



Research article

Machine learning predicts cancer-associated venous thromboembolism using clinically available variables in gastric cancer patients

Qianjie Xu^{a,1}, Haike Lei^{b,1}, Xiaosheng Li^{b,1}, Fang Li^b, Hao Shi^b, Guixue Wang^c, Anlong Sun^{b,**}, Ying Wang^{b,***}, Bin Peng^{a,*}^a Department of Health Statistics, School of Public Health, Chongqing Medical University, Chongqing, 400016, China^b Chongqing Cancer Multi-omics Big Data Application Engineering Research Center, Chongqing University Cancer Hospital, Chongqing, 400030, China^c MOE Key Lab for Biorheological Science and Technology, State and Local Joint Engineering Laboratory for Vascular Implants, College of Bioengineering Chongqing University, Chongqing, 400030, China

ARTICLE INFO

Keywords:

Gastric cancer
Venous thromboembolism
Prediction model
Machine learning

ABSTRACT

Stomach cancer (GC) has one of the highest rates of thrombosis among cancers and can lead to considerable morbidity, mortality, and additional costs. However, to date, there is no suitable venous thromboembolism (VTE) prediction model for gastric cancer patients to predict risk. Therefore, there is an urgent need to establish a clinical prediction model for VTE in gastric cancer patients. We collected data on 3092 patients between January 1, 2018 and December 31, 2021. And after feature selection, 11 variables are reserved as predictors to build the model. Five machine learning (ML) algorithms are used to build different VTE predictive models. The accuracy, sensitivity, specificity, and AUC of these five models were compared with traditional logistic regression (LR) to recommend the best VTE prediction model. RF and XGB models have selected the essential characters in the model: Clinical stage, Blood Transfusion History, D-Dimer, AGE, and FDP. The model has an AUC of 0.825, an accuracy of 0.799, a sensitivity of 0.710, and a specificity of 0.802 in the validation set. The model has good performance and high application value in clinical practice, and can identify high-risk groups of gastric cancer patients and prevent venous thromboembolism.

1. Introduction

Gastric cancer (GC) is one of China's most common malignant tumors with high morbidity and mortality. It is also the third leading cause of cancer-related death worldwide [1], causing a severe burden to people worldwide. Venous thromboembolism (VTE) refers to the abnormal clotting of blood in the vein, which makes the blood vessels entirely or incompletely blocked. It is a venous return disorder, including deep vein thrombosis (DVT) and pulmonary thromboembolism (PT) [2]. It is a prevalent and potentially fatal

* Corresponding author.

** Corresponding author.

*** Corresponding author.

E-mail addresses: 387714294@qq.com (A. Sun), 13996412826@163.com (Y. Wang), pengbin@cqmu.edu.cn (B. Peng).¹ Contributions to this work have been equally between the authors, with each sharing first authorship.<https://doi.org/10.1016/j.heliyon.2022.e12681>

Received 7 October 2022; Received in revised form 20 December 2022; Accepted 21 December 2022

Available online 3 January 2023

2405-8440/© 2022 Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

disease, killing approximately more than 3 million people yearly [3].

Previous studies have confirmed that VTE is a risk factor for many cancers, such as lung cancer [4], pancreatic cancer, and gastroesophageal cancer [5]. VTE is the most common complication in cancer patients and the second leading cause of death in cancer patients [6]. In terms of morbidity, it is estimated that the annual incidence of VTE in cancer patients is as high as 0.5% compared to 0.1% in the general population [7]. In addition, patients receiving hormone therapy, central catheters, and surgery are at higher risk for VTE [8]. Tetzlaff et al. reported that for localized gastroesophageal cancer, survival was significantly longer in patients without VTE events than in those with VTE events (32 months vs. 17.7 months) [9]. Fuentes et al. reported that patients with gastric cancer have a higher risk of developing VTE than other cancer types [10]. And VTE is associated with a worse prognosis for patients with gastric cancer. In a study of 191 patients with advanced gastric cancer treated in clinical trials, those who developed VTE had a significantly shorter OS (3.9 months vs. 8.7 months, $P < 0.01$) [11]. Therefore, it is crucial to correctly identify whether gastric cancer patients have VTE to prolong their survival time.

