Research article

A multi-center big-data approach for precise PICC-RVT prognosis and identification of major risk factors in clinical practice

Yue Li [a,b], Ting Li [a,*], Hengjie Su [a], Xin Zhang [a], Jiangbo Pu [a], Hong Sun [c], Qiong Liu [d], Bowen Zhang [a], Biao Sun [e], Jia Li [f], Xinxin Yan [g], Laiyou Wang [h]

[a] Institute of Biomedical Engineering, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[b] School of Electronics and Information Engineering, TianGong University, Tianjin, China
[c] Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China
[d] Department of Ultrasound, First Hospital of Shanxi Medical University, Shanxi, China
[e] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[f] Macquarie University, Australia
[g] Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[h] Department of Clinical Pharmacy, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

A R T I C L E I N F O

A B S T R A C T

Background: The Peripherally Inserted Central Catheter (PICC) is a widely used technique for delivering intravenous fluids and medications, especially in critical care units. PICC may induce venous thrombosis (PICC-RVT), which is a frequent and serious complication. In clinical practice, Color Doppler Flow Imaging (CDFI) is regarded as the gold standard for diagnosing PICC-RVT. However, CDFI not only requires prominent time and effort from experienced healthcare professionals, but also relies on the formation and development of PICC-RVT, especially at early stages of PICC-RVT, when PICC-RVT is not apparent. A prognosis tool for PICC-RVT is crucial to bridge the gap between its diagnosis and treatment, especially in resource-limited settings, such as remote healthcare facilities.
Objective: Evaluate over 14,885 models from various machine learning techniques to identify an effective prognostic model (referred to as PRAD - PICC-RVT Assessment via Deep-learning) for quantifying the risks associated with PICC-RVT.
Methods: To tackle the challenges associated with PICC-RVT diagnosis, we gathered a comprehensive dataset of 5,272 patients from 27 healthcare centers across China. From a pool of 14885 models from various machine learning techniques, we systematically screened a data-driven prognostic model to quantify the risks associated with PICC-RVT. This model aims to provide objective evidence, and facilitate timely interventions.
Results: The proposed model displayed exceptional predictive accuracy, achieving an accuracy of 86.4 % and an AUC of 0.837. Based on the prognosis model, we further incorporated a weight analysis to identify the major contributing factors for PICC-RVT risk during catheterization. Albumin levels, primary diagnosis, hemoglobin levels, platelet levels, and education level are emphasized as important risk factors.
Conclusions: Our method excels in predicting early PICC-RVT risks, especially in asymptomatic patients. The findings in this paper offers insights into controllable PICC risk factors that could benefit vast patients and reduce disease burden through stratification and early intervention.

* Corresponding author.
  E-mail address: liting@bme.pumc.edu.cn (T. Li).

1. Introduction

Peripherally inserted central catheters (PICCs) are commonly used in clinical settings to establish venous access for a range of treatments, including antibiotics, chemotherapy medications, and total parenteral nutrition [1]. Within high-risk infection environments, such as the critical care units, PICCs are frequently employed to deliver therapies [2]. Research indicates that critically ill patients face a risk of PICC-related venous thrombosis (PICC-RVT) ranging from 13 % to a staggering 91 % [3]. Approximately 30 % of these cases can result in severe complications from thrombosis, such as pulmonary embolism, post-thrombotic syndrome, limb paralysis, and even mortality [4]. Immediate standard anticoagulant protocols should be initiated upon identifying PICC-RVT to prevent these adverse clinical outcomes. A distinctive feature of PICC-RVT is the development of thrombosis within the vascular system [4]. Many PICC-RVT cases are either asymptomatic or nonspecific, representing 50 % or more of the adult population [5]. Color Doppler ultrasound is the primary diagnostic tool for PICC-RVT [6]. However, it is used administered when patients demonstrate specific symptoms, which can lead to an under-diagnosis of PICC-RVT. While consistent monitoring of thrombosis in patients is crucial, its practical implementation in clinical settings is often challenging [7]. Those evidences underscore the urgent need for a reliable and objective prognosis tool to quantitatively assess the likelihood of PICC-RVT onset.

Artificial Neural Networks (ANNs) excel at identifying complex patterns within large datasets. which can automatically learn from data via multi-layered feature representations, making them ideal for tasks like automatic categorization, recognition, and prediction. Recently, extensive research has focused on identifying factors related to PICC-RVT, such as lumen count, patient comorbidities, previous chemotherapy sessions, instances of deep vein thrombosis, and history of metastatic cancer [8–10]. Various predictive approaches have been explored [11]. However, most datasets used for these predictions have been limited to patients with similar medical conditions or confined to a single medical institution. The aim of this study was to overcome the limitations of previous research on peripherally PICC-RVT, which had small sample sizes and heterogeneous criteria.

We collected data from 27 medical centers across China, involving 5,272 patients with diverse clinical presentations. To objectively assess the risk of PICC-RVT for each patient, we evaluated more than 14,885 models using various machine learning techniques. Through rigorous screening, we narrowed our focus to four prominent models—PRAD, Random Forest, AE, and SVM—chosen for their exceptional performance and strong relevance to our dataset. Despite the recognition of these models, our comprehensive approach necessitated evaluating a broad range of models to identify the optimal combination for performance. Subsequently, we applied the SHAP [12] method to analyze the contribution of different features to the predictions made by the PRAD model. These insights not only establish a foundation for the quantitative classification of patients at risk for PICC-RVT but also have important implications for advancing personalized medicine, improving the allocation of healthcare resources, and alleviating the burden of PICC-RVT.

2. Material and methods

2.1. Patients and data sources

This study was a comprehensive retrospective multicenter cohort analysis conducted between May 2015 and July 2023. Retrospective studies typically utilize pre-existing data, which, in this case, was anonymized and managed in compliance with institutional and ethical guidelines to ensure patient confidentiality. Ethical approval for the study was obtained from Beijing Hospital, with the approval document included as an attachment. Furthermore, the study has been registered with the Chinese Clinical Trial Registry (Registration Number: ChiCTR2300070265).

During the data collection phase, each participating medical institution collated records of patients with PICC insertions, culminating in data from 5272 individuals (Supplementary Table S1). These patients underwent regular catheter maintenance and color Doppler ultrasonographic examinations. However, upon further examination of this data, we identified inconsistencies and deviations from our research benchmarks. Consequently, we set exclusion criteria for the data. The inclusion criteria for patients are: (a) aged over 18; (b) requiring intravenous administration of antineoplastic drugs or nutritional support; (c) a life expectancy of more than 30 days post-PICC insertion. The exclusion criteria are incomplete relevant data or record data anomaly.

2.2. Catheter insertion, maintenance, and monitoring

The PICC insertion was handled by nurses certified in venous therapy by China's provincial or city Nursing Association. Before the study began, they received comprehensive training on PICC-RVT-related knowledge and diagnostic criteria. Once the catheter was accurately placed, a sterile transparent dressing covered the insertion site, and expert physicians verified the correct positioning of the catheter's tip. Moving forward, skilled nurses provided weekly maintenance for the catheter. The dressing was first replaced within 24 h of insertion and then every 5–14 days. For patients receiving chemotherapy, the catheter was consistently flushed with 10 ml of saline before and after each medication administration.

