

Development and Validation of a Machine Learning Model for Predicting Pulmonary Embolism Using Routine Clinical Features

Jing Gao

Sir Run Run Shaw Hospital

Cheng Lin

21418078@zju.edu.cn

Sir Run Run Shaw Hospital

Research Article

Keywords: Pulmonary embolism, machine learning, XGBoost, SHAP, clinical prediction model, risk stratification

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11054963/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Pulmonary embolism (PE) is a life-threatening cardiovascular emergency with non-specific presentations, making timely diagnosis challenging. Over-reliance on computed tomography pulmonary angiography (CTPA) has raised concerns about radiation exposure and healthcare costs. Traditional clinical prediction rules, such as the Wells and Geneva scores, have limited specificity and fail to capture complex interactions among risk factors. Machine learning (ML) offers a promising alternative, but many existing models lack interpretability or depend on features not routinely available in emergency settings.

Objective

To develop and internally validate a parsimonious, interpretable ML model for predicting PE using routinely available clinical variables, and to apply SHapley Additive exPlanations (SHAP) to enhance model transparency.

Methods

We conducted a retrospective study of 519 patients suspected of PE at a single tertiary referral center. Eight independent predictors—age, hyperlipidemia, limb circumference, D-dimer, oxygen saturation, dyspnea, immobility, and malignancy—were identified via univariate and multivariable logistic regression on the training set. Three ML algorithms (Lasso, Random Forest, and XGBoost) were trained and evaluated on a 70/30 train-test split. Model performance was assessed using AUC-ROC, calibration plots, and Brier score. SHAP analysis was performed for global and local interpretability.

Results

On the testing set, Lasso achieved an AUC of 0.766 (95% CI: 0.691–0.840), while XGBoost showed a comparable AUC of 0.764 (95% CI: 0.690–0.837). XGBoost was selected as the primary model due to its superior interpretability via SHAP. At the optimal threshold of 0.526, XGBoost achieved a sensitivity of 74.4%, specificity of 66.7%, and overall accuracy of 71.0%. Calibration was moderate (Brier score = 0.224 vs. null model 0.247; Hosmer-Lemeshow $P = 0.563$). SHAP analysis identified limb circumference, D-dimer, and malignancy as the top three contributors to PE risk.

Conclusions

The XGBoost model based on eight routine clinical features demonstrated moderate discriminative ability and good interpretability for predicting PE. This tool may serve as a decision-support aid for

emergency physicians in risk stratification, potentially reducing unnecessary CTPA utilization. However, the high PE prevalence (55.6%) in our tertiary cohort limits generalizability; external validation in broader, low-prevalence populations is essential before clinical implementation.

Trial registration

: Not applicable (retrospective observational study).

Background

Worldwide, Pulmonary embolism (PE) stands out prominently among causes of death and disability linked to circulatory diseases. Its yearly occurrence is estimated at 1–2 cases per one thousand members of the general public [1]. The symptomatic profile of this disorder is extremely ambiguous, ranging from accidental detection on scans for other reasons to sudden lethal arrhythmic events, which perpetually complicates prompt correct management within crisis medicine and urgent treatment environments [2]. Failure to identify it, or undue delays, may prove catastrophic; conversely, superfluous diagnostic procedures carry threats of ionizing radiation, contrast-triggered renal impairment, and inflated fiscal outlays [3]. Optimizing the sequence of assessments for individuals with suspected PE represents a vital goal for everyday healthcare delivery.

Currently, Computed Tomography Pulmonary Angiography (CTPA) represents the benchmark radiologic test for establishing a PE diagnosis. Nonetheless, its extensive deployment has generated concerns about overdiagnosis and the squandering of radiological capacities [4]. In an attempt to curtail improper CTPA referrals, standardized risk-prediction tools, exemplified by the Wells score and the updated Geneva rule, have been crafted and are now extensively utilized [5, 6]. These instruments stratify subjects into lower, middle, or upper suspicion bands, aiding practitioners in deciding between performing a D-dimer assay and proceeding straight to radiographic examination. Still, these approaches harbor clear deficiencies: they hinge upon an inflexible assortment of yes/no or categorical items, cannot grasp nuanced nonlinear connections among multiple determinants, and regularly show unsatisfactory specificity in authentic emergency ward populations [7]. In reality, numerous patients still undergo avoidable CTPA scanning, whereas particular high-risk individuals remain overlooked.

Over the past few years, machine learning has arisen as an encouraging strategy for enhancing clinical prognostic models, harnessing vast stores of electronic health records and independently detecting nuanced, multidimensional correlations [8]. Unlike traditional regression techniques, ML approaches including random-forest and XGBoost obviate the prerequisite of preselecting a mathematical equation to represent nonlinear effects and variable cross-talk [9]. Various research projects have applied ML to PE prediction, incorporating diverse clinical, laboratory, and radiographic elements, and have documented discriminative abilities that are on a par with, or superior to, established risk scoring methods [10, 11]. Despite such progress, many existing ML systems depend upon an excessively broad

array of inputs—some of which are infrequently obtainable during standard emergency evaluations—or lack sufficient clarity, thereby obstructing their adoption in clinical workflows [12].

To surmount these hurdles, we designed and internally verified a compact yet robust ML-based PE predictor that uses only eight routinely captured clinical variables, chosen through a disciplined two-stage feature reduction approach (univariate examination followed by multivariable logistic modeling). Three specific algorithms were selected for comparison—Lasso (as a benchmark), random-forest, and XGBoost—and their efficacy was gauged using the AUC, calibration curves, and the Brier index. We additionally deployed the SHAP method to demystify the final algorithm's decision-making, offering both overarching rankings of variable relevance and tailored risk narratives per case. We hypothesized that the XGBoost classifier would yield an AUC greater than 0.75, with SHAP-derived attributions demonstrating clinical coherence. Our overarching goal is to furnish a straightforward, transparent, and potentially actionable support tool to aid emergency room clinicians in triaging suspected PE cases.

Methods

Study Design and Data Source

This investigation employed a retrospective, single-center observational design at Sir Run Run Shaw Hospital, covering the period from January 2020 through December 2025. Protocol approval was obtained from the hospital's Institutional Review Board (IRB; Approval No.: 2026-2131-02). Owing to its retrospective character, the IRB waived the need for written informed consent. All information, drawn from the facility's electronic medical record (EMR) system, underwent complete de-identification prior to any analytic procedures.

