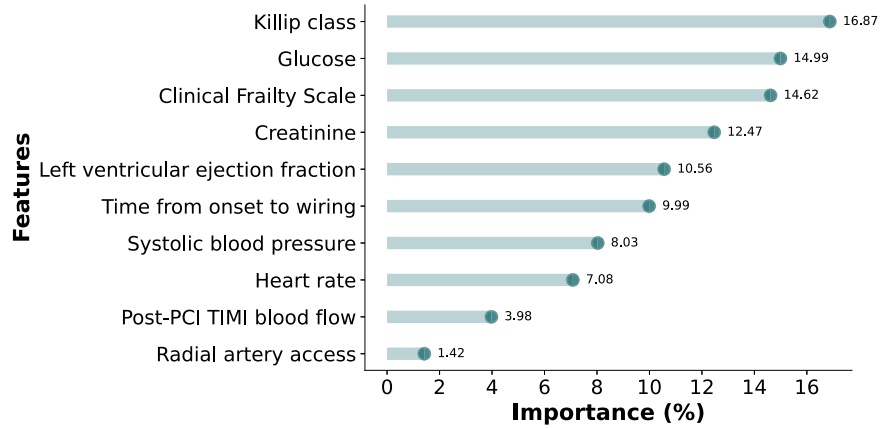


Fig. 6 | Post-PCI model. Random Forest model with Ridge feature selection versus Logistic Regression model with Bayesian feature selection versus traditional models. **a** Receiver operating characteristic (ROC) curves, **b** calibration plots, **c** precision-recall curves. AUC area under the curve, CI confidence interval, PR precision recall.

importance of procedural outcomes in refining mortality risk assessments and optimizing patient care.

Previous studies have consistently demonstrated the superior performance of ML models over traditional risk scores. In Malaysia, ML models outperformed the TIMI score, achieving AUCs of 89% in the general population and 93% in females with STEMI, compared to

TIMI's 81%^{8,9}. Similarly, in Korea, the XGBoost model achieved an AUC of 93.8%, surpassing TIMI (86.6%) and GRACE (92.1%)¹⁰. In Israel, a full-feature Random Forest model reached an AUC of 80.4%, significantly exceeding the GRACE score of 76.4%, while a simple Random Forest model achieved an AUC of 78.7%, comparable to GRACE¹¹.

Fig. 7 | Post-PCI model. SHAP summary plot and causal effect of Random Forest model with Ridge feature selection. **a** SHAP summary plot, **b** causal effect. PCI percutaneous coronary intervention.



Feature Importance, SHAP values, and CRF offer complementary insights into the predictive modeling process, each addressing distinct yet interconnected aspects of feature interpretation and application. Feature Importance provides an initial perspective on the contribution of each feature to the model after the feature selection process is completed; however, it does not account for interactions between variables, nor can it specify the directionality or elucidate the underlying mechanisms of these contributions. In contrast, SHAP leverages game theory to elucidate the magnitude and direction of each feature’s impact on individual predictions and account for interactions between variables, thereby offering nuanced interpretability at the level of individual patients or cases. As a result, SHAP provides a more detailed and comprehensive understanding of each variable’s influence within the model compared to the Feature Importance method. However, the overall ranking and contribution levels of features remained largely consistent between the two methods in both the pre-PCI and post-PCI models. Notable shifts in feature order predominantly occurred among variables with similar levels of importance—for instance, the interchange in rank between the CFS and blood glucose in the pre-PCI model or between left ventricular ejection fraction and the time from symptom onset to wire crossing in the post-PCI model. This demonstrates the stability of feature contributions, whether assessed in isolation or with interactions between variables taken into account.

However, neither Feature Importance nor SHAP alone can reliably infer causality, which can only be thoroughly established through randomized controlled trials. To address this gap in the context of real-world settings, CRF explicitly estimates conditional causal effects, enabling practitioners to assess how deliberate interventions on certain features can affect clinical outcomes or prognoses. Specifically, a variable may demonstrate

only a limited association with the outcome at a given cross-sectional point—for instance, post-PCI TIMI blood flow, which had the lowest SHAP impact in the post-PCI model—yet holds the potential to produce the most substantial change in the outcome if effectively modified through intervention, as indicated by its large effect size in the CRF analysis. This highlights that SHAP and CRF provide distinct yet complementary insights into the influence of individual features on the model, each contributing a unique perspective to the interpretation process.

Determining the optimal number of variables for inclusion in a predictive model requires a careful trade-off between maximising performance and ensuring clinical practicality. While models with fewer variables are inherently more straightforward and user-friendly in real-world settings, they often fall short in predictive power. Traditional risk scores, for instance, typically rely on only 6–8 variables, which contributes to their ease of use but also partly explains their modest performance, with AUCs seldom exceeding 0.9—both in their original development studies and their external validations^{3–7,26–28}. On the other end of the spectrum, models enriched with numerous features generally achieve higher accuracy but at the cost of complexity that undermines their clinical usability. More critically, such models are highly susceptible to overfitting, limiting their generalization ability across various populations. Recent ML models serve as illustrative examples for this pattern, as they often include more than ten predictors and consistently outperform traditional risk scores regarding predictive accuracy^{8–14,29}. However, this increase in complexity comes with trade-offs. The sheer number of features further compounds the inherent complexity of ML models, making them even less interpretable and more challenging to integrate into routine clinical practice. Although the STEMI-OP project acknowledges the limitation of arbitrarily selecting ten features for both the