After PICC insertion, weekly CDFI assessments were performed to continuously monitor for potential thrombosis. Further details about CDFI are provided in reference (15). If patients presented with symptoms such as pain, swelling, erythema, or increased skin temperature in the limbs near the PICC site or in adjacent areas like the shoulder, neck, or chest, along with significant venous distension in the affected limb, further evaluation was necessary. Strong indicators of PICC-RVT, as identified through color Doppler ultrasound, included: (a) a non-compressible venous lumen, (b) absence of flow signals within the lumen, (c) visible solid echoes inside the lumen, (d) loss of the phasic characteristics of venous blood flow, (e) a diminished or absent Fajer response, and (f) reduced or
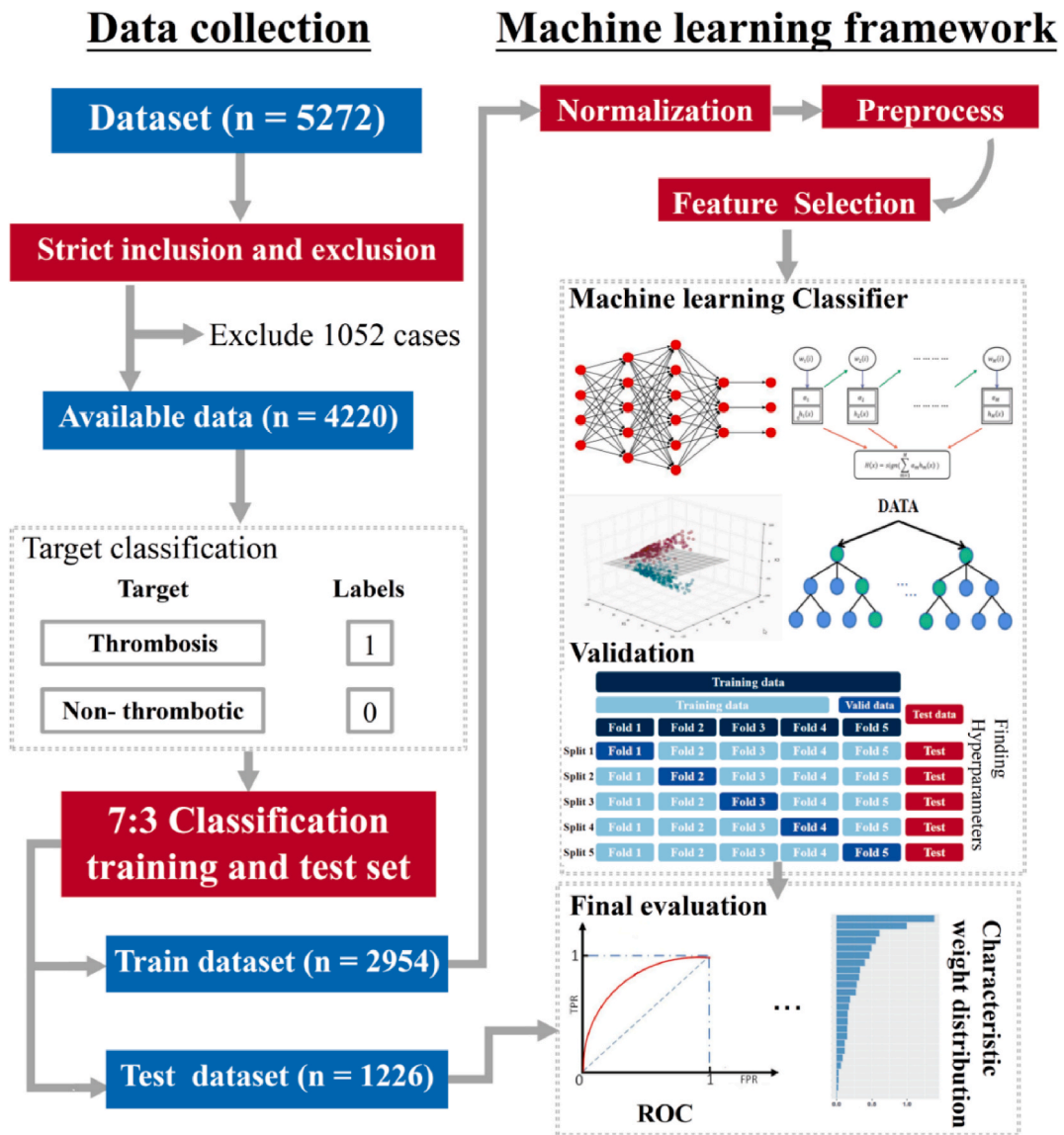
Data collection

Dataset (n = 5272)

↓

Strict inclusion and exclusion

→ Exclude 1052 cases

↓

Available data (n = 4220)

↓

Target classification

Target	Labels
Thrombosis	1
Non- thrombotic	0

↓

7:3 Classification training and test set

Train dataset (n = 2954)

Test dataset (n = 1226)

Machine learning framework

Normalization → **Preprocess**

Feature Selection

↓

Machine learning Classifier

DATA

Validation

Finding Hyperparameters

	Training data					
	Training data				Valid data	Test data
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test

Final evaluation

ROC

Characteristic weight distribution

Fig. 1. Flow chart of machine learning model building.

absent flow enhancement upon compression of the distal corneal rim. A detailed description of the Fajer response is available in reference (15).

2.3. Data processing

In this study, we undertook data cleaning, which involved removing duplicate entries, addressing missing values, and ensuring data accuracy and consistency. Following this process, we split the cleaned dataset into training and testing subsets with a 7:3 ratio. Continuous variables were summarized using mean and standard deviation, while categorical variables were presented as percentages. The training set was utilized for developing deep learning models and their machine learning counterparts, while the test set was reserved exclusively for evaluating the performance of these models. Python (version 3.8.0) was employed to manage data cleaning, model development, training, and predictive analysis. Machine learning models were compared using FeAture Explorer Pro 0.5.4 (FAE, as cited in reference [13], link: https://github.com/salan668/FAE), and statistical analyses, along with related variable calculations, were conducted using SPSS (from IBM).

During our detailed data analysis, we noticed a significant imbalance in the sample distribution. Negative samples (patients without thrombosis) far outnumbered the positive samples (thrombotic patients). Training on such skewed data could make the model biased towards the majority class, compromising prediction accuracy for the minority class. To address this, we incorporated resampling techniques in our training approach, including oversampling, under sampling, and the Synthetic Minority Over-sampling

Technique (SMOTE) (details on these training methods can be found in references [14–16]). These methods are well-recognized in medical research for their effectiveness. To ensure model reliability, we only applied these techniques during the training phase.

2.4. Model design

To predict the risk of PICC-RVT in patients undergoing PICC treatment, we developed a Deep Neural Network (DNN). The DNN comprises three main components: the input layer, hidden layers, and the output layer. The input layer is responsible for processing raw data features from the patients. It is followed by three hidden layers: the first has 32 neurons, while the subsequent two contain 16 neurons each. All neurons in these layers use the ReLU activation function. The output layer, designed to provide a predictive probability for a patient's PICC-RVT diagnosis, contains a single neuron. It's fully connected, ensuring it considers all outputs from the hidden layers. To ensure the output represented probability values summing to 1, we emploed the softmax function.

In this study, we aimed to evaluate the risk of PICC-RVT by leveraging various machine learning models and established clinical data processing techniques. Our methodology included data preprocessing, feature selection, and ML algorithm application. We used three normalization methods, three techniques for feature dimensionality reduction, and four feature selection approaches. By combining 40 relevant variables with 10 ML algorithms, we constructed 14,400 models through permutation and combination (the process is illustrated in Supplementary Fig. S1). Following this, we built approximately 485 neural network models in different forms. In total, we constructed 14,885 predictive models and subsequently conducted an exhaustive comparison of each algorithm.

2.5. Model training and optimization

During the data manipulation phase, we allocated 2954 samples to the training set with a positive to negative ratio of 408–2546. Meanwhile, the testing dataset comprised 1266 samples with a positive to negative ratio of 174–1092. To balance the datasets, we employed data augmentation techniques, including random oversampling, SMOTE, and Adaptive Synthetic Sampling (ADASYN), aiming to enhance the number of positive samples. In the training regimen, we configured 600 epochs, each with a batch size of 32. We utilized the Adam optimizer, setting an initial learning rate at 0.001. Both L1 and L2 regularization methods were employed, and a dropout rate of 0.2 was set for the Dropout layer. This approach ensured that the network could recognize complex patterns in the training data while preventing overfitting and preserving the model's adaptability to new data. We introduced an early stopping mechanism to avoid overfitting and saved the model when it reached its best validation performance.