Study Population and Eligibility Criteria

Initial recruitment comprised consecutive cases with suspected PE who had undergone either D-dimer assays or CTPA scans during the study window. Three inclusion criteria governed formal enrolment: (1) age ≥ 18 years, (2) availability of complete records for all target variables, and (3) a definitive PE status—ascertained via positive CTPA or excluded through clinical evaluation combined with negative imaging or D-dimer results. Patients missing essential data were removed after preliminary review. The final analytical cohort contained 519 individuals, among whom 289 (55.6%) received a PE diagnosis.

Data Collection and Variable Definition

From the EMR, we extracted demographic features, clinical presentations, comorbid conditions, vital signs, and laboratory measurements. The initial variable pool encompassed sex, age, BMI, diabetes, hypertension, smoking history, hyperlipidemia, limb circumference, D-dimer, C-reactive protein (CRP), peripheral oxygen saturation (SpO_2), heart rate, dyspnea, pain, surgical history, immobilization,

malignancy, alongside the Wells and Geneva scoring tools. Limb circumference was measured in centimeters. A difference of > 3 cm compared to the contralateral limb was defined as "abnormal" (consistent with the Wells score criteria for DVT clinical signs). PE status served as the binary outcome, coded as 1 for a positive CTPA finding and 0 for a negative one. The datasets generated and/or analyzed during the current study are not publicly available due to hospital privacy regulations but are available from the corresponding author on reasonable request.

Feature Selection and Statistical Analysis

To guard against overfitting and isolate genuinely informative predictors, we executed a rigorous two-step selection procedure confined to the training subset. Univariate logistic regression first screened every candidate variable; those yielding $P < 0.1$ advanced to the next phase. A bidirectional stepwise algorithm, driven by the Akaike information criterion, was then applied within a multivariate logistic framework. Eight independent determinants ultimately emerged for subsequent machine-learning modelling: age, hyperlipidemia, limb circumference, D-dimer, oxygen saturation, dyspnea, immobilization, and malignancy. Effect sizes for these retained factors were expressed as odds ratios with 95% confidence intervals. All preliminary statistical computations relied on R version 4.5.3, specifically the glm and MASS packages. This study is reported in accordance with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement.

Machine Learning Model Development

After feature selection, the entire dataset underwent a stratified random split (70% training, 30% testing) to preserve the original PE prevalence across both partitions. The training portion supported five-fold cross-validated model building, whereas the testing portion stayed untouched for unbiased final evaluations. Three distinct algorithms were constructed and compared: Lasso-penalized logistic regression (with lambda.min chosen via cross-validation), Random Forest (500 trees, mtry = 3), and XGBoost (tree-based booster with a binary logistic objective). For XGBoost, specific hyperparameters—eta = 0.05, max_depth = 3, subsample = 0.8, and colsample = 0.8—were pre-specified based on previous literature and preliminary experiments to balance performance and overfitting, given the modest sample size. The optimal number of boosting rounds was determined by early stopping, using validation AUC with a patience of 10 rounds.

Model Evaluation and Performance Metrics

On the held-out test set, each model's predictive capacity was quantified from multiple angles. Discriminatory performance was assessed via the area under the receiver operating characteristic curve (AUC-ROC), with 95% confidence intervals derived through DeLong's method. Calibration was inspected by plotting decile-based predicted probabilities against observed event rates, supplemented by the Brier score as an overall measure of prediction error. To explore practical utility, we identified an optimal

probability threshold using the Youden index. At that cut-point, we computed sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy.

Model Interpretation using SHAP

Given its superior overall performance on the test set, the final XGBoost model was selected for interpretability analysis via SHapley Additive exPlanations (SHAP), implemented with the shapviz package. SHAP values offered a granular breakdown of each of the eight features' contributions to individual risk predictions. Global interpretations are presented through a beeswarm summary plot—which simultaneously ranks variable importance and indicates directional effects—alongside a bar chart of mean absolute SHAP values. For local clarity, a waterfall plot for a representative patient concretely illustrates how specific feature values shaped that particular individual's estimated probability.

Statistical Software

All analytic steps—from data preprocessing and feature selection through machine-learning development to visualization—were carried out in R 4.5.3, supported by the following core packages: tidyverse, caret, glmnet, randomForest, xgboost, pROC, shapviz, and rmda. Except where explicitly noted, a two-sided $P < 0.05$ marked statistical significance for the logistic regression analyses.

Results

Study Population and Baseline Characteristics

This investigation recruited 519 participants overall. Among them, 289 cases (55.6%) received a PE diagnosis on the basis of positive CTPA results. The cohort's age median stood at 69 years (interquartile range (IQR): 60–76), with males constituting 54.5% of the sample. Table 1 provides a complete breakdown of baseline demographic and clinical profiles, grouped according to PE status.

Table 1

Baseline characteristics of the study population stratified by pulmonary embolism (PE) status

Variable	No Pulmonary Embolism (n = 230)	Pulmonary Embolism (n = 289)	P value
Age, years, mean (SD)	68.20 (13.07)	65.11 (15.44)	0.016
BMI, kg/m ² , mean (SD)	23.58 (3.45)	24.05 (3.46)	0.127
D-dimer, mg/L, mean (SD)	5.87 (6.04)	8.00 (6.77)	< 0.001
CRP, mg/L, mean (SD)	26.05 (41.57)	31.06 (40.48)	0.167
SpO ₂ , %, mean (SD)	97.52 (1.28)	97.05 (1.99)	0.002
Heart rate, bpm, mean (SD)	74.66 (11.32)	74.70 (12.01)	0.965
Male, n (%)	124 (53.9)	159 (55.0)	0.871
Diabetes mellitus, n (%)	43 (18.7)	69 (23.9)	0.188
Hypertension, n (%)	123 (53.5)	154 (53.3)	1.000
Smoking history, n (%)	67 (29.1)	96 (33.2)	0.367
Hyperlipidemia, n (%)	35 (15.2)	59 (20.4)	0.158
Limb circumference abnormality, n (%)	31 (13.5)	137 (47.4)	< 0.001
Dyspnea, n (%)	14 (6.1)	44 (15.2)	0.002
Chest pain, n (%)	3 (1.3)	10 (3.5)	0.201
Recent surgery, n (%)	39 (17.0)	49 (17.0)	1.000
Immobility, n (%)	35 (15.2)	33 (11.4)	0.253
Cancer history, n (%)	15 (6.5)	52 (18.0)	< 0.001
Data are presented as mean (standard deviation), median (interquartile range), or n (%), as appropriate.			
P-values were calculated using Student's t-test or Mann–Whitney U test for continuous variables and chi-square test or Fisher's exact test for categorical variables.			
PE, pulmonary embolism; body mass index (BMI); SpO ₂ , peripheral oxygen saturation; CRP, C-reactive protein.			