Several common VTE prediction tools, such as Capri and Khorana scores, have been validated in gastric cancer. They were mainly from the Caucasian database and did not consider surgery, anti-cancer treatment, and supportive care. So these tools are not a good representation of the Chinese population [12,13]. Xu Cheng et al. used multivariate logistic regression and stepwise logistic regression to establish models to predict the occurrence of VTE after robot-assisted radical prostatectomy [14]. However, the sample size of this study was small, and only 351 samples were included. In addition, logistic regression is a very traditional statistical method, which is not fully applicable to the increasingly updated clinical data compared with the new machine learning methods. Machine learning (ML) is a branch of artificial intelligence (AI) that aims to build systems that can learn or improve performance based on the data they use. ML does not refer to a mathematical method in particular but rather to a wide range of statistical models with high flexibility and the ability to discern subtle nonlinear patterns in the data [15]. Nudel et al. constructed artistic neural networks (ANN) and gradient boosting machines (XGB) to predict the occurrence of VTE after bariatric surgery, compared the two models with traditional logistic regression (LR) and found that the prediction accuracy of ANN and XGB was outperformed that of LR [15].

Therefore, the purpose of this study is mainly on two aspects: First, the five ML algorithms of random forest (RF), support vector machine (SVM), back propagation neural network (BPNN), naïve Bayes (NB), and gradient boosting machines (XGB) are used to establish different ML models for predicting the occurrence of VTE in gastric cancer patients. Second, compare these five models with traditional LR in terms of AUC value, Accuracy, Sensitivity, and Specificity, and select the best predictive model.

2. Materials and methods

2.1. Data sources and study population

The data in this study were based on the database of Chongqing University Cancer Hospital, covering almost all diagnosed gastric cancer patients in Chongqing since 2018. A total of 3281 patients with gastric cancer diagnosed from January 1, 2018, to December 31, 2021, were selected. The following information was collected: Patient-associated factors: sex, age, Body Mass Index (BMI), Karnofsky Performance Status (KPS), angiocardopathy, blood transfusion history; Cancer-associated factors: clinical stage, lymph vessel invasion; treatment-associated factors: chemotherapy, central venous catheterization (CVC) via jugular vein, subclavian vein, and femoral

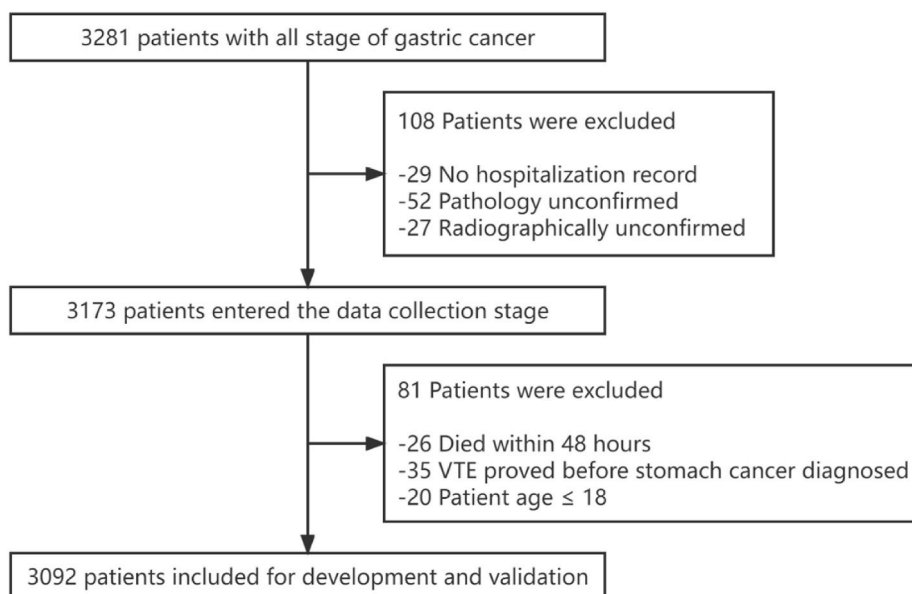


Fig. 1. The flow diagram outlining the search progress.

vein, operation; biomarker: leukocyte count, platelet count, hemoglobin, albumin concentration, serum creatinine, fibrinogen degradation product (FDP), D-dimer. The blood samples were collected from the antecubital vein and stored in vacuum tubes containing EDTA (Ethylene Diamine Tetraacetic Acid). All the blood tests were done at the Chongqing University Cancer Hospital laboratory. The serum prognostic markers under investigation were LDH and β 2-microglobulin. Kit manufacturers state that serum LDH and β 2-microglobulin have normal upper limits of 245 g/L and 2.5 g/L, respectively. Continuous variables such as BMI, white blood cell count, platelet count, hemoglobin, albumin concentration, serum creatinine, FDP, and D-dimer were transformed into categorical variables by specific critical points. Age is divided into three categories by P25 and P75.