During the construction phase of the comparative model, we began by normalizing the data to ensure all features operated on the same scale. This not only accelerated the algorithm's convergence rate but also enhanced the model's stability. We then implemented feature dimension reduction to eliminate potential noise and redundant information. Building upon this, we used feature selection techniques to identify and retain only those features with significant contributions to model predictions. This step was crucial for improving model interpretability and minimizing the risk of overfitting. After refining the features, rigorous training was conducted using machine learning algorithms to predict the target categories accurately. A 5-fold cross-validation was performed on the training data to gauge the model's optimal predictive capability. Based on the performance metrics from the validation dataset, we carefully selected the model's hyperparameters (Fig. 1).

2.6. Evaluation metrics

The DNN model's hyperparameters were determined based on its performance on the validation set, using a 5-fold cross-validation technique. While accuracy was initially used as a primary metric to evaluate the model's predictions, the challenge posed by imbalanced datasets led us to realize that solely relying on accuracy could provide a skewed view of the model's capabilities. Thus, we incorporated additional metrics such as sensitivity and specificity for a more comprehensive assessment.

The model's AUC-ROC value was also calculated. In disease prediction, the ROC curve is highly regarded, typically appearing above the $y = x$ line. The area under this curve, or AUC, ranges between values of 0.5 and 1.0. A value closer to 1.0 indicates a high precision in disease detection, while a value near 0.5 suggests the prediction is no better than random guessing. Given that real-world clinical datasets often present a noticeable imbalance between positive and negative samples, the advantage of AUC is its ability to offer a balanced evaluation, even in such scenarios. Thus, for imbalanced datasets in disease prediction, the AUC proves more insightful than mere accuracy, suggesting the model's competence in determining the presence or absence of a disease.

Furthermore, the Odds Ratio (OR) was computed for each variable, complemented by its 95 % Confidence Interval (CI) and the P-value. These combined metrics provide a detailed evaluation of the research's statistical significance and practical implications. By incorporating these benchmarks, we ensure a thorough assessment of the model's performance, highlighting its potential for clinical applications.

3. Results

3.1. Features of subjects

From October 2015 to July 2023, clinical data was collected from 5,272 patients undergoing peripheral central catheterization across 27 tertiary-level hospitals. After applying strict inclusion and exclusion criteria, 1,052 patients who didn't meet the inclusion standards were excluded. Among them: 21 were excluded due to advanced age, 147 due to data record anomalies, and 884 because of

Table 1

Comparison of demographic information between thrombus and non-thrombotic patients.

Variable	Non-PICC-RVT (n = 3638)	PICC-RVT (n = 582)
Age(years)	52.47 ± 12.26	55.92 ± 11.99
Height(cm)	162.19 ± 8.55	164.67 ± 8.67
Weight(kg)	62.54 ± 11.46	63.92 ± 12.47
White blood cell(10 × 10 ^{9}/L)	6.87 ± 14.43	6.32 ± 3.41
Hemoglobin(g/L)	126.15 ± 33.13	127.73 ± 38.08
Platelet(10^9/L)	246.46 ± 87.10	127.73 ± 38.08
D-dimer(mg/L)	1.10 ± 3.98	1.25 ± 4.58
Prothrombin time (s)	12.96 ± 3.12	12.71 ± 2.86
Fibrinogen(g/L)	3.48 ± 1.39	3.46 ± 1.42
Activated partial prothrombin time(s)	29.29 ± 11.69	27.48 ± 5.34
Albumin(g/L)	32.80 ± 17.23	38.59 ± 9.58
C reactive protein (mg/L)	11.64 ± 26.07	12.29 ± 23.08
Catheter-to-vessel diameter ratio	0.34 ± 0.08	0.35 ± 0.08
Catheter dwell time (days)	99.65 ± 106.71	89.08 ± 83.36
Gender (Male/Female)	1121 (30.8 %)/2517 (69.2 %)	321 (55.2 %)/264 (44.8 %)
Hypertension (No/Yes)	2947 (81.3 %)/691 (18.7 %)	489 (84.4 %)/93 (15.6 %)
Diabetes (No/Yes)	3340 (91.8 %)/298 (8.2 %)	540 (92.8 %)/42 (7.2 %)
Coronary heart disease (No/Yes)	3565 (98 %)/73 (1.2 %)	576 (99 %)/6 (1 %)
Past deep vein thrombosis and thrombophlebitis (No/Yes)	3598 (98.9 %)/40 (1.1 %)	570 (98.0 %)/12 (2.0 %)
History of deep vein catheterization (No/Yes)	3423 (94.1 %)/214 (5.9 %)	531 (91.3 %)/51 (8.7 %)
Catheter lumen number (Single/Others)	3630 (99.8 %)/8 (0.2 %)	582 (100.0 %)/0 (0.0 %)
Puncture site (Left/Right)	1884 (51.8 %)/1754 (48.1 %)	288 (49.6 %)/294 (50.4 %)
Number of punctures (1/>1)	3477 (95.6 %)/161 (3.8 %)	568 (97.6 %)/14 (2.2 %)
Prestored sealing fluid (No/Yes)	2419 (66.5 %)/1219 (33.5 %)	492 (84.6 %)/90 (15.4 %)
Type of sealing fluid (Normal saline/Heparin)	3565 (98.0 %)/73 (2.0 %)	562 (96.7 %)/20 (3.3 %)
Education status:		
No education background and No record	2248 (61.8 %)	436 (74.9 %)
Primary school/Middle school	728 (20.7 %)/364 (10.6 %)	87 (15.2 %)/40 (7.1 %)
Bachelor degree or above	254 (6.9 %)	19 (2.8 %)
Main diagnosis of admission:		
Respiratory system/Circulatory system	582 (16.1 %)/4 (0.1 %)	174 (30.3 %)/0 (0.0 %)
Digestive system/Motion system	764 (21.2 %)/18 (0.5 %)	168 (28.8 %)/1 (0.2 %)
Nervous system/Urinary system	40 (1.1 %)/25 (0.7 %)	3 (0.6 %)/0 (0.0 %)
Reproductive system/Mammary glands	222 (6.1 %)/1593 (43.8 %)	31 (5.4 %)/111 (19.1 %)
Blood system/Ear, nose, throat, head, neck and mouth	181 (5.2 %)/138 (3.8 %)	32 (5.6 %)/44 (7.8 %)
Others	71 (2.0 %)	22 (4.0 %)
Tumor staging:		
None	72 (1.5 %)	9 (1.5 %)
I/II	222 (6.1 %)/726 (20.1 %)	24 (4.2 %)/71 (12.3 %)
III/IV	1127 (31.0 %)/1491 (41.2 %)	210 (36.2 %)/268 (46.3 %)
Reason for catheter retention:		
Chemotherapy drugs/Hypertonic fluid	3175 (87.2 %)/193 (5.3 %)	512 (87.9 %)/39 (6.7 %)
Long-term infusion/Others	261 (7.2 %)/9 (0.3 %)	29 (5.0 %)/2 (0.4 %)
Position of catheter tip:		
T7 and Lower than T7/Upper than T6	1640 (45 %)/1998 (54.9 %)	257 (44.1 %)/325 (55.8 %)
Chemotherapy cycle:		
No chemotherapy/0-4	1288 (35.4 %)/1308 (35.9 %)	124 (21.3 %)/310 (53.2 %)
4-8/>8	854 (23.5 %)/188 (5.1 %)	132 (22.6 %)/16 (2.7 %)
Conduit mode:		
Bard PICC Catheter 4F	3070 (84.4 %)	452 (77.7 %)
Bard-771405-4Fr/Others	527 (14.5 %)/41 (1.1 %)	124 (21.4 %)/6 (0.9 %)
Type of PICC:		
Tri-valve valvular	3452 (94.9 %)	479 (82.3 %)
Front opening/Others	73 (2.0 %)/113 (3.1 %)	58 (10.2 %)/45 (7.5 %)
PICC catheterization method:		
Under ultrasound guidance	3201 (88.0 %)	549 (94.4 %)
Ultrasound Guidance Combined with ECG/Others	284 (7.8 %)/153 (4.2 %)	15 (2.6 %)/18 (2.9 %)
Puncture blood vessel:		
Basilic vein/Brachial vein	3321 (91.3 %)/222 (6.1 %)	543 (93.3 %)/28 (4.8 %)
Cephalic vein/Others	65 (1.8 %)/30 (0.9 %)	132 (22.6 %)/8 (1.5 %)
Puncture complications:		
None	3438 (94.5 %)	572 (98.1 %)
Ectopic catheter/Others	170 (4.7 %)/30 (0.9 %)	9 (1.7 %)/1 (0.2 %)
Positioning method of catheter tip:		
X-ray	2576 (70.8 %)	552 (94.8 %)
CG + X ray/Others	327 (9.0 %)/735 (20.2 %)	18 (3.2 %)/12 (2.0 %)
Catheter-to-vessel diameter ratio:		