Feature Selection and Independent Risk Factors

Univariate logistic screening initially yielded ten candidate variables meeting the $P < 0.1$ threshold. These were then fed into a multivariable logistic framework that employed stepwise (bidirectional) elimination. Ultimately, eight independent determinants persisted within the final model: age, hyperlipidemia, limb circumference, D-dimer, oxygen saturation (SpO₂), dyspnea, immobilization, and malignancy (cancer).

Table 2 reports the corresponding odds ratios, their 95% confidence intervals, and associated P-values for each of these eight predictors. Of all retained factors, limb circumference exhibited the most robust relationship with PE risk, yielding an OR of 5.87 (95% CI: 3.341–10.756, $P < 0.001$).

Table 2
Multivariable logistic regression analysis of independent risk factors for pulmonary embolism

Variable	OR (95% CI)	P-value
Age	0.983 (0.966–1.001)	0.0631
Hyperlipidemia	1.648 (0.88–3.138)	0.122
Limb circumference	5.87 (3.341–10.756)	< 0.001
D-dimer	1.046 (1.008–1.087)	0.0192
SpO ₂	0.876 (0.752–1.013)	0.0808
Dyspnea	2.839 (1.264–6.817)	0.0144
Immobility	0.442 (0.216–0.88)	0.022
Cancer	3.627 (1.688–8.387)	0.0015
OR, odds ratio; CI, confidence interval. Variables were selected via univariate screening ($P < 0.1$) followed by stepwise multivariable logistic regression. Hosmer–Lemeshow goodness-of-fit test $P = 0.5626$. All categorical variables were coded as 1 = Yes, 0 = No.		

Machine Learning Model Performance

Discrimination (AUC-ROC)

When evaluated on the withheld test subset ($n = 156$), every one of the three ML algorithms demonstrated satisfactory discriminatory power for PE prediction. Lasso achieved the highest AUC of 0.766 (95% CI: 0.691–0.840), comparable to XGBoost (AUC = 0.764, 95% CI: 0.690–0.837) and Random Forest (AUC = 0.729, 95% CI: 0.649–0.808). XGBoost was selected as the primary model for its superior interpretability via SHAP. Table 3 summarizes the detailed performance metrics of all three models, including sensitivity, specificity, PPV, NPV, and accuracy at the optimal threshold. Figure 1 depicts the ROC curves for all three models.

Table 3
Performance comparison of three machine learning models for predicting PE risk

Model	AUC	AUC_CI	Sensitivity	Specificity	PPV	NPV	Accuracy
Lasso	0.766	0.691–0.840	NA	NA	NA	NA	NA
Random Forest	0.729	0.649–0.808	NA	NA	NA	NA	NA
XGBoost	0.764	0.690–0.837	74.40%	66.70%	73.60%	67.60%	71.00%
AUC, area under the receiver operating characteristic curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.							
All metrics were evaluated on the independent testing set (n = 156).							
For the XGBoost model, sensitivity, specificity, PPV, NPV, and accuracy were calculated at the optimal probability threshold of 0.526, determined by the Youden index (maximizing sensitivity + specificity – 1).							
Lasso and Random Forest models are reported for discrimination comparison only; their clinical metrics were not calculated because a threshold was not optimized.							

Calibration and Overall Performance

Plotting predicted probabilities against actual outcomes (calibration curve) for the XGBoost model revealed generally reasonable concordance, although minor deviations appeared in the intermediate-risk zone (around 0.5 predicted probability)—a pattern probably owing to the modest cohort size (Fig. 2). The Brier score for this model amounted to 0.224, whereas the null model's Brier score equalled 0.247, translating into a 9.3% relative gain. This reduction confirms that the model furnishes additional prognostic information beyond a simple average-based estimate.

SHAP Analysis

To interpret the XGBoost model, SHAP values were calculated for all eight predictors. Figure 3 presents the beeswarm summary plot, which ranks feature importance and displays the direction of effect for each predictor. Limb circumference, D-dimer, and malignancy were identified as the top three contributors to PE risk prediction. Figure 4 shows the mean absolute SHAP values, confirming the same ranking of feature importance. Figure 5 illustrates a waterfall plot for a representative patient, demonstrating how each feature contributed to the individual predicted risk; the final predicted log-odds of – 0.163 corresponded to a risk probability of approximately 45.9%, which was below the optimal threshold (0.526) and classified as low risk.

Discussion

This retrospective work involved constructing and internally testing three ML classifiers—Lasso, Random Forest, and XGBoost—to predict PE risk from a set of eight routinely obtainable clinical variables. Among

these, XGBoost demonstrated comparable performance to Lasso and was selected as the primary model for its superior interpretability, registering an AUC of 0.764 (95% CI: 0.690–0.837), along with 74.4% sensitivity and 66.7% specificity at the chosen cutoff of 0.526. SHAP-based variable attribution singled out limb circumference as the dominant driver of predicted risk, followed in descending order by D-dimer, malignancy, dyspnea, age, oxygen saturation, immobility, and hyperlipidemia. These outcomes indicate that a lean collection of clinical measures, when coupled with a boosting algorithm, can furnish moderate discriminative power for PE triage, potentially helping clinicians decide which patients warrant expedited imaging.