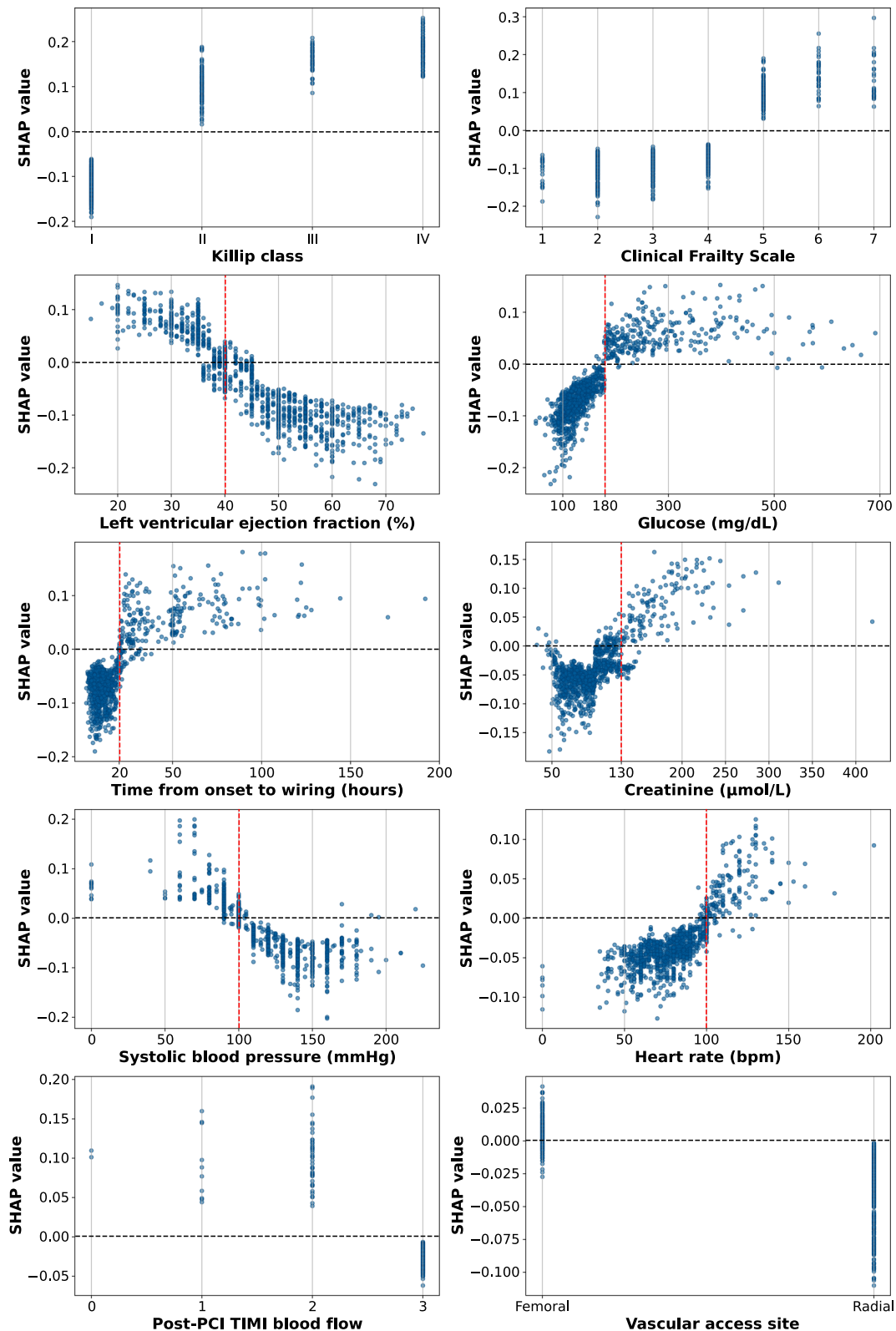


Fig. 8 | Post-PCI model. SHAP dependence plot of the 10 features of the Random Forest model with Ridge feature selection. bpm beat per minute, PCI percutaneous coronary intervention.

pre- and post-PCI models, this decision was intentional. The primary objective was to balance enhancing model performance over traditional risk scores and maintaining simplicity, interpretability, and clinical feasibility. Moreover, by constraining the number of features, the model design aims to minimize the risk of overfitting, thereby increasing the likelihood that the

models will preserve acceptable performance during external validation across diverse patient populations in future applications.

The STEMI-OP project represented the first known attempt to develop ML models specifically to predict in-hospital mortality in older patients following primary PCI. Both conventional models and recent ML

approaches have been developed and validated on data from the general population, with a relatively low representation of older patients^{3–14,26–29}. In addition to key features such as the Killip class, blood pressure, heart rate, LVEF, glucose, and creatinine levels, age consistently emerged in both traditional and ML models as one of the most critical predictors influencing model performance^{8–14,29}. However, there is increasing emphasis on geriatric assessments, driven by the aging population and evidence supporting the prognostic significance of geriatric syndromes, particularly frailty, in predicting cardiovascular outcomes in older adults^{16–19,30}. Recently, with deeper insights into geriatrics, frailty, as a marker of biological age, proved to be a more accurate predictor of poor outcomes, underscoring the greater importance of frailty over chronological age in risk assessment for older patients^{16–19}. Frailty is defined by a breakdown in the coordination of multiple physiological systems, disrupting the body's ability to maintain dynamic homeostasis, depleting its physiological reserves, and rendering it more vulnerable to illness and death³¹. It often manifests as an impaired or dysfunctional response to stressors, such as acute events like myocardial infarction, setting off a damaging cycle of functional decline. This cascade of decline amplifies the harmful effects of acute events, escalating the risk of severe adverse outcomes and mortality. In both the STEMI-OP pre- and post-PCI models, age, historically a strong predictor in traditional and ML models^{3–7,26–28}, lost much of its predictive power when frailty, measured by the CFS, was introduced. The CFS, following the Killip class, consistently ranked as the second most significant predictor in the majority of pre- and post-PCI ML models.