(continued on next page)

Table 1 (*continued*)

Variable	Non-PICC-RVT (n = 3638)	PICC-RVT (n = 582)
Positive pressure joint	3132 (86.1 %)	422 (72.5 %)
Heparin cap/Others	440 (12.1 %)/66 (1.8 %)	151 (26.0 %)/9 (1.5 %)
Catheter fixation method:		
Transparent dressing	3558 (97.8 %)	544 (93.5 %)
Hold high the platform/Others	10 (0.3 %)/70 (1.9 %)	29 (5.0 %)/9 (1.5 %)
Type of disinfectant:		
Dextrose Chlorhexidine Ethanol Solution	728 (20.7 %)	122 (21.0 %)
Iodine Tincture/Others	2878 (79.1 %)/32 (0.2 %)	460 (79.0 %)/0 (0.0 %)
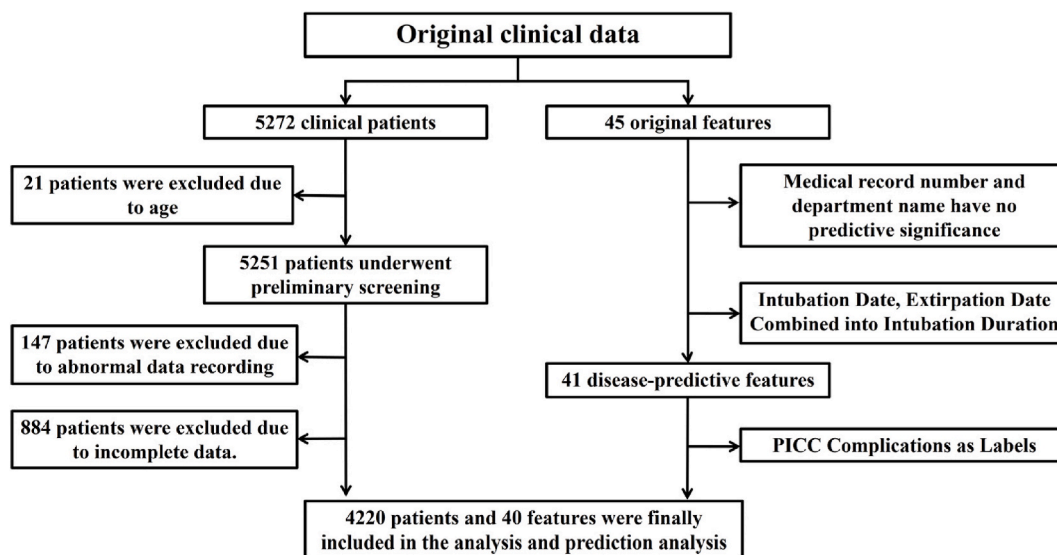
Fig. 2. Inclusion and exclusion chart.

data incompleteness. Initially, the dataset contained 45 features. After the data cleansing process, five features were removed, namely: patient record number, departmental designation, catheter instillation date, date of removal, and the reason for extraction. The dates indicating catheter insertion and removal were replaced by catheter dwell duration. Both the medical record identifiers and departmental tags, often used as patient identifiers within hospitals, were deemed non-predictive. The main variable for prediction was PICC-related complications. After these adjustments, 40 patient features remained, detailed further in Table 1. Patient-specific information, such as biological characteristics and prior medical histories, was sourced from medical records or inquiry files. Cancer staging data came from clinical and pathological assessments at diagnosis. Data focused on catheter-specific features originated from surgical records and notes taken after catheter insertion (Fig. 2).

Out of the initial cases, 21 patients were excluded due to age, while 147 were excluded because of data anomalies. Further, due to heterogeneity in data recording among participating centers, another 884 patients were omitted because of incomplete or incorrect information provided during the study. In terms of patient features, we removed the medical record number and department name since they lacked research significance. We combined the dates of intubation and extubation to form new features. Additionally, PICC complications were utilized as prediction labels instead of being used as model training features.

A total of 4,220 patients were analyzed in this study, with their demographic details outlined in Table 1. Among these participants, 582 experienced thrombosis. The participants consisted of 1,442 males (33.87 %) and 2,781 females (66.12 %), and they had an average age of 52.47 ± 12.26 years. Notably, 677 patients (16 %) had attained an education level of high school or higher. The majority, 3,893 (92.2 %), were diagnosed with malignant tumors, and among them, 3,096 (73.3 %) were in stage III or beyond. Furthermore, 205 participants (5.94 %) had prior experience with central venous catheterization, while 41 (1.18 %) had previously faced deep vein thrombosis and thrombophlebitis.

3.2. Statistical analysis results

Upon statistical analysis, certain variables exhibited pronounced statistical significance. Specifically, variables including Gender, Age, Height, Activated partial thromboplastin time, Albumin, Education level, Primary diagnosis upon admission, Tumor staging, Chemotherapy cycles, Catheterization method, Connector type, Catheter fixation method, and prestored sealing fluid had notable statistical differences with $p < 0.001$. Meanwhile, other factors such as Weight, Platelets, Fibrinogen, C-reactive protein, Catheter-to-

Table 2
Univariate analysis of the risk factors of PICC-RVT.

Variable	Univariate analysis		
	Odds ratio	95 % CI	*P*-value
Age	1.025	1.016–1.033	<0.001
Height	1.034	1.019–1.047	<0.001
Weight	1.010	1.002–1.019	0.017
White blood cell	0.985	0.950–1.014	0.375
Hemoglobin	1.002	0.999–1.004	0.319
Platelet	0.998	0.996–0.999	0.020
D-dimer	0.998	0.966–1.026	0.411
Prothrombin time	0.960	0.930–0.991	0.072
Fibrinogen	1.053	0.990–1.124	0.017
Activated partial prothrombin time	0.962	0.946–0.977	<0.001
Albumin	1.025	1.019–1.032	<0.001
C reactive protein	1.000	0.997–1.00	0.589
Catheter vessel diameter ratio	1.503	0.621–3.464	0.046
Retention time	0.998	0.996–1.000	0.029
Position of catheter tip	1.020	0.974–1.077	0.402
Gender	0.376	0.308–0.452	<0.001
Hypertension	0.740	0.562–0.939	0.091
Diabetes	0.811	0.526–1.156	0.451
Coronary heart disease	0.256	0.209–0.320	0.176
History of deep vein thrombosis and thrombophlebitis	1.311	0.590–2.398	0.066
History of deep vein catheterization	1.543	1.109–2.126	0.011
Education	0.789	0.721–0.858	<0.001
Main diagnosis on admission	0.921	0.889–0.953	<0.001
Tumor staging	1.165	1.056–1.299	<0.001
Reason for catheter retention	0.936	0.780–1.093	0.268
Chemotherapy cycle	1.002	0.971–1.029	<0.001
PICC Catheter Models	1.134	0.868–1.935	0.103
PICC Catheter Lumen Number	0.652	0.497–0.874	0.536
Type of PICC	0.527	0.459–0.595	<0.001
PICC catheterization method	0.652	0.514–0.804	<0.001
Puncture site	1.119	0.922–1.342	0.621
Number of punctures	1.265	1.015–1.731	0.03
Blood vessel puncture	0.928	0.928–1.165	0.071
Puncture complications	0.828	0.614–1.115	0.01
Catheter Tip Positioning Method	1.412	1.243–1.604	0.01
Connector type	1.698	1.368–2.086	<0.001
Catheter fixation method	2.440	1.393–5.464	<0.001
Prestored sealing fluid	0.351	0.255–0.456	<0.001
Sealing liquid prefilling type	1.440	0.720–2.582	0.047
Type of disinfectant	1.122	0.892–1.427	0.737

vessel diameter ratio, Days of catheter retention, History of deep vein catheterization, Catheter model, Number of punctures, and Method of catheter tip positioning showed significant associations with P < 0.05. These findings are detailed in Table 2.