Turning to the SHAP importance rankings, they offer clinically plausible interpretations. Limb circumference took the top position—an expected and clinically compelling finding. Increased limb circumference typically reflects lower extremity deep vein thrombosis (DVT), which is the predominant source of pulmonary emboli. This strong association underscores the clinical importance of physical examination in PE risk stratification and aligns with the well-established paradigm that DVT and PE represent two manifestations of the same thromboembolic disease [13]. D-dimer claimed second place; this biomarker remains the cornerstone of non-imaging exclusion strategies, and its elevated SHAP value reaffirms its pivotal role even within a non-linear modelling framework, consistent with earlier reports [14]. Interestingly, immobility exhibited an inverse association with PE in the multivariable analysis (OR = 0.442, P = 0.022), which contradicts established clinical knowledge. This unexpected finding may be explained by confounding by indication—immobilized patients in our cohort were more likely to receive prophylactic anticoagulation, thereby reducing their actual PE risk—or by selection bias inherent to the retrospective design. Therefore, this association should be interpreted with caution and does not imply a true protective effect of immobility. Malignancy and immobilisation, both classic constituents of the Wells and Geneva scores, surfaced in our final predictor set, further bolstering confidence in the clinical soundness of our selection procedure. Hyperlipidemia contributed the least, implying that chronic metabolic disturbances exert a weaker influence on acute PE episodes than do symptomatic and laboratory parameters—a pattern probably reflecting the emergency-oriented composition of our study population.

How does our model compare with preceding ML efforts for PE prediction? The AUC (0.764) falls within the range documented by [15]. More importantly, whereas some investigators have drawn upon high-dimensional data sources—including comprehensive electronic health record (EHR) extractions or natural language processing (NLP)-processed notes [16]—our solution demands merely eight items uniformly captured in urgent care settings. This parsimony boosts transportability and curbs overfitting, facilitating potential adoption in lower-resource clinical environments. Nevertheless, our AUC trails behind figures obtained with richly annotated research registries or deep learning models that analyze computed tomography pulmonary angiography (CTPA) images [17], illustrating the inevitable tension between algorithmic complexity and real-world feasibility.

From a practical vantage point, the proposed tool supplies emergency physicians with a workable risk-stratification aid. Applying the 0.526 threshold, the algorithm caught 74.4% of truly PE-positive

individuals while correctly excluding 66.7% of those without the disease. In absolute terms, roughly two of every three unnecessary CTPA orders might be spared through such a triage step—a non-trivial reduction, considering the widely recognised dangers of radiation, contrast-associated nephropathy, and escalating expenditures linked to overuse of CTPA [18]. The accompanying SHAP waterfall visualisation further enables case-level risk decomposition, allowing attending doctors to explain to patients, in an accessible manner, why a particular prediction emerged as high risk. Still, we stress that this instrument is designed as a supplementary decision aide, never as a replacement for seasoned clinical reasoning. We emphasize that this model is intended as a decision-support tool for risk stratification, rather than a standalone diagnostic or rule-out test. Its optimal use would be to identify patients at sufficiently low risk to safely defer CTPA, while those in the intermediate/high-risk groups would still require clinical judgment and further testing. Sensitivity could be improved at the cost of specificity by lowering the threshold, depending on the clinical setting

Several caveats deserve mention. The single-centre, retrospective design inevitably introduces selection biases and restricts external applicability. Our sample size proved modest—especially the count of PE-positive cases, approximately 289—which probably contributed to mild overoptimism during cross-validation and to the wiggles observed in the calibration curve. External validation using an independent cohort has not been performed, yet such a step remains indispensable before any prospective clinical deployment. The Brier score of 0.224, although better than the null (0.247), still lies somewhat above the ideal range, signalling that calibration could be refined—a shortfall likely linked to limited data and the elevated disease prevalence in the test partition. The PE prevalence in our cohort (55.6%) was substantially higher than that typically observed in unselected emergency department patients (10%–20%). This reflects the fact that our institution is a tertiary referral center, which predominantly receives patients with high clinical suspicion of PE. Therefore, our findings may not be directly generalizable to primary or community emergency settings with lower PE prevalence. External validation in low-prevalence cohorts is warranted. Additionally, we omitted several candidate predictors that might have enhanced performance, such as arterial blood gas results, echocardiographic measures, or detailed medication histories. Future investigations should prioritise multicentre prospective designs, incorporate supplementary biomarkers, and embed the model within clinical decision-support systems to assess its tangible effects on patient management.

Conclusions

This study has constructed and internally tested a lean XGBoost-based prediction tool that relies on eight commonly collected clinical variables—age, hyperlipidemia, limb circumference, D-dimer, oxygen saturation, dyspnea, immobility, and malignancy—for estimating PE likelihood. The model exhibited satisfactory discriminatory power (AUC = 0.764, 95% CI: 0.690–0.837) coupled with moderate calibration. At the chosen cut-off, sensitivity reached 74.4% and specificity 66.7%. SHAP analysis was applied to boost clinical transparency, revealing limb circumference and D-dimer as the most potent contributors to predicted risk.

Such an instrument could function as a decision-support aid for emergency physicians in triage, potentially curbing unnecessary CTPA referrals while maintaining diagnostic safety. Nevertheless, the retrospective and single-centre character of this work limits generalizability; external validation using larger, prospectively collected multi-centre datasets remains a prerequisite for eventual clinical adoption. Subsequent efforts should also consider embedding the algorithm within EHR platforms to facilitate real-time bedside decisions and to evaluate its effects on both patient outcomes and healthcare expenditures.

Abbreviations

<i>Abbreviation</i>	<i>Full Term</i>
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
CI	Confidence interval
CRP	C-reactive protein
CTPA	Computed tomography pulmonary angiography
DVT	Deep vein thrombosis
EHR	Electronic health record
EMR	Electronic medical record
IQR	Interquartile range
IRB	Institutional Review Board
Lasso	Least Absolute Shrinkage and Selection Operator
ML	Machine learning
NLP	Natural language processing
NPV	Negative predictive value
OR	Odds ratio
PE	Pulmonary embolism
PPV	Positive predictive value
ROC	Receiver operating characteristic
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SpO ₂	Peripheral oxygen saturation
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
XGBoost	Extreme Gradient Boosting

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the Institutional Review Board of Sir Run Run Shaw Hospital (Approval No.: 2026-2131-02).

The requirement for written informed consent was waived by the IRB due to the retrospective nature of the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to hospital privacy regulations but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

JG and CL conceived and designed the study. JG and CL collected and curated the data. JG performed the statistical analysis and machine learning modeling. CL drafted the initial manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

JG is a nurse at Sir Run Run Shaw Hospital, affiliated with Zhejiang University School of Medicine. CL is a vascular surgeon at the same institution.