Core clinical variables—such as Killip classification, systolic blood pressure, heart rate, duration of ischemia, and serum creatinine—have long stood as pillars in risk stratification, consistently featured across both traditional scoring systems and more recent ML models^{3–14,26–29}. In our study, these variables remained among the most influential predictors of outcomes in both top-performing pre- and post-PCI models. Among them, the Killip class stands out as a particularly potent prognostic indicator, persistently included in virtually all prior models. Interestingly, in addition to frailty, the STEMI-OP models also incorporated several more predictors that were either underrepresented or entirely absent in conventional risk scores^{3–7,26–28}. For example, hemoglobin in the pre-PCI model or post-PCI TIMI blood flow in the post-PCI model—seen only in the CADILLAC score⁵—or radial artery access in the post-PCI model—exclusive to the ALPHA score⁶—are notable additions. This discrepancy may partly be attributed to the more constrained variable selection strategies employed in traditional models, which often relied on a limited set of 6–8 predictors and largely overlooked procedure-specific features related to PCI. Consequently, these models may have inadvertently omitted features that, although individually weak predictors, still contributed meaningfully to overall model performance. As a result, conventional risk scores may have fallen short in capturing the full complexity of patient outcomes, potentially explaining their relatively limited predictive performance. On the other hand, modern ML-based approaches embrace a broader spectrum of features, enabling the inclusion of nuanced clinical and procedural variables and, in turn, achieving substantial gains in performance. By integrating such predictors—despite their relatively modest individual impact—into both pre- and post-PCI models, the STEMI-OP models enriched the predictive scope of traditional risk scores without introducing excessive complexity, offering a balanced yet more powerful alternative.

A notable strength of the STEMI-OP project lay in its prospective data collection approach, which allowed for the proactive collection of all feature data across individual patients, significantly reducing missing data. Most prior studies developing ML models have utilized retrospective data, where certain features were excluded due to a high proportion of missing values^{8–14}. The subsequent use of imputation algorithms in these cases resulted in datasets that did not fully express the true nature of the features. Additionally, evaluating frailty using the CFS is impractical when relying on retrospective medical records. Frailty assessment has only gained attention in recent years, and as such, older medical records lacked any documented frailty information.

In summary, the STEMI-OP models, with the combination of advanced ML algorithms, robust feature selection methods, SHAP and CRF for model interpretability, the novel inclusion of frailty as a feature, and the use of prospective data, have proven to be a highly effective strategy for predicting in-hospital mortality in older patients undergoing primary PCI. These models not only outperformed traditional risk scores but also offered the necessary transparency for clinical adoption, marking a significant advancement in the application of artificial intelligence in geriatric healthcare.

Despite the innovations in approach and implementation, the STEMI-OP project had several limitations. First, constraints in resources and time allocated to this study have limited the advancement of our models into fully functional platforms designed for direct engagement with clinicians, such as interactive visualization tools or mortality risk assessment applications. Second, the STEMI-OP project was limited to several tertiary cardiovascular centers in southern Vietnam, which may restrict the generalizability of the models to the broader older population in the country. While employing Leave-One-Center-Out Cross-Validation as a rigorous internal validation technique, the absence of external validation remains a limitation. Validating these STEMI-OP models using independent datasets from different geographic regions and healthcare settings would ensure their generalizability. Prospective validation in real-world clinical environments would further strengthen the models' credibility by assessing their performance in real-time decision-making scenarios. Expanding research to encompass diverse populations and performing external validation studies across various regions of Vietnam and Southeast Asia would enhance the applicability and robustness of the findings, increasing their relevance to a broader demographic. Third, although the STEMI-OP project did not implement Local Interpretable Model-agnostic Explanations (LIME), it remains a promising tool for improving interpretability. Future research could explore LIME as part of visualization tools to provide localized insights for each individual prediction, complementing SHAP's global and local feature explanations and further enhancing model transparency and clinical applicability. Another limitation is the shift in the distribution of P2Y12 inhibitor usage following guideline updates, which may have contributed to differences in mortality between the two timelines, potentially introducing bias in model development and interpretation. However, during model development, the P2Y12 inhibitor type was incorporated into the feature selection process to evaluate its influence on the model. This represents an attempt, to some extent, to mitigate the potential bias associated with this issue. Lastly, frailty assessment in older patients remains under-prioritized in Vietnam, creating substantial barriers to the application of these models in clinical practice.

The implementation of ML models in real-world clinical settings requires addressing critical challenges such as scalability, cost-effectiveness, and compatibility with existing healthcare systems. To maximize the clinical utility of the STEMI-OP models, future research should prioritize creating advanced visualization tools that translate complex predictions into clear, actionable insights for clinicians. For instance, interactive dashboards combining SHAP plots with causal effect summaries could enable practitioners to delve into how key predictors shape individual risk profiles. Such tools would act as a critical interface, bridging the gap between sophisticated ML algorithms and their practical clinical application while fostering trust and enhancing usability. The STEMI-OP models could be implemented either as an exclusive application on smart devices or as components of clinical decision support systems embedded within electronic health record platforms. Such integration facilitates real-time access to patient data, enabling the generation of instant mortality risk prediction that can be used to inform treatment strategies.

An increasing body of evidence suggests the potential to prevent and reverse frailty through several straightforward interventions based on physical activity (e.g., multicomponent exercise incorporating resistance training) or nutritional strategies (e.g., the Mediterranean diet)^{32,33}. These approaches simultaneously represent lifestyle modifications associated with improved outcomes for patients undergoing