3.3. Model performance comparison

In our study, we plotted the models' training accuracy and loss curves during training (Fig. 3A), as well as a heat map of factors affecting prediction results during testing (Fig. 3B). We highlighted our PRAD prediction model, which achieved the highest AUC value, and compared it with the next three best-performing models: RF, AE, and SVM. Specifically, in Fig. 4, we present the AUC training curves of the PRAD prediction model (Fig. 4a) and the top three models—RF (Fig. 4b), AE (Fig. 4c), and SVM (Fig. 4d).

The figure shows the training curve, validation curve, 5-fold cross-validation (CV) curve, and test curve. These include PRAD operating characteristic curve, random forest operating characteristic curve, AE operating characteristic curve, and SVM operating characteristic curve.

Delving into variable analysis, we employed the cooperative game theory for intuitive visualization interpretation of PRAD. Within this framework, predictive factors are considered as "contributors." For each sample, the model generates a prediction value, represented by SHAP (SHapley Additive exPlanation value) values assigned to each feature[30]. These values reveal the positive and negative impact of each feature in the sample (Fig. 5). The results showed that factors like albumin levels, primary diagnosis, hemoglobin levels, platelets levels, and education level significantly influence the onset of PICC-RVT.

We highlighted the modeling methods with the highest AUC values for each ML algorithm in the test set (Table 3). To ensure clarity, we adopted a systematic naming convention for the preprocessing of our model: normalization_preprocessing_feature-selection. This approach standardizes the process of data preprocessing.
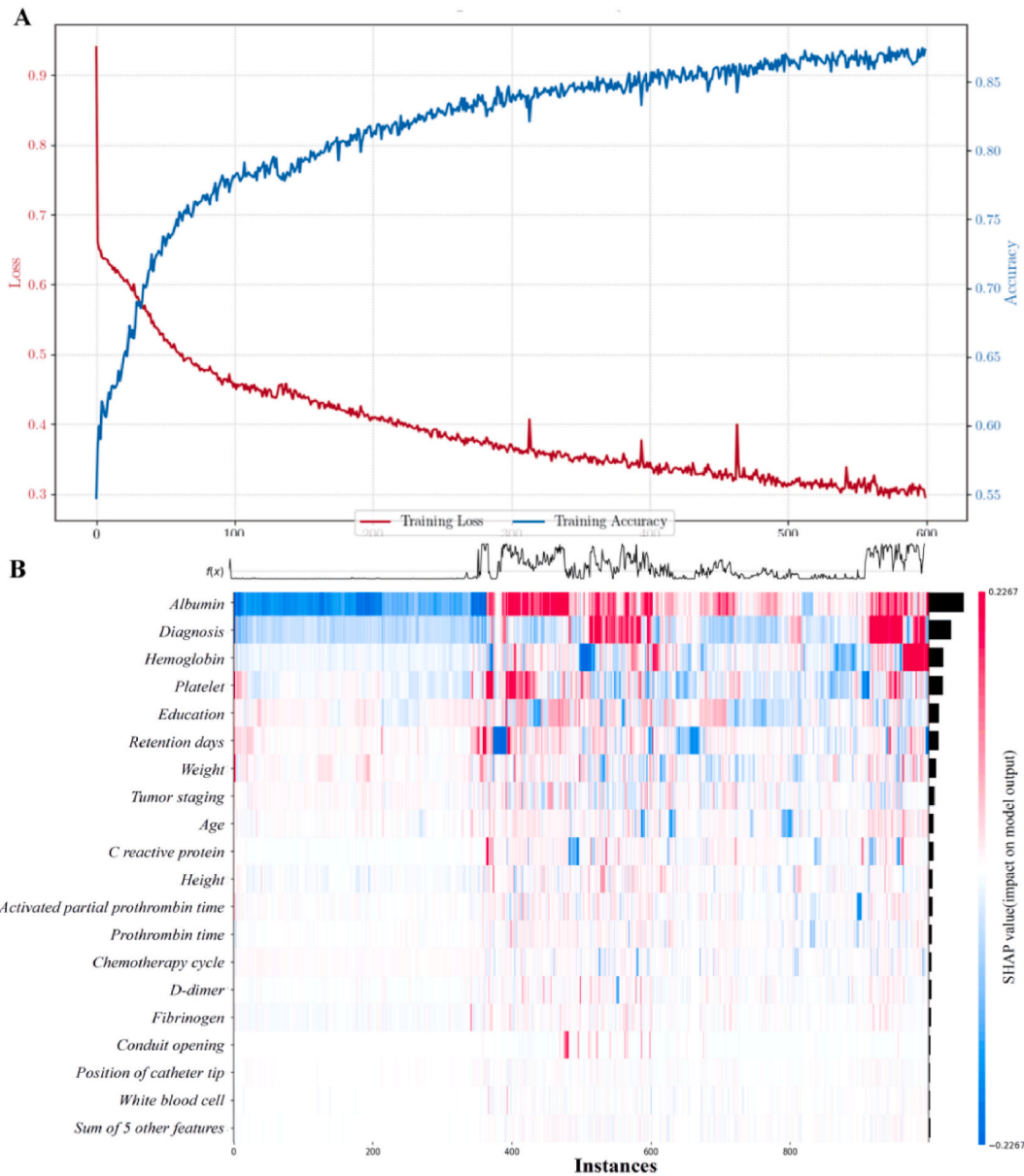
Fig. 3. A represents the PRAD model training diagram and B represents the heat map of the influencing factors of the PRAD model on the test set.

4. Discussion

The PICC is globally recognized as a cornerstone for intravenous infusion. Current studies reveal a significant variation in the incidence rate of PICC-RVT, ranging from 2 % to 67 %. For critically ill patients, this rate can soar to a distressing 91 %. Over time, the risk escalates, with complications becoming increasingly likely the longer the catheter remains in place [17]. If left untreated, these complications can progress to post-thrombotic syndrome or even life-threatening pulmonary embolism. Specifically in cancer patients, mortality rates attributable to PICC-RVT vary between 12 % and 34 %, marking it as the second leading cause of death following the cancer itself [18,19]. In an ideal scenario, regular ultrasonic monitoring would preemptively address PICC-RVT. However, due to staffing constraints, equipment availability, and budgetary challenges, this optimal strategy often proves unfeasible in real-world settings [20]. The key lies in the early detection and specialized care for high-risk patients, aiming to minimize the impact of PICC-RVT. Consequently, there exists an urgent need for dependable models capable of pinpointing these high-risk individuals. Crafting such diagnostic tools could revolutionize patient care, guaranteeing improved outcomes and a more judicious allocation of medical resources.