References

1. Zuin M, Becattini C, Klok FA, Bikdeli B, Barco S, Farmakis I, et al. Pulmonary embolism epidemiology: trends, risk factors, comorbidities, and clinical implications. *Eur Heart J*. 2026;47(27):3555–71.
2. Essien EO, Rali P, Mathai SC. Pulmonary Embolism. *Med Clin North Am*. 2019;103(3):549–64.
3. van der Hulle T, Cheung WY, Kooij S, Beenen LFM, van Bommel T, van Es J, et al. Simplified diagnostic management of suspected pulmonary embolism (the YEARS study): a prospective, multicentre, cohort study. *Lancet*. 2017;390(10091):289–97.

4. Zantonelli G, Cozzi D, Bindi A, Cavigli E, Moroni C, Luvara S, Grazzini G, Danti G, Granata V, Miele V. Acute Pulmonary Embolism: Prognostic Role of Computed Tomography Pulmonary Angiography (CTPA). *Tomography*. 2022;8(1):529–39.
5. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *N Engl J Med*. 2003;349(13):1227–35.
6. Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H, Perrier A. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. *Ann Intern Med*. 2006;144(3):165–71.
7. Di Marca S, Cilia C, Campagna A, D'Arrigo G, Abd ElHafeez S, Tripepi G, Puccia G, Pisano M, Mastrosimone G, Terranova V, et al. Comparison of Wells and Revised Geneva Rule to Assess Pretest Probability of Pulmonary Embolism in High-Risk Hospitalized Elderly Adults. *J Am Geriatr Soc*. 2015;63(6):1091–97.
8. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, Kannampallil T, Abraham J. Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications. *JAMA Netw Open*. 2021;4(3):e212240.
9. Khodadadi F, Taghizadeh F, Hashemi Baghi A, Ayyoubzadeh SM, Dadashzadeh S, Haeri A. Leveraging ensemble machine learning models (XGBoost and random forest) and genetic algorithms to predict factors contributing to the liposomal entrapment of therapeutics. *Nanoscale*. 2025;17(38):22271–90.
10. Hunsaker AR. Deep Learning and Risk Assessment in Acute Pulmonary Embolism. *Radiology*. 2022;302(1):185–6.
11. Li Y, Zhang L, Liu H, Li Y, Liu Z. Research progress of artificial intelligence and machine learning in pulmonary embolism. *Front Med (Lausanne)*. 2025;12:1577559.
12. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell*. 2019;1(5):206–15.
13. Di Nisio M, van Es N, Buller HR. Deep vein thrombosis and pulmonary embolism. *Lancet*. 2016;388(10063):3060–73.
14. Giannitsis E, Mills NL, Mueller C. D-Dimer in suspected pulmonary embolism. *Eur Heart J Acute Cardiovasc Care*. 2023;12(10):721–2.
15. Yoon SH, Kwon C, Choi Y, Kim HJ, Kim J, Kim YH. Machine learning improves prediction of pulmonary thromboembolism and reduces unnecessary computed tomography scans in the emergency department. *Sci Rep*. 2026;16(1):4935.
16. Hassan SM, Mylvaganam R, Didebulidze T, Khalid M, Nardelli P, Piazza G, San Jose Estepar R, Cuttica MJ, Johnson S, Dao N, et al. Leveraging Hybrid Natural Language Processing Techniques for Large-Scale Pulmonary Embolism Identification. *JACC Adv*. 2025;4(11 Pt 2):101845.
17. Li Y, Chen S, Chen G, Ren B, Hu S, Zong M, Jia Z, Xue T, Feng G, Li D. A deep learning model for the interpretable identification of pulmonary thromboembolism from computed tomography pulmonary

angiography. *Front Cardiovasc Med.* 2026;13:1864916.

18. Mirabile A, Lucarelli NM, Sollazzo EP, Stabile Ianora AA, Sardaro A, Mirabile G, Lorusso F, Racanelli V, Maggialetti N, Scardapane A. CT pulmonary angiography appropriateness in a single emergency department: does the use of revised Geneva score matter? *Radiol Med.* 2021;126(12):1544–52.