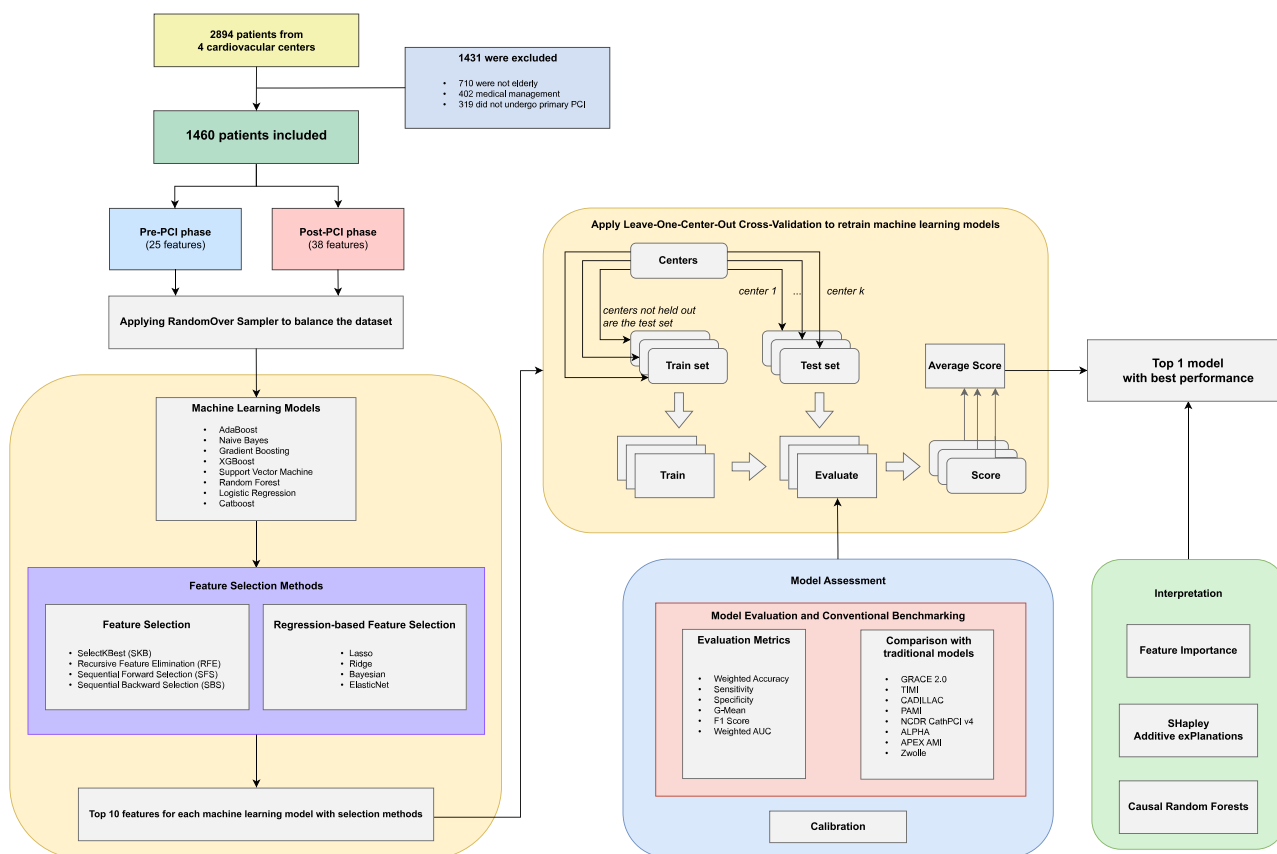


Fig. 9 | The STEMI-OP project workflow and applied methodology. AUC area under the curve, PCI percutaneous coronary intervention, STEMI ST-segment elevation myocardial infarction.

primary PCI. Consequently, reversing or decelerating frailty progression may contribute significantly to improving the prognosis of older patients after primary PCI. Our proposed prognostic model may serve as a critical reminder to clinicians of a novel prognostic geriatric-specific factor that significantly impacts mortality and holds the potential to be preventable and even reversible. This emphasizes the necessity of early identification and timely intervention to address this factor, thereby advancing healthcare quality for the aging population.

Methods

Study design and patient cohorts

The STEMI-OP project included 1460 older patients diagnosed with STEMI who underwent primary PCI. Among them, 426 patients were prospectively enrolled from two national cardiovascular centers in southern Vietnam between February 2017 and April 2018, while an additional 1034 patients were recruited prospectively from two other cardiovascular centers in the same region between February 2021 and May 2024.

The study included individuals aged 60 years or older, aligning with the definition of older adults established by the United Nations³⁴. This definition is uniformly recognized across all ASEAN member states³⁵.

STEMI was defined in line with the 2023 European Society of Cardiology guideline for acute coronary syndrome management³⁰. Patients were excluded if they had: (a) age <60; (b) suspected STEMI but did not undergo coronary angiography; or (c) culprit lesions but did not undergo primary PCI.

All primary PCI procedures and medications followed standard techniques and adhered to the 2023 European Society of Cardiology guideline for acute coronary syndrome management³⁰. The study received approval from the local ethics committee of the University of Medicine and Pharmacy at Ho Chi Minh City.

The STEMI-OP project workflow is illustrated in Fig. 9, detailing each phase, from data collection to the development of ML models for predicting in-hospital mortality.

Data collection

Financial barriers in developing countries like Vietnam consistently limit access to advanced medical procedures such as primary PCI. Health insurance typically provides insufficient coverage for high-cost, specialized treatments, leaving patients with significant financial burdens for life-saving interventions. Consequently, patients and their families face difficult financial decisions under time pressure, underscoring the need for the STEMI-OP project to develop reliable pre-PCI predictive models to assess in-hospital mortality and inform treatment decisions.

To address this issue, pre-PCI predictive models were developed using basic admission data, including clinical characteristics, comorbidities, and fundamental laboratory tests. These models aim to predict mortality risk before PCI, providing critical information for decision-making. Additionally, the STEMI-OP project developed post-PCI models to improve predictive accuracy by incorporating more detailed data, including serum high-sensitivity troponin levels, echocardiographic findings, and PCI procedural characteristics. Comprehensive data on demographic information, comorbidities, and clinical presentation were thoroughly documented upon admission.

Frailty, a key geriatric-specific predictor of adverse outcomes, was assessed using the CFS version 2.0 at admission¹⁵. The CFS is a nine-tier framework grounded in the cumulative deficit frailty model. It provides a concise clinical representation of biological age by incorporating key factors such as illness severity, comorbidities, functional impairments, and cognitive deficits¹⁵. Notably, the CFS has been recognized as a highly effective tool for evaluating frailty in patients with acute coronary syndrome (ACS), offering robust predictive validity and ease of use in clinical practice^{36,37}.