Based on differences in thrombosis incidence between Asian and Western populations, including genetic susceptibility,
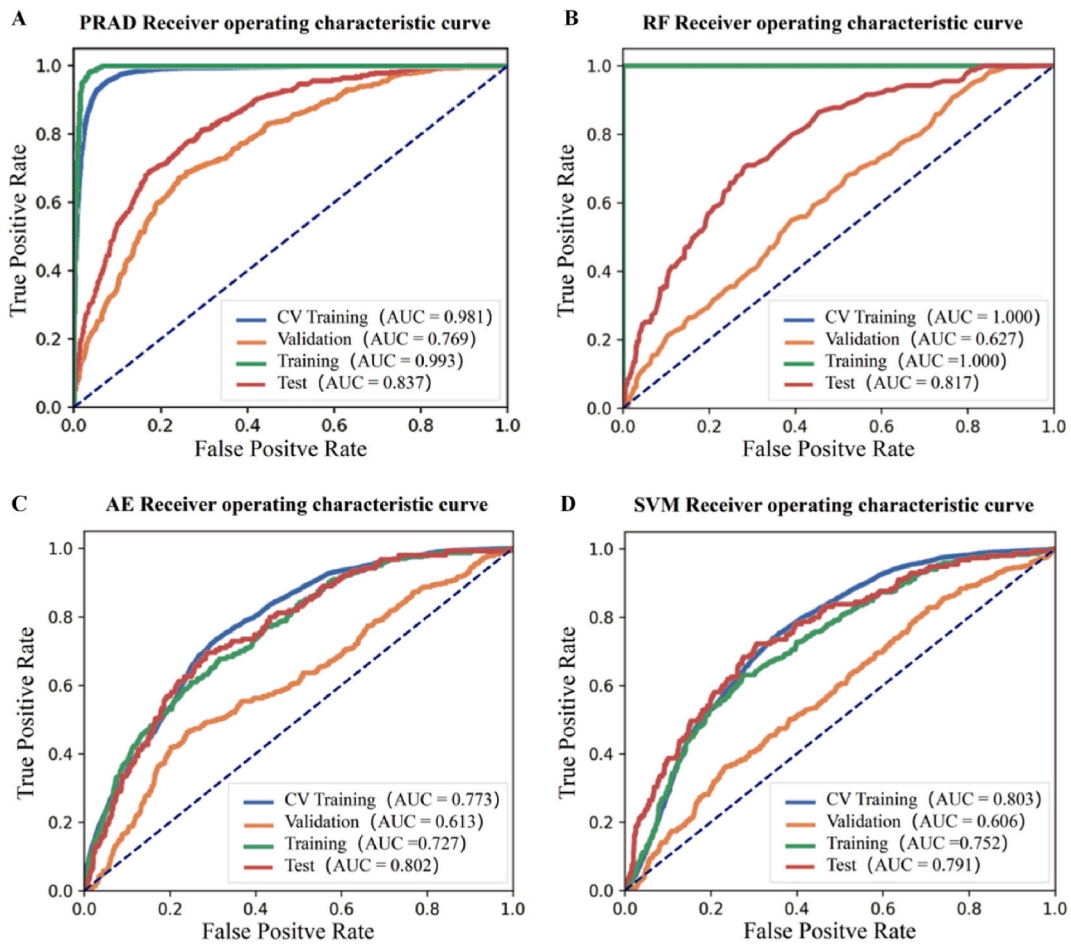
Fig. 4. Model training curves for the top 3 prediction performance of the PRAD model and other machine learning models.

underdiagnosis, and asymptomatic VTE, as noted by Wang K L et al. [21], we have tailored our study to the distribution characteristics of the Chinese population. We have presented data from 27 medical centers across northern, southern, and southwestern China, representing over 80 % of densely populated cities, covering 5,272 PICC-treated patients. Originally, each patient has been recorded with 45 unique features covering diverse demographics. Data from 4,220 patients with 40 features were included in the data analysis, where 582 patients has been identified with PICC-RVT. A total of 14,885 machine learning models have been generated for comparison during analysis, with the DNN predictive model showing the best performance (AUC = 0.837, Accuracy = 86.4 %). The results of this paper highlight the potential of the proposed prognosis model in the prognosis of PICC-RVT. Our approach have the potential to revolutionize patient stratification, personalized treatment, resource allocation, and the management of PICC-RVT, thereby improving patient outcomes and healthcare system efficiency.

Previous research has introduced predictive models for PICC-RVT with varied success. These models often faced limitations due to factors like being based on data from a single medical center, constraints of intensive care equipment, or relying on consistent pathological diagnoses. In contrast, our study utilizes a more expansive dataset encompassing a broader range of patient attributes. Our introduction of the DNN model to forecast PICC-RVT outcomes represents a significant advancement in this field. A study by Li S et al., somewhat paralleling our approach, gathered data from 348 cancer patients undergoing PICC treatment. They employed ML models to predict PICC-RVT outcomes and achieved an AUC of 0.787, outperforming the conventional Seeley scoring system. Nonetheless, it's vital to note that Liu S et al.'s research was restricted to one medical institution. When juxtaposing performance metrics, our study demonstrates a superior AUC and Sen compared to Liu S et al. (AUC: 0.837 vs 0.809; Sen: 85.6 % vs 71.4 %). Another noteworthy study by Fu Jie [22] and team collected data from 1,844 breast cancer patients receiving PICC therapy. They developed an artificial neural network with a singular hidden layer containing 18 neurons for thrombosis prediction, relying on merely 8 features. By contrast, our model discerningly selects features, effectively sidelining irrelevant ones by allocating them a zero weight. Given these distinctions, our study appears to offer improved predictive accuracy, flexibility, and wider relevance. Considering the constrained data volume and unique data sources in earlier research, their findings might be prone to data inconsistencies, casting doubt on the universal applicability of their models. For a comprehensive comparison of PICC-RVT predictive models, readers can refer to Table 4.

In our study, the identified independent predictive factors for PICC-RVT was consistent with previous research [8–10]. However,
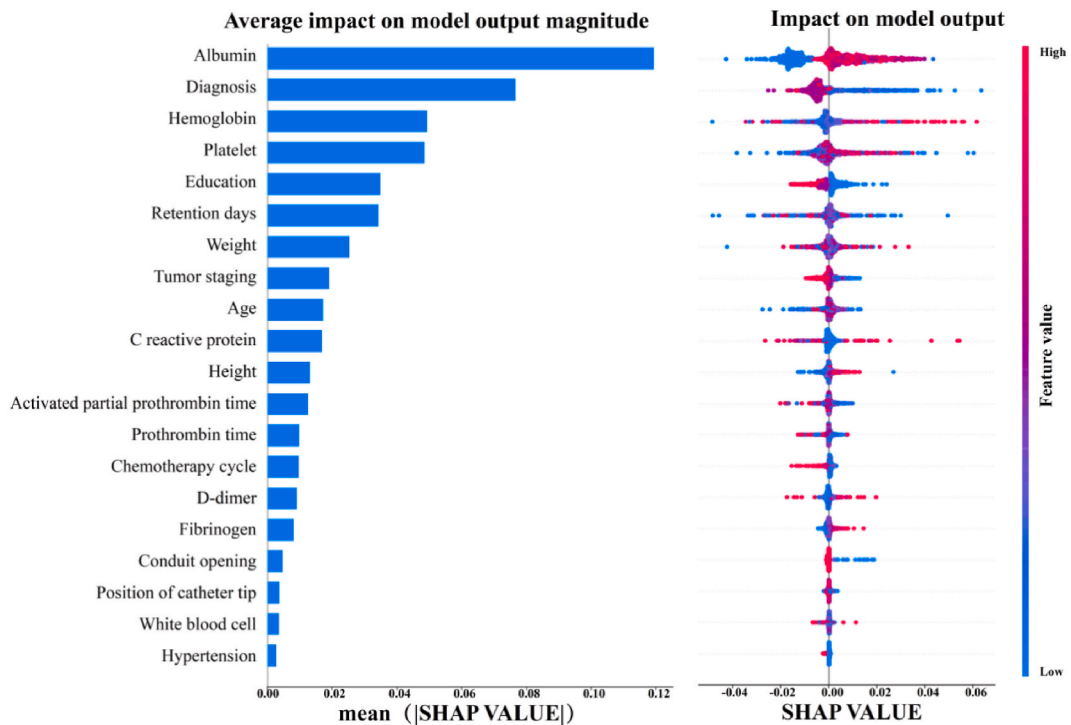
Fig. 5. Visualization of PRAD prediction of PICC-RVT related factors using SHAP method on all data.

Table 3
Performance comparison of various machine learning models in this study.