Figures

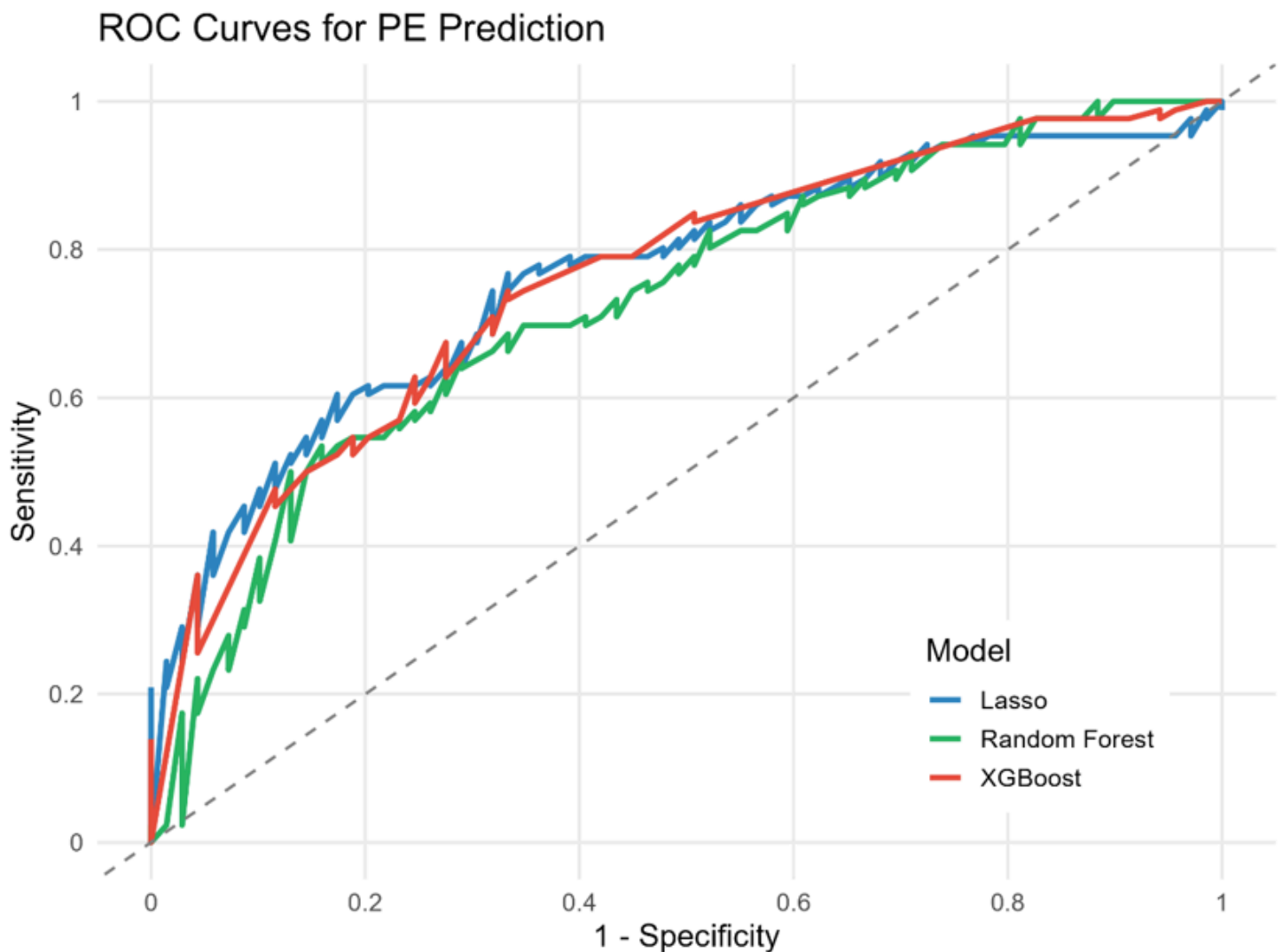


Figure 1

Receiver operating characteristic (ROC) curves of three machine learning models for predicting pulmonary embolism

Legend: ROC curves comparing the predictive performance of Lasso, Random Forest, and XGBoost models on the independent testing set (n = 156). The diagonal dashed line represents the performance of a random classifier (AUC = 0.5). Lasso achieved the highest AUC of 0.766 (95% CI: 0.691–0.840),

comparable to XGBoost (AUC = 0.764, 95% CI: 0.690–0.837) and Random Forest (AUC = 0.729, 95% CI: 0.649–0.808). XGBoost was selected as the primary model for its superior interpretability via SHAP.

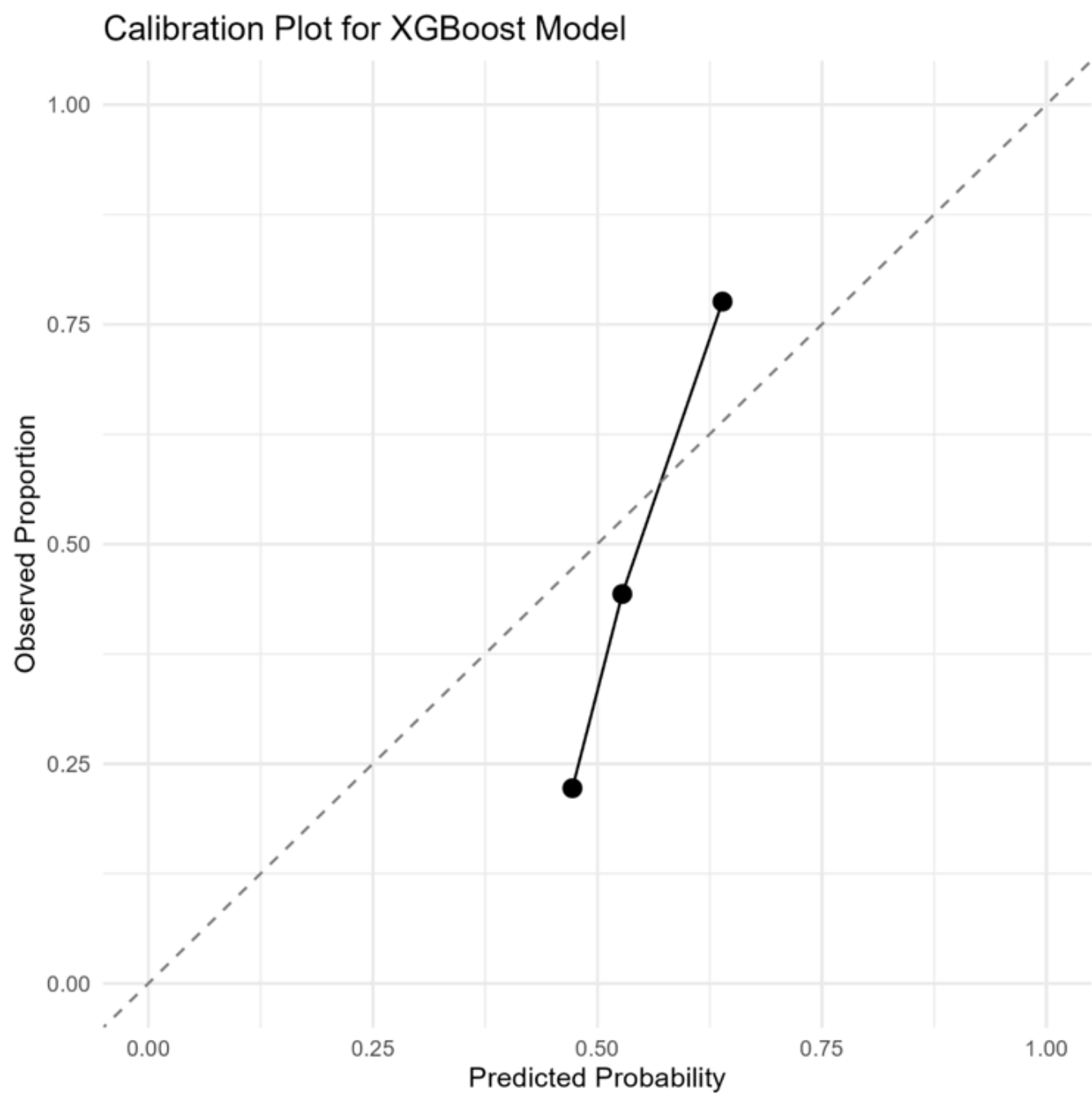


Figure 2

Calibration curve of the XGBoost model for predicting pulmonary embolism risk

Legend: Calibration plot assessing the agreement between predicted probabilities (x-axis) and observed proportions (y-axis) of pulmonary embolism in the testing set. The dashed diagonal line represents perfect calibration (predicted probability = observed proportion). The solid red line shows the calibration performance of the XGBoost model. A generally acceptable calibration was observed, with slight

deviations in the mid-risk range (around 0.5 predicted probability), which may be attributed to the limited sample size. The Brier score of the model was 0.224, compared to the null model Brier score of 0.247, showing a modest improvement over the null model (0.247). This indicates moderate calibration, which may be attributed to the relatively small test set and the high prevalence of PE in our cohort.