Frailty levels were assessed using the CFS through direct patient interviews or consultations with their relatives. The evaluation targeted the patient's condition two weeks before hospital admission to minimize the influence of hospitalization-related health deterioration. Unlike other frailty assessments that require domain-specific instruments, the CFS relies exclusively on questionnaire data derived from recollections of the patient's basic and instrumental activities of daily living. This retrospective approach was deemed feasible, as patients were presumed to have stable health statuses before the onset of STEMI.

Trained geriatric resident physicians conducted the CFS evaluations, documenting their findings independently in medical records. To ensure objectivity, the collected frailty data were blinded to the principal investigators, who subsequently gathered clinical outcomes. The questionnaire was administered directly to patients when possible, as this method yielded the most reliable data. However, for patients unable to communicate—such as those in cardiogenic shock or requiring intubation—information was obtained from relatives. If patients later regained communicative ability during their hospital stay, any data previously collected from relatives were superseded by new data directly provided by the patients.

Electrocardiographic features and essential laboratory tests were recorded along with these clinical features. Information regarding medications that could influence PCI outcomes and mortality, such as aspirin, types of P2Y12 inhibitors, and statins, was also documented before the procedure. Post-PCI data, including procedural details, echocardiographic findings, and troponin levels, were then gathered further to improve the predictive power of the pre-PCI models. The pre- and post-PCI model development features are outlined in Supplementary Table S7.

The STEMI-OP final ML models were evaluated by comparing their performance to eight established prediction models, including GRACE 2.0, TIMI, Zwolle, CADILLAC, PAMI, NCDR CathPCI v4, ALPHA, and APEX AMI, with relevant features from these models also being collected^{3-7,26-28}.

Data imbalance management

To address the issue of class imbalance in the dataset, the RandomOverSampler (ROS) technique is employed³⁸. This method operates by increasing the representation of minority classes in the data through random duplication of samples. By replicating instances of the underrepresented class, the process ensures a more balanced distribution of classes, which is crucial for improving the performance and fairness of classification models. ROS is particularly effective in scenarios where the imbalance could lead to biased predictions, allowing ML algorithms to learn equally from all classes. This strategy enhances the model's generalization ability by mitigating the bias introduced by imbalanced data, improving predictive accuracy across all categories. By applying ROS solely to the training set, we tried to minimize the risk of biased predictions on the test set and improved the model's ability to generalize across all classes and to the real-world population, thereby enhancing its overall predictive accuracy.

ML models

The STEMI-OP project selected several ML algorithms for their robustness in classification tasks, particularly in healthcare predictions. The models included AdaBoost³⁹, Naive Bayes⁴⁰, Gradient Boosting⁴¹, XGBoost⁴¹, Support Vector Machine⁴², Random Forest⁴³, Logistic Regression⁴⁴, and Catboost⁴⁵. AdaBoost and Gradient Boosting aim to iteratively reduce errors, while Naive Bayes offers a probabilistic approach well-suited for high-dimensional data. Random Forest and XGBoost excel in managing large and complex datasets with built-in feature importance mechanisms. SVM was included due to its capability to handle high-dimensional spaces, and Logistic Regression was chosen for its simplicity and extensive use in medical research. CatBoost was selected for its ability to handle categorical features effectively and mitigate target leakage, making it particularly robust in structured data tasks common in healthcare.

Leave-one-center-out cross-validation

Leave-one-center-out cross-validation (LOCO-CV) is a robust validation method commonly used in multi-center studies to assess the generalizability and stability of predictive models across heterogeneous datasets⁴⁶. In this study, LOCO-CV was implemented using data collected from four different cardiovascular centers. Each center represented distinct characteristics regarding data collection methods, patient demographics, and treatment protocols. The LOCO-CV procedure involved standardizing data from the centers to minimize differences due to data formatting or collection methods, training the models on data from three centers, and evaluating the performance of the remaining hospital's data. This process was repeated four times, with each center serving as the test set once. The results from each iteration were aggregated to provide a comprehensive evaluation of model performance. This approach not only tested the model's stability across the four centers but also simulated real-world scenarios where a particular center may challenge the model's robustness. Although LOCO-CV can be computationally expensive, the implementation of the data from these four centers was manageable, and the results provided reliable insights into the model's generalizability across diverse clinical environments.

In addition, LOCO-CV was concurrently applied to eight traditional risk scores for a fair comparison. Importantly, the original published coefficients and variables for each traditional risk score were used without refitting on the test fold in each LOCO-CV iteration, treating these scores as externally validated models. This approach mirrors real-world applications, where traditional risk scores are typically used on new populations without retraining. By consistently applying LOCO-CV across all models, we ensured an equal evaluation framework, allowing for fair and comprehensive comparisons between ML models and established clinical scores.

Feature selection and feature importance

The determination of the optimal number of features incorporated into the models was driven by the balance between model performance and clinical applicability. Conventional risk scores such as GRACE, TIMI, and CADILLAC typically incorporate only 6–8 variables^{45,7}, while recent ML models demonstrate that expanding to 10–15 features markedly improves models' predictive accuracy^{8-11,47,48}. While incorporating additional features often improves performance, an excessive number of variables can be impractical in clinical practice and may induce overfitting, thus limiting the model's generalizability to various populations. For these reasons, the STEMI-OP project selected the ten most important features for model development.

Four feature selection techniques were employed, including Recursive Feature Elimination⁴⁹, SelectKBest⁴⁹, Sequential Forward Selection⁵⁰, and Sequential Backward Selection⁵¹, to reduce dimensionality while preserving key features, improving model performance and interpretability.

For models capable of generating feature importance metrics (e.g., Random Forest, Gradient Boosting), the top ten features were retained based on their respective importance scores. In contrast, for models lacking inherent feature importance mechanisms, mutual information was utilized to select the ten features with the highest mutual information values⁵².

To further support model-specific feature selection strategies, regression-based analyses were conducted in parallel to identify the ten features most significantly associated with the outcome. Lasso, Ridge, ElasticNet, and Bayesian regression-based feature selection methods were deliberately chosen for their regularization capabilities, which mitigate overfitting while simultaneously ranking features by importance⁵³. These regression techniques are particularly advantageous in scenarios with multicollinearity or high-dimensional datasets, as they assign weights to features, enabling precise identification of those with the greatest predictive value. Following the selection of their respective sets of ten features by the four regression-based techniques, these variable subsets were subsequently employed in ML models to generate combinations between each ML model and its corresponding feature selection method.