2.2. VTE diagnosis

We refer to the VTE diagnostic criteria in the Guidelines for the Prevention and Treatment of Tumor-Associated Venous Thromboembolism (2019 Edition) issued by the Cancer and Thrombosis Expert Committee of the Chinese Society of Clinical Oncology. The diagnosis of VTE mainly includes the diagnosis of DVT and PE. According to the guidelines, we use vascular pressure Doppler ultrasound or venography to diagnose DVT; PE was diagnosed by CT pulmonary angiography (CTPA) or nuclide lung ventilation/perfusion imaging. And there were no false positive cases based on imaging in this study.

2.3. Inclusion and exclusion criteria

The inclusion criteria of this study were (1) Patients aged ≥ 18 years; (2) At least one hospitalization record; (3) Pathology-confirmed GC; (4) Radiographically-confirmed VTE. Exclusive criteria were: (1) VTE proved before GC diagnosis; (2) Died within 48 h after the entry into the hospital; (3) The clinical data are seriously missing and incomplete. After passing the inclusion and exclusion criteria, 3092 patients were included in the model construction, as shown in Fig. 1.

Table 1
Patient demographics and clinical characteristics.

Characteristic	Level	Overall (n = 3092)	No VTE (n = 2987)	VTE (n = 105)	P-value
Sex (%)	Female	868	832(95.85)	36(4.15)	0.183
	Male	2224	2155(96.9)	69(3.1)	
Age (%)	<55	722	708(98.06)	14(1.94)	0.026
	55–70	1652	1593(96.43)	59(3.57)	
	>70	718	686(95.54)	32(4.46)	
BMI (%)	<18.5	580	552(95.17)	28(4.83)	0.078
	18.5–23.9	1852	1792(96.76)	60(3.24)	
	≥ 24	660	643(97.42)	17(2.58)	
KPS (%)	<70	244	222(90.98)	22(9.02)	<0.001
	≥ 70	2848	2765(97.09)	83(2.91)	
Angiocardopathy (%)	NO	2509	2434(97.01)	75(2.99)	0.014
	YES	583	553(94.85)	30(5.15)	
Blood transfusion (%)	NO	2138	2087(97.61)	51(2.39)	<0.001
	YES	954	900(94.34)	54(5.66)	
Clinical stage (%)	I-II	688	676(98.26)	12(1.74)	<0.001
	III	996	979(98.29)	17(1.71)	
	IV	1408	1332(94.60)	76(5.40)	
	NO	2785	2686(96.45)	99(3.55)	
Lymph vessel invasion (%)	YES	307	301(98.05)	6(1.95)	0.193
	NO	2053	1979(96.40)	74(3.60)	
Chemotherapy (%)	NO	1039	1008(97.02)	31(2.98)	0.427
	YES	2955	2863(96.89)	92(3.11)	
CVC (%)	NO	137	124(90.51)	13(9.49)	<0.001
	YES	1206	1179(97.76)	27(2.24)	
Operation (%)	NO	1886	1808(95.86)	78(4.14)	0.006
	YES	182	166(91.21)	16(8.79)	
Leukocyte count (%)	<11	2910	2821(96.94)	89(3.06)	<0.001
	≥ 11	2763	2670(96.63)	93(3.37)	
Platelet count (%)	<350	329	317(96.35)	12(3.65)	0.916
	≥ 350	869	820(94.36)	49(5.64)	
Hemoglobin (%)	<100	2223	2167(97.48)	56(2.52)	<0.001
	≥ 100	225	206(91.56)	19(8.44)	
Albumin concentration (%)	<30	2867	2781(97.00)	86(3.00)	<0.001
	≥ 30	3034	2935(96.74)	99(3.26)	
Serum creatinine (%)	≤ 133	58	52(89.66)	6(10.34)	0.009
	>133	2293	2248(98.04)	45(1.96)	
FDP (%)	≤ 5	799	739(92.49)	60(7.51)	<0.001
	>5	1284	1270(98.91)	14(1.09)	
D-dimer (%)	≤ 0.5	1808	1717(94.97)	91(5.03)	<0.001
	>0.5				

Abbreviation: BMI: body mass index; KPS: karnofsky performance status; CVC: central venous catheterization; FDP: fibrinogen degradation product.