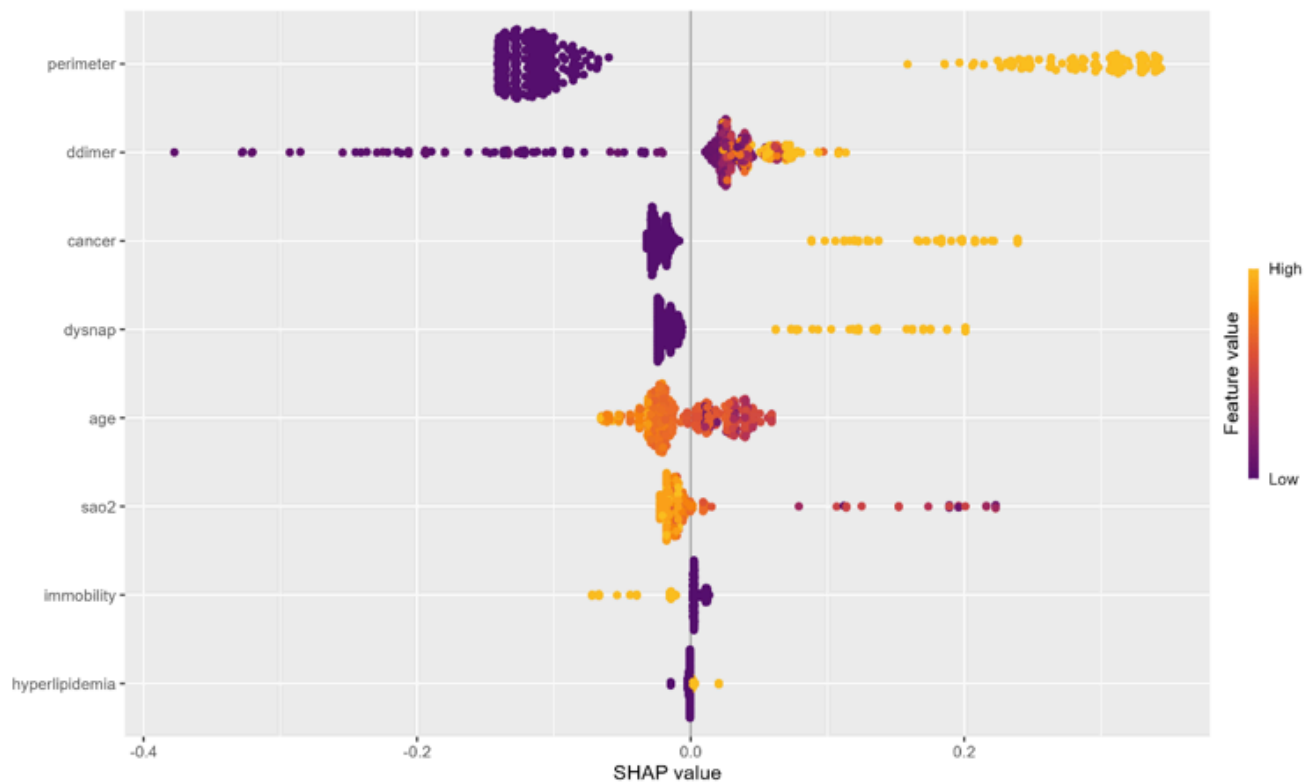


Figure 3

SHAP feature importance summary plot (beeswarm) for the XGBoost model

Legend: Beeswarm summary plot of SHAP values for the eight predictors in the XGBoost model. Each point represents a patient, with the x-axis indicating the SHAP value (positive = increased PE risk, negative = decreased PE risk) and the y-axis showing features ranked by their mean absolute SHAP value (top = most important). The color scale represents the feature value (red = high, blue = low). Limb circumference , D-dimer, and malignancy (cancer) were the top three contributors to PE risk prediction. The plot demonstrates that higher limb circumference, elevated D-dimer, presence of malignancy, dyspnea, lower oxygen saturation, and immobility showed mixed directional effects, possibly reflecting confounding in the data, while older age and hyperlipidemia showed more complex directional effects. SHAP, SHapley Additive exPlanations.