By integrating these regression-based methods and model-specific feature selection strategies, tailored feature sets were created for each ML

model. This holistic approach ensures the inclusion of the most relevant features, bolstering the robustness and generalizability of the feature selection process across diverse models.

Model performance

Model evaluation. The STEMI-OP model performance was evaluated using key metrics, including weighted accuracy, sensitivity, specificity, G-Mean, F1 score, and weighted AUC. For the threshold-dependent metrics (e.g., accuracy, sensitivity, specificity, G-Mean, and F1 score), a probability threshold of 0.5 was used to dichotomize predicted probabilities into class labels for metric calculation. AUC is considered a more robust metric than accuracy, especially in imbalanced datasets. Accuracy, while commonly used, can be misleading when one class dominates. It measures the overall correctness of predictions but does not account for the probability estimates of classifiers. In contrast, AUC provides a more nuanced evaluation by analyzing the trade-off between true positive and false positive rates across varying thresholds, making it especially useful for class-imbalanced datasets or when probability estimates are critical⁵⁴. Similarly, the F1 score, which harmonizes precision and recall, is particularly valuable for imbalanced datasets, as it captures the balance between correctly identifying positive instances and minimizing false positives⁵⁵. Weighted AUC and weighted accuracy further refine performance evaluation by accounting for the class imbalance in the dataset, ensuring that each class contributes proportionally to the overall metric⁵⁶. Consequently, these metrics, particularly weighted AUC, have become preferred tools for ranking and comparing models due to their ability to offer deeper insights into predictive performance.

In our study, these metrics, with weighted AUC serving as the decisive criterion, were used to identify the top-performing ML model and to compare its performance with traditional risk scores. Given that the logistic regression model is commonly preferred in clinical settings due to its simplicity and interpretability—contrasting with the more complex, non-linear nature of ML models—a top-performing logistic regression model was likewise identified. This model was then directly compared to the best ML model for both predictive performance and alignment with observed outcomes.

To further ensure the validity of comparisons, pairwise comparisons of AUC values were performed using DeLong's test, with the top-performing ML model serving as the reference. This non-parametric approach accounts for the correlation between predictions on the same dataset, offering a rigorous statistical evaluation of differences in model performance.

Model calibration. Calibration in ML ensures that predicted probabilities closely reflect actual outcome probabilities⁵⁷. Calibration plots are commonly used to compare predicted probabilities with observed class probabilities. Ideally, a well-calibrated model will produce a 45-degree diagonal line on this plot, indicating that predicted probabilities align with empirical outcomes. This process is crucial for real-world applications where both model accuracy and the reliability of probability estimates are essential. Calibration is necessary for probabilistic classifiers, where miscalibrated predictions could affect decision-making and risk assessments.

In our study, calibration plots were employed to benchmark the predictive accuracy of the top-performing ML model against the best logistic regression model, allowing us to determine which model best approximated the actual mortality outcomes.

SHapley additive exPlanations

Although conventional feature importance metrics offer some understanding of how individual features affect a model's output, they frequently fail to capture feature interactions or account for the specific contribution of each feature to a particular prediction. To overcome this limitation, SHapley Additive exPlanations was employed, providing both global and individual insights into feature contributions of the STEMI-OP models⁵⁸. In our study, SHAP was used on the top-performing ML model of both the pre- and post-

PCI phases to quantify the contribution level of each predictor. The SHAP summary plot uses Shapley values to give a comprehensive view of features' impact across the entire dataset. It visualizes the distribution of Shapley values for each variable, showing the intensity and direction of each feature's impact on the model's predictions. This plot helps identify the most influential features and how their contributions vary across different samples. On the other hand, the SHAP dependence plot focuses on the relationship between a single predictor and the model's output. It uses Shapley values to illustrate how the value of a specific feature affects the prediction while also considering interactions with other variables. This plot allows us to observe how the impact of a feature changes with its value and how it interacts with other predictors in the model.

Causal Random Forests

To explore causal relationships between features and in-hospital mortality of the STEMI-OP models, Causal Random Forests were applied to the top-performing ML model for both pre- and post-PCI phases⁵⁵. The CRF model estimates the Conditional Average Treatment Effect, which quantifies the effect of a feature on the outcome for each individual in the dataset. Unlike standard random forests, which prioritize prediction accuracy, causal random forests estimate heterogeneous treatment effects. This allows for the identification of how specific features produce variations in outcomes while accounting for interactions of covariates. In essence, this method helps pinpoint which features have the most significant causal effects on the outcome and how those effects may differ based on other factors in the dataset.

Statistics

Continuous variables were summarized either as means with standard deviations or as medians accompanied by interquartile ranges (IQR: 25%–75%). Categorical variables were reported as counts and percentages. To compare categorical variables, the Chi-square test or Fisher's exact test was applied, while continuous variables were evaluated using either the Student's *t* test or the Mann–Whitney *U* test, depending on the normality of the data distribution.

In the dataset, certain features exhibited a missing data rate of 1%–1.5%, including LVEF in 23 cases (1.6%), creatinine in 17 cases (1.2%), high-sensitivity troponin T in 24 cases (1.6%), and smoking history in 21 cases (1.4%). Missing data were assumed to be completely at random. To handle these missing values, a multiple imputation approach was performed using the 'mice' package in R⁵⁹.

Data availability

The data that support the findings of the STEMI-OP project are available from Pham Ngoc Thach University of Medicine, but restrictions apply to the availability of these data, which were used under license for the current research and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of Pham Ngoc Thach University of Medicine.

Code availability

The underlying code for this study [and the dataset] is not publicly available but may be made available to qualified researchers at a reasonable request from the corresponding author.