Model	Preprocessing	Features	AUC	Acc (%)	Sen (%)	Spe (%)
PRAD	–	40	0.837	86.44	85.67	79.33
RF	Mean + PCC + ANOVA	29	0.813	75.74	77.48	67.69
AE	Mean + PCC + ANOVA	28	0.802	73.53	76.77	63.46
SVM	Z-score + PCC + KW	27	0.791	72.87	72.26	69.46
NB	Mean + PCC + RFE	22	0.772	69.58	85.16	55.09
LR-Lasso	Z-score + PCA + KW	11	0.764	64.97	75.48	63.12
GP	Mean + PCA + ANOVA	16	0.757	71.70	71.61	71.72
AdaBoost	Z-score + PCA + KW	8	0.756	75.77	62.58	77.83
LR	Z-score + PCA + ANOVA	14	0.751	69.87	70.32	69.80
LDA	Mean + PCA + ANOVA	14	0.751	69.68	72.26	69.32
DT	MinMax + PCC + RFE	3	0.708	67.66	66.45	67.87

PRAD,PICC-RT Assessment via Deep-learning; AUC, the area under the receiver operating characteristic; Acc, Accuracy; Sen, Sensitivity; Spe, Specificity; CI, Confidence Interval; RF, Random Forest; AE, Autoencoder; SVM, Support Vector Machines; NB, Naive Bayes; LR,Logistic Regression; GP, Gaussian process; LR,Logistic Regression; LDA,Linear Discriminant Analysis; DT,Decision tree.

factors like the number of catheter lumens, puncture sites, and disinfectant types showed weak correlations, thus having limited impact on the predictions of PICC-RVT. In our deeper exploration of the model, we employed cooperative game theory to investigate the impact of various predictive factors on PICC-RVT predictions. Within this framework, each predictive factor is considered an underlying "contributor." For every prediction sample, the model generates a value, represented by SHAP values assigned to each feature in the sample (Fig. 5) [12,27]. This approach quantitatively evaluated the contribution of each feature to the model's prediction for PICC-RVT. Our research findings indicate that albumin is a major influencing factor in PICC-RVT, with higher albumin levels correlating with increased thrombosis probability [28]. Previous studies also suggest that albumin may reduce antiplatelet aggregation, thereby elevating thrombosis risk [29]. Hemoglobin and platelets show similar trends, where increased levels potentially elevated blood viscosity and thrombosis risk [30,31]. For the key factor of diagnosis, we observed major impacts from the circulatory and hematological systems during data processing which were assigned as 3 and 4 using one hot encoder method(please see Fig. 5, second bar under Diagnosis). Interestingly, the PRAD model places significant emphasis on education level, suggesting that higher education correlates with a lower probability of PICC-RVT, possibly due to better patient-doctor cooperation and clearer communication of health status, leading to more timely and accurate diagnoses [22]. However, it is important to note that the majority of analyzed individuals had no education or no recorded education (61.8 % for Non-PICC-RVT and 74.9 % for PICC-RVT). This very unbalanced

Table 4
Comparison among the prediction models of PICC-RT.

Research Group	Predicted Performance	Catheterization/ Thrombosis	Patient Category	Data Source	Number of risk factors	Prediction Method
Song X. et al. [12]	AUC = 0.579–0.822 Sen = 0.729	339/56	PICC therapy for cancer	Single center (National Cancer Center of china)	3–5	Logistic regression
Lin S. et al. [21]	AUC = 0.757	647/244	PICC therapy for cancer	Single center (Ningbo First Hospital)	36	Logistic regression
Liu S. et al. [23]	AUC = 0.809 Sen = 0.71, Spe = 0.90	348/-	PICC therapy for cancer	Single center (West China Hospital)	8	Machine learning
Lin Y. et al. [24]	AUC = 0.615 Sen = 0.74, Spe = 0.44	201/108	PICC therapy for cancer	Single center (First Affiliated Hospital of Fujian Medical University)	15	Caprini
Yue J. et al. [25]	AUC = 0.808 Sen = 0.63, Spe = 0.73	117/19	PICC therapy for hematologic malignancies	Single center (Mianyang Centarl Hospital)	7	Logistic regression
Fu J. et al. [26]	AUC = 0.742, Acc = 0.715 Sen = 0.727, Spe = 0.71	1844/256	PICC therapy for breast cancer	Single center (Fujian Medical University Union Hospital)	8	ANN
Ours	*AUC = 0.837, Acc = 0.864 Sen = 0.856, Spe = 0.793*	*4220/582*	*Any PICC therapy*	*27 medical centers (Supplementary Material* Table S1)	*40*	*DNN*

PICC, Peripherally inserted central catheter; AUC, the area under the receiver operating characteristic; ANN, Artificial Neural Network; Acc, accuracy; Sen, Sensitivity; Spe, Specificity.DNN, Deep Neural Network.

distribution might have significantly contributed to education level emerging as one of the main factors for PICC-RVT onset.

This study provides valuable insights into risk factors associated with PICC-RVT progression. With these findings, healthcare practitioners can better guide patients, empowering them to manage potential risks more effectively. Established research confirms that timely and standardized thrombosis interventions can maintain catheter retention and curb the recurrence of thromboembolism. Our developed PICC-RVT predictive model boasts an impressive accuracy rate of 86.4 % in identifying high-risk patients. Suitable for all patients undergoing PICC insertion, this model enables continuous, quantitative, and objective monitoring, aiding in the early identification of complications and ensuring swift intervention. The widespread adoption of this model has the potential to elevate patient care standards, yielding enhanced outcomes and streamlining healthcare operations. In resource-limited settings, where access to advanced imaging, clinical tests, and expert pathology might be limited, PRAD offers a valuable online tool for monitoring potential PICC-RVT cases. For pathologists, PRAD aids in quickly identifying top-tier differential diagnoses related to PICC-RVT risks. This then guides the subsequent necessary ancillary testing or clinical correlations to finalize a diagnosis. The workflow supported by PRAD brings several benefits. It reduces reliance on clinical tests, speeds up detection, and offers standardized predictive accuracy. PRAD, in our view, can significantly accelerate the diagnosis of complex PICC-RVT cases, conserving precious resources in the process. Furthermore, when the risks of PICC-RVT are unclear, PRAD serves as a vigilant online monitor. If its high-confidence predictions differ from clinical evaluations, it prompts healthcare professionals to reevaluate. Such divergence might also lead to the consideration of other overlooked hypotheses.

5. Conclusion

Our study sourced data from 27 medical institutions, encompassing a diverse group of 5,272 patients, each characterized by 45 unique features. Upon refining the dataset, we closely analyzed 4,220 patients based on 40 specific attributes. Following an extensive assessment of pre-existing predictive models and the creation of 14885 comparative models, the deep-learning-based PRAD model demonstrated superior performance, and achieved an AUC of 0.837 and an accuracy of 86.4 %. These results lay the groundwork for the prognosis of PICC-RVT, which could contribute to streamlined clinical monitoring and prompt intervention for patients with high PICC-RVT risks. Deploying such a model has the potential to significantly improve patient care quality and the efficiency of resource utilization in the healthcare system.

Funding

CRediT authorship contribution statement

Yue Li: Writing – review & editing, Writing – original draft, Supervision, Methodology. **Ting Li:** Funding acquisition, Formal analysis. **Hengjie Su:** Writing – review & editing. **Xin Zhang:** Writing – review & editing. **Jiangbo Pu:** Writing – review & editing. **Hong Sun:** Data curation. **Qiong Liu:** Writing – original draft, Supervision, Methodology. **Bowen Zhang:** Writing – original draft. **Biao Sun:** Methodology, Funding acquisition. **Jia Li:** Writing – review & editing. **Xinxin Yan:** Writing – review & editing. **Laiyou Wang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e39178.

References

[1] G.J. Schears, N. Ferko, I. Syed, et al., Peripherally inserted central catheters inserted with current best practices have low deep vein thrombosis and central line–associated bloodstream infection risk compared with centrally inserted central catheters: a contemporary meta-analysis, J. Vasc. Access 22 (1) (2021) 9–25.

[2] S.O. Trerotola, S.W. Stavropoulos, J.I. Mondschein, et al., Triple-lumen peripherally inserted central catheter in patients in the critical care unit: prospective evaluation, Radiology 256 (1) (2010) 312–320, https://doi.org/10.1148/radiol.10091860.