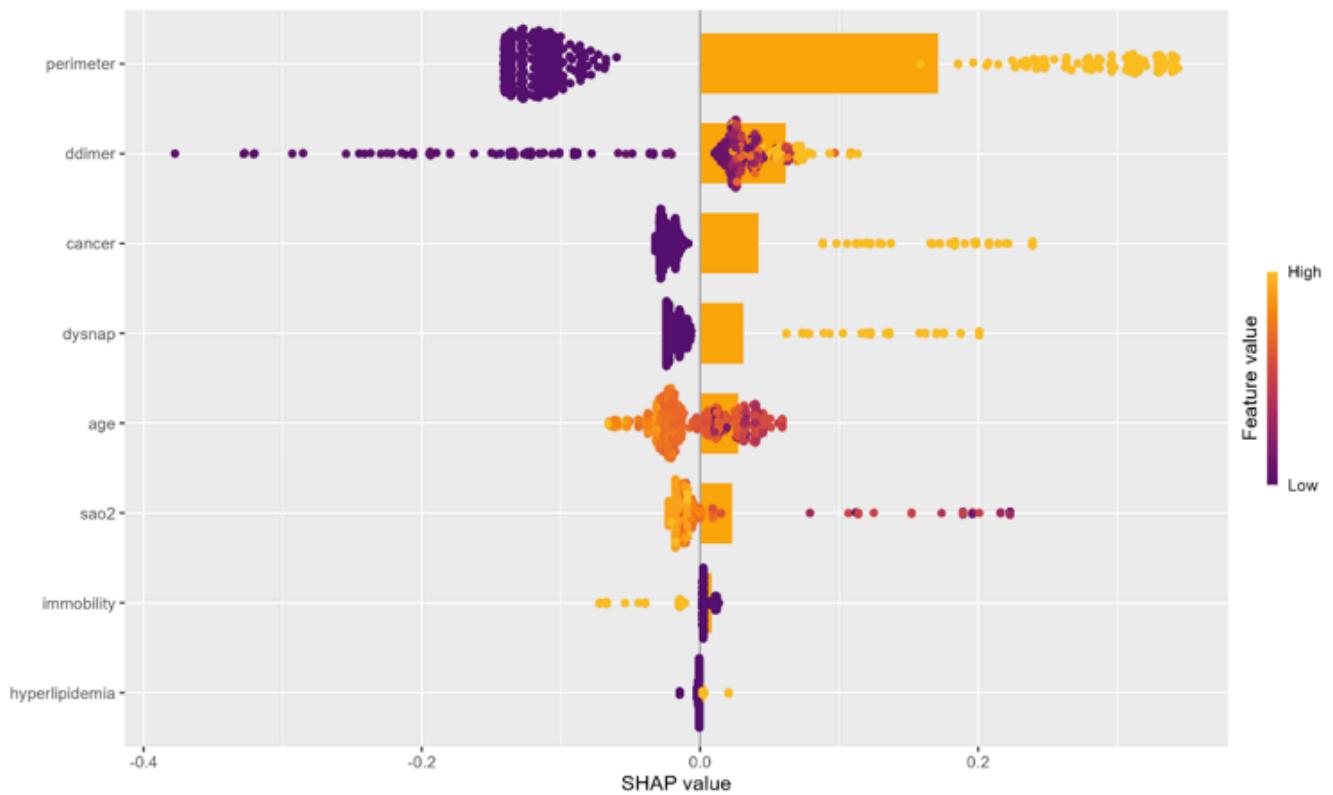


Figure 4

SHAP mean absolute importance bar plot

Legend: Bar plot showing the global feature importance of the XGBoost model based on the mean absolute SHAP values. Bars represent the average contribution of each feature to the absolute prediction value across all patients in the testing set. Longer bars indicate greater overall influence on the model's predictions. The ordering is consistent with the beeswarm plot, confirming that limb circumference was the most influential predictor, followed by D-dimer and malignancy. SHAP, SHapley Additive exPlanations.

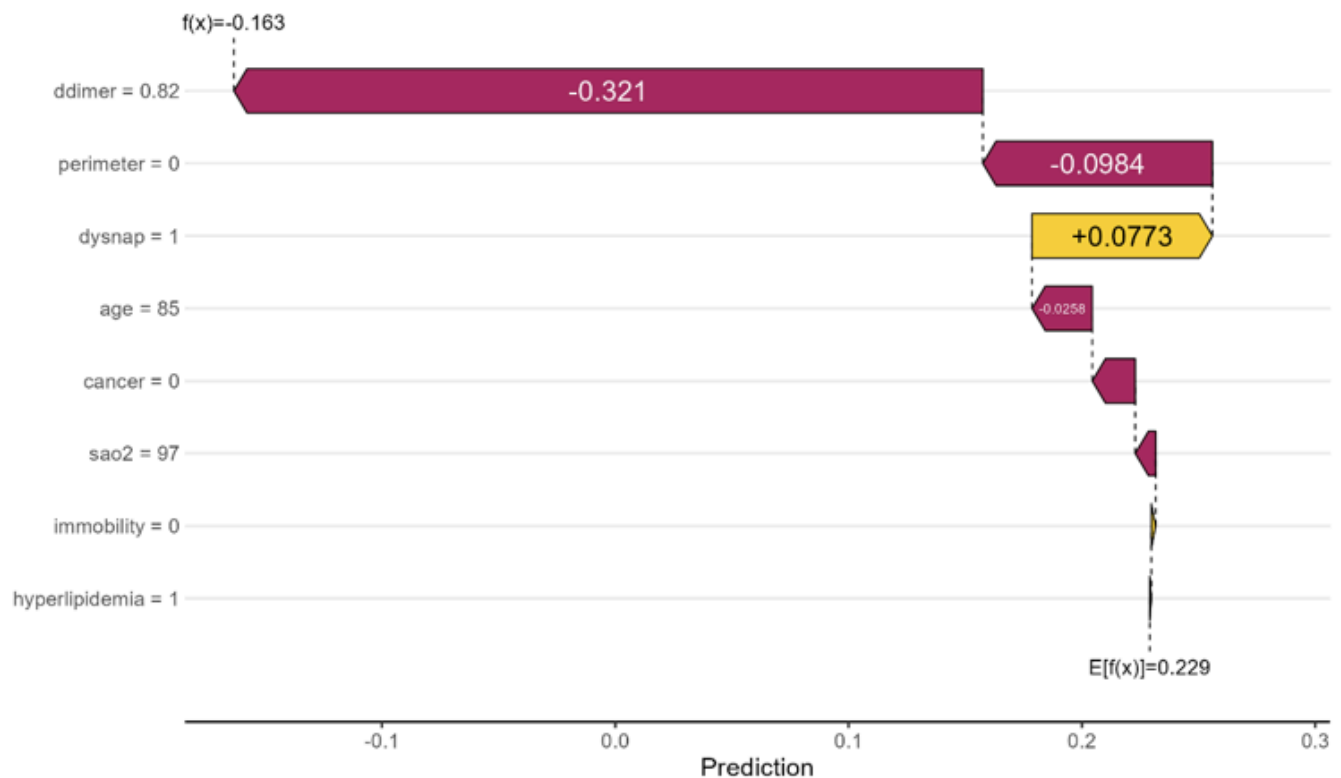


Figure 5

SHAP waterfall plot for an individual patient (example)

Legend: Waterfall plot illustrating the individualized contribution of each feature to the final prediction for a representative patient in the testing set. The baseline value ($E[f(x)]$) represents the average log-odds of PE across the cohort. Each arrow shows how a specific feature shifts the prediction away from the baseline: red arrows indicate features that increase the predicted risk (positive contribution), and blue arrows indicate features that decrease the predicted risk (negative contribution). The final predicted log-odds of -0.163 corresponds to a risk probability of approximately 45.9%, which is below the optimal threshold (0.526), classifying this patient as low risk. In this case, limb circumference was the primary driver increasing the predicted risk (red arrow), while a relatively low D-dimer value partially offset this effect. This visualization demonstrates the clinical interpretability of the model at the individual level.