Received: 8 October 2024; Accepted: 12 May 2025;

Published online: 06 June 2025

References

- Riehle, L. et al. Implementation of the ESC STEMI guidelines in female and elderly patients over a 20-year period in a large German registry. *Clin. Res. Cardiol.* **112**, 1240–1251 (2023).
- Tong, J. et al. Clinical outcomes of elderly South-East Asian patients in primary percutaneous coronary intervention for ST-elevation myocardial infarction. *J. Geriatr. Cardiol.* **13**, 830–835 (2016).

3. Addala, S. et al. Predicting mortality in patients with ST-elevation myocardial infarction treated with primary percutaneous coronary intervention (PAMI risk score). *Am. J. Cardiol.* **93**, 629–632 (2004).
4. Fox, K. A. et al. Should patients with acute coronary disease be stratified for management according to their risk? Derivation, external validation and outcomes using the updated GRACE risk score. *BMJ Open* **4**, e004425 (2014).
5. Halkin, A. et al. Prediction of mortality after primary percutaneous coronary intervention for acute myocardial infarction: the CADILLAC risk score. *J. Am. Coll. Cardiol.* **45**, 1397–1405 (2005).
6. Hizoh, I. et al. A novel risk model including vascular access site for predicting 30-day mortality after primary PCI: The ALPHA score. *Cardiovasc Revasc. Med.* **18**, 33–39 (2017).
7. Morrow, D. A. et al. TIMI risk score for ST-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation: an intravenous nPA for treatment of infarcting myocardium early II trial substudy. *Circulation* **102**, 2031–2037 (2000).
8. Kasim, S. et al. Ensemble machine learning for predicting in-hospital mortality in Asian women with ST-elevation myocardial infarction (STEMI). *Sci. Rep.* **14**, 12378 (2024).
9. Aziz, F. et al. Short- and long-term mortality prediction after an acute ST-elevation myocardial infarction (STEMI) in Asians: a machine learning approach. *PLoS ONE* **16**, e0254894 (2021).
10. Lee, W. et al. Machine learning enhances the performance of short and long-term mortality prediction model in non-ST-segment elevation myocardial infarction. *Sci. Rep.* **11**, 12886 (2021).
11. Hadanny, A. et al. Predicting 30-day mortality after ST elevation myocardial infarction: machine learning- based random forest and its external validation using two independent nationwide datasets. *J. Cardiol.* **78**, 439–446 (2021).
12. Zhao, J., Zhao, P., Li, C. & Hou, Y. Optimized machine learning models to predict in-hospital mortality for patients with ST-segment elevation myocardial infarction. *Ther. Clin. Risk Manag.* **17**, 951–961 (2021).
13. Deng, L. et al. Machine learning to predict no reflow and in-hospital mortality in patients with ST-segment elevation myocardial infarction that underwent primary percutaneous coronary intervention. *BMC Med. Inform. Decis. Mak.* **22**, 109 (2022).
14. Li, Y. Diagnostic model of in-hospital mortality in patients with acute ST-segment elevation myocardial infarction used artificial intelligence methods. *Cardiol. Res. Pract.* **2022**, 8758617 (2022).
15. Rockwood, K. & Theou, O. Using the clinical frailty scale in allocating scarce health care resources. *Can. Geriatr. J.* **23**, 210–215 (2020).
16. Patel, A. et al. Frailty and outcomes after myocardial infarction: insights from the CONCORDANCE registry. *J. Am. Heart Assoc.* **7**, e009859 (2018).
17. Ekerstad, N. et al. Clinical Frailty Scale classes are independently associated with 6-month mortality for patients after acute myocardial infarction. *Eur. Heart J. Acute Cardiovasc. Care* **11**, 89–98 (2022).
18. Yoshioka, N. et al. Impact of the clinical frailty scale on mid-term mortality in patients with ST-elevated myocardial infarction. *Int J. Cardiol. Heart Vasc.* **22**, 192–198 (2019).
19. Hosseini, F. et al. Association of frailty with in-hospital and long-term outcomes among STEMI patients receiving primary percutaneous coronary intervention. *CJC Open* **6**, 1004–1012 (2024).
20. Obermeyer, Z. & Emanuel Ezekiel, J. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219.
21. Zhou, Z.-H. *Ensemble methods: foundations and algorithms*. 1st edn. (Chapman and Hall/CRC, 2012).
22. Kalule, R., Abderrahmane, H. A., Alameri, W. & Sassi, M. Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. *Sci. Rep.* **13**, 9855 (2023).
23. Friedman, J. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29** (2000).
24. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *Arxiv* <https://arxiv.org/abs/1802.03888> (2018).
25. Stefan Wager, S. A. Estimation and inference of heterogeneous treatment effects using random forests. *Arxiv* <https://doi.org/10.48550/arXiv.1510.04342> (2017).
26. De Luca, G. et al. Prognostic assessment of patients with acute myocardial infarction treated with primary angioplasty: implications for early discharge. *Circulation* **109**, 2737–2743 (2004).
27. Stebbins, A. et al. A model for predicting mortality in acute ST-segment elevation myocardial infarction treated with primary percutaneous coronary intervention: results from the Assessment of Pexelizumab in Acute Myocardial Infarction Trial. *Circ. Cardiovasc. Inter.* **3**, 414–422 (2010).
28. Brennan, J. M. et al. Enhanced mortality risk prediction with a focus on high-risk percutaneous coronary intervention: results from 1,208,137 procedures in the NCDR (National Cardiovascular Data Registry). *JACC Cardiovasc. Inter.* **6**, 790–799 (2013).
29. Shetty, M. K. et al. Machine learning based model for risk prediction after ST-elevation myocardial infarction: insights from the North India ST elevation myocardial infarction (NORIN-STEMI) registry. *Int. J. Cardiol.* **362**, 6–13 (2022).
30. Byrne, R. A. et al. 2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of acute coronary syndromes of the European Society of Cardiology (ESC). *Eur. Heart J.* **44**, 3720–3826 (2023).
31. Fried, L. P. et al. From bedside to bench: research agenda for frailty. *Sci. Aging Knowl. Environ.* **2005**, pe24 (2005).
32. Brennan, T. H., Lewis, L. K., Gordon, S. J. & Prichard, I. Effectiveness of interventions to prevent or reverse pre-frailty and frailty in middle-aged community dwelling adults: a systematic review. *Prev. Med.* **185**, 108008 (2024).
33. Eidam, A. et al. Interventions to prevent the onset of frailty in adults aged 60 and older (PRAE-Frail): a systematic review and network meta-analysis. *Eur. Geriatr. Med.* **15**, 1169–1185 (2024).
34. Older persons. <https://emergency.unhcr.org/protection/persons-risk/older-persons> (2024).
35. Secretariat, A. Syntax of referencing. In *Old Age Poverty and Active Ageing in ASEAN: Trends and Opportunities 7–37* (ASEAN Secretariat, 2023).
36. Alonso et al. The role of frailty in acute coronary syndromes in the elderly. *Gerontology* **64**, 422–429 (2018).
37. Theou, O. et al. What do we know about frailty in the acute care setting? A scoping review. *BMC Geriatr.* **18**, 139 (2018).
38. Hayati, M., Mutmainah, S. & Ghufuran, S. Random and synthetic over-sampling approach to resolve data imbalance in classification. *Int. J. Artif. Intell. Res.* **4**, 86 (2021).
39. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
40. Beja-Battais, P. Overview of AdaBoost : Reconciling its views to better understand its dynamics. *Arxiv* <https://arxiv.org/abs/2310.18323> (2023).
41. Jozdani, S. E., Johnson, B. A. & Chen, D. Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. *Remote Sens.* **11**, 1713 (2019).
42. Tong, S. & Koller, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2**, 45–66 (2001).
43. Al Amrani, Y., Lazaar, M. & El Kadiri, K. E. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Comput. Sci.* **127**, 511–520 (2018).
44. Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **96**, 3–14 (2002).

45. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: unbiased boosting with categorical features. <https://doi.org/10.48550/arXiv.1706.09516> (2019).
46. Berrar, D. Syntax of referencing. In *Encyclopedia of Bioinformatics and Computational Biology* (ed. Ranganathan, S., Nakai, K., & Schonbach, C.) 542–545 (Academic Press, 2019).
47. Shouval, R. et al. Machine learning for prediction of 30-day mortality after ST elevation myocardial infarction: an acute coronary syndrome Israeli survey data mining study. *Int. J. Cardiol.* **246**, 7–13 (2017).
48. Shakhgeldyan, K. I., Kuksin, N. S., Domzhalov, I. G., Rublev, V. Y. & Geltser, B. I. Interpretable machine learning for in-hospital mortality risk prediction in patients with ST-elevation myocardial infarction after percutaneous coronary interventions. *Comput. Biol. Med.* **170**, 107953 (2024).
49. Ferri, F. J., Pudil, P., Hatef, M. & Kittler, J. Syntax of referencing. In *Machine Intelligence and Pattern Recognition* (ed. Edzard S. Gelsema & Laveen S. Kanal) 403–413 (North-Holland, 1994).
50. Liu, W. & Wang, J. Recursive elimination–election algorithms for wrapper feature selection. *Appl. Soft. Comput.* **113**, 107956 (2021).
51. Darst, B. F., Malecki, K. C. & Engelman, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **19**, 1–6 (2018).
52. Krier, C., François, D., Wertz, V. & Verleysen, M. Feature scoring by mutual information for classification of mass spectra. <http://hdl.handle.net/2078.1/90752> (2006).
53. Trevor Hastie, R. T., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, 2009).
54. Ling, C. X., Huang, J. & Zhang, H. Syntax of referencing. In *Advances in Artificial Intelligence* (ed. Xiang, Y., & Brahim, C. D) 329–341 (Springer, 2017).
55. Juba, B. & Le, H. S. Precision-recall versus accuracy and the role of large data sets. *Proc. AAAI Conf. Artif. Intell.* **33**, 4039–4048 (2019).
56. Dong, Y., Li, X., Li, J. & Zhao, H. Analysis on weighted AUC for imbalanced data learning through isometrics. *J. Comput. Inf. Syst.* **8**, 371–378 (2012).
57. Bella, A., Ferri, C., Hernández-Orallo, J. & Ramírez-Quintana, M. J. Syntax of referencing. In *Machine Learning: Concepts, Methodologies, Tools and Applications* (ed Information Resources Management Association) 32–49 (IGI Global, 2012).
58. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. <https://arxiv.org/abs/1705.07874> (2017).
59. Li, P., Stuart, E. A. & Allison, D. B. Multiple imputation: a flexible tool for handling missing data. *JAMA* **314**, 1966–1967 (2015).

Acknowledgements

The STEMI-OP project was funded by Nguyen Tat Thanh University and Pham Ngoc Thach University of Medicine, Ho Chi Minh City, Vietnam. The funders played no role in the study design, data collection, data analysis, interpretation, or this manuscript's writing.

Author contributions

T.V.N., Q.T.N. and T.D.L. contributed equally to the design of the STEMI-OP project, supervision of clinical aspects, data analysis, technical feedback provision, result interpretation, final editing, and final approval of the manuscript. H.Q.N. was responsible for technical aspects, data analysis, technical result interpretation, and drafting of the method and the result sections. N.T.N. developed ethical protocols, ensured the relevance of the clinical data, offered critical insights, and guided data collection to maintain adherence to clinical and ethical guidelines. K.D.L. and L.H.D.T. drafted the abstract, introduction, and discussion section. T.C.N. and T.H.V. prepared the figures. P.H.L. and P.T.T. prepared the tables and supplementary information. K.D.L., L.H.D.T., T.C.N., T.H.V., P.H.L. and P.T.T. contributed to data collection and verification, analysis, and interpretation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41514-025-00238-9>.

Correspondence and requests for materials should be addressed to Quyen The Nguyen or Thanh Dinh Le.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025, corrected publication 2025