[3] L. Pan, Q. Zhao, X. Yang, Risk factors for venous thrombosis associated with peripherally inseRVTed central venous catheters, Int. J. Clin. Exp. Med. 7 (12) (2014) 5814.

[4] L.E. Flinterman, F.J.M. Van Der Meer, F.R. Rosendaal, et al., Current perspective of venous thrombosis in the upper extremity, J. Thromb. Haemostasis 6 (8) (2008) 1262–1266, https://doi.org/10.1111/j.1538-7836.2008.03017.x.

[5] L.B. Kreuziger, J. Jaffray, M. Carrier, Epidemiology, diagnosis, prevention and treatment of catheter-related thrombosis in children and adults, Thromb. Res. 157 (2017) 64–71, https://doi.org/10.1016/j.thromres.2017.07.002.

[6] J. Kang, W. Chen, W. Sun, et al., Peripherally inserted central catheter-related complications in cancer patients:a prospective study of over 50,000 catheter days, J. Vasc. Access 18 (2) (2017) 153–157, https://doi.org/10.5301/jva.5000670.

[7] M.I. Jordan, T.M. Mitchell, Machine learning:Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260, https://doi.org/10.1126/science. aaa8415.

[8] V. Chopra, S. Anand, A. Hickner, et al., Risk of venous thromboembolism associated with peripherally inserted central catheters:a systematic review and meta-analysis, Lancet 382 (9889) (2013) 311–325, https://doi.org/10.1016/S0140-6736(13)60592-9.

[9] V. Chopra, S. Kaatz, A. Conlon, et al., The Michigan Risk Score to predict peripherally inserted central catheter-associated thrombosis, J. Thromb. Haemostasis 15 (10) (2017) 1951–1962, https://doi.org/10.1111/jth.13794.

[10] H. Chen, L. Tao, X. Zhang, et al., The effect of systemic and local risk factors on triggering peripherally inserted central catheter–related thrombosis in cancer patients:A prospective cohort study based on ultrasound examination and structural equation modeling, Int. J. Nurs. Stud. 121 (2021) 104003, https://doi.org/10.1016/j.ijnurstu.2021.104003.

[11] D. Kort, N. van Rein, F.J.M. van der Meer, et al., Relationship between neighborhood socioeconomic status and venous thromboembolism:results from a population-based study, J. Thromb. Haemostasis 15 (12) (2017) 2352–2360, https://doi.org/10.1111/jth.13868.

[12] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186, https://doi.org/10.1145/3375627.3375830.

[13] Y. Song, J. Zhang, Y. Zhang, Y. Hou, X. Yan, Y. Wang, M. Zhou, Y. Yao, G. Yang, FeAture Explorer (FAE): a tool for developing and comparing radiomics models, PLoS One (2020), https://doi.org/10.1371/journal.pone.0237587.

[14] S.F. Sung, L.C. Hung, Y.H. Hu, Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning, Int. J. Med. Inf. 152 (2021) 104505, https://doi.org/10.1016/j.ijmedinf.2021.104505.

[15] A. Lopez-Del Rio, S. Picart-Armada, A. Perera-Lluna, Balancing data on deep learning-based proteochemometric activity classification, J. Chem. Inf. Model. 61 (4) (2021) 1657–1669, https://doi.org/10.1021/acs.jcim.1c00086.

[16] Y. Wu, Y. Fang, Stroke prediction with machine learning methods among older Chinese, Int. J. Environ. Res. Publ. Health 17 (6) (2020) 1828, https://doi.org/10.3390/ijerph17061828.

[17] F.T.M. Bosch, M.D. Nisio, H.R. Büller, et al., Diagnostic and therapeutic management of upper extremity deep vein thrombosis, J. Clin. Med. 9 (7) (2020) 2069, https://doi.org/10.3390/jcm9072069.

[18] T.K. Liem, K.E. Yanit, S.E. Moseley, G.J. Landry, T.G. Deloughery, C.A. Rumwell, E.L. Mitchell, G.L. Moneta, Peripherally inserted central catheter usage patterns and associated symptomatic upper extremity venous thrombosis, J. Vasc. Surg. 55 (3) (2012 Mar) 761–767, https://doi.org/10.1016/j.jvs.2011.10.005. PMID: 22370026.

[19] V. Zochios, I. Umar, N. Simpson, N. Jones, Peripherally inserted central catheter (PICC)-related thrombosis in critically ill patients, J. Vasc. Access 15 (5) (2014 Sep-Oct) 329–337, https://doi.org/10.5301/jva.5000239. Epub 2014 Apr 25.

[20] J. Willan, H. Katz, D. Keeling, The use of artificial neural network analysis can improve the risk-stratification of patients presenting with suspected deep vein thrombosis, Br. J. Haematol. 185 (2) (2019) 289–296, https://doi.org/10.1111/bjh.15780.

[21] K.L. Wang, E.S. Yap, S. Goto, et al., The diagnosis and treatment of venous thromboembolism in Asian patients, Thromb. J. 16 (2018) 1–12.

[22] J. Fu, W. Cai, B. Zeng, et al., Development and validation of a predictive model for peripherally inserted central catheter-related thrombosis in breast cancer patients based on artificial neural network:A prospective cohort study, Int. J. Nurs. Stud. 135 (2022) 104341, https://doi.org/10.1016/j.ijnurstu.2022.104341.

[23] S. Peng, T. Wei, X. Li, et al., A model to assess the risk of peripherally inserted central venous catheter-related thrombosis in patients with breast cancer: a retrospective cohort study, Support. Care Cancer 30 (2) (2021) 1127–1137, https://doi.org/10.1007/s00520-021-06073-4.

[24] S. Liu, F. Zhang, L. Xie, et al., Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer, Int. J. Med. Inf. 129 (2019) 175–183, https://doi.org/10.1016/j.ijmedinf.2019.06.001.

[25] Y. Lin, Z. Zeng, R. Lin, et al., The Caprini thrombosis risk model predicts the risk of peripherally inserted central catheter-related upper extremity venous thrombosis in patients with cancer, Journal of Vascular Surgery:Venous and Lymphatic Disorders 9 (5) (2021) 1151–1158, https://doi.org/10.1016/j.jvsv.2020.12.075.

[26] J. Yue, Y. Zhang, F. Xu, et al., A clinical study of peripherally inserted central catheter-related venous thromboembolism in patients with hematological malignancies, Sci. Rep. 12 (1) (2022) 9871, https://doi.org/10.1038/s41598-022-13916-5.

[27] D. Janzing, L. Minorics, P. Bloebaum, Feature relevance quantification in explainable AI: a causal problem, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics 108 (2020) 2907–2916, https://doi.org/10.48550/arXiv.1910.13413. PMLR.

[28] A. Gigante, B. Barbano, L. Sardo, et al., Hypercoagulability and nephrotic syndrome, Curr. Vasc. Pharmacol. 12 (3) (2014) 512–517, https://doi.org/10.2174/157016111203140518172048.

[29] Y. Song, L.A. Bienvenu, V. Bongcaron, et al., Platelet-targeted thromboprophylaxis with a human serum albumin fusion drug: preventing thrombosis and reducing cardiac ischemia/reperfusion injury without bleeding complications, Theranostics 14 (8) (2024) 3267–3281, https://doi.org/10.7150/thno.97517.

[30] M. Warny, J. Helby, H.S. Birgens, et al., Arterial and venous thrombosis by high platelet count and high hematocrit: 108,521 individuals from the Copenhagen General Population Study, J. Thromb. Haemostasis 17 (11) (2019) 1898–1911, https://doi.org/10.1111/jth.14555.

[31] V. Nair, S. Singh, M.Z. Ashraf, et al., Epidemiology and pathophysiology of vascular thrombosis in acclimatized lowlanders at high altitude: a prospective longitudinal study, The Lancet Regional Health-Southeast Asia (2022) 3, https://doi.org/10.1016/j.lansea.2022.100043.