2.4. Feature selection

An excellent clinical prediction model does not require too many predictors, so we optimized the collected data and selected the most appropriate features as predictors to achieve the best balance between model performance and clinical applicability. Feature screening is divided into two steps: First, the data is divided into two groups according to whether the patient has VTE, and the demographic and clinical characteristics of the patients between the two groups are compared (Table 1). Considering the statistical significance and clinical significance, we relax the threshold of P value, select the variables with P value less than 0.2, incorporate them into the stepwise logistic regression, and screen the variables according to the principle of minimum AIC. Finally, we retain model features based on stepwise logistic regression results, literature reviews, and clinical experience. A total of 11 variables including age, angiocardopathy, blood transfusion history, clinical stage, CVC, operation, leukocyte count, albumin concentration, serum creatinine, FDP, and D-dimer were retained as predictors.

2.5. Model development

To ensure a more objective evaluation of the model, we use the 'createDataPartition' function in the 'caret' package to randomly split the data set into a training set and a validation set. The function ensures that the proportion of each factor in the training set and the validation set is the same as that in the original data set. The training set accounts for 70% and the validation set accounts for 30%. The proportion of the training set and the verification set is set to 7:3, which is most suitable for the sample size of this study. Statistical tests showed no significant differences between the two queues ($P > 0.05$). Since VTE events occur in only a small number of samples in the data set, our data may have a class imbalance. In this regard, we intend to use two techniques to deal with imbalanced data. The first type is sample-based resampling (oversampled, undersampled, and mixed). Another class of methods is the ROSE algorithm and SMOTE algorithm, which combines sampling and algorithm techniques. All these methods positively impact solving the problem of imbalanced class data, but in this study, mixed sampling has the best effect. Each model was trained by using ten-fold cross-validation method for hyperparameter tuning, and the optimal hyperparameters and training set data were used to train the final model.

Similarly, in the process of model training, we also investigated the contribution of each feature, that is, the importance score of the predictors included in the model. This score is mainly from two models: RF and XGB. In RF, the importance score of each feature is calculated by how much each feature improves the error rate of the classifier. Finally, the importance score of all trees on the random forest is averaged to get the final score of each feature. In XGB, the SHAP value is used to measure the importance of the feature.

2.6. Model evaluation

In model evaluation, the accuracy rate is often used to judge the ability of the model to predict correctly. However, in the case of class imbalance, the model will predict that all samples will not suffer from VTE. In this case, the model has considerable accuracy but no practical significance or application value. Therefore, we still need to use sensitivity and specificity to judge the model's merits further. A confusion matrix is a cross-table that summarizes the prediction results of a classification model, and in the form of a matrix, the samples in the dataset are cross-summarized according to the two criteria of natural categories and categories predicted by the classification model. According to the confusion matrix, the accuracy, sensitivity, specificity, and other indicators can be calculated to evaluate the effectiveness of the model, and the calculation formula is as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

In addition, we calculated and compared the area under the receiver operating characteristic curve (AUC). Since the AUC can be interpreted as the probability that the model scores higher for randomly selected patients with VTE than for randomly selected patients without VTE. The higher the AUC value, the stronger the generalization of the model, and the higher the clinical application value of the model. When the AUC value is less than 0.5, the model is meaningless.

2.7. Data analysis

Missing values are filled in by the "Mice package" based on multiple interpolations. Enumeration data were described by frequency and percentage and compared by the chi-square test. All models, including RF, SVM, BPNN, NB, XGB and LR, were implemented in R (version 4.2.1; <https://www.r-project.org/>) and R studio (version 2022.07.1-554; <https://rstudio.com/products/rstudio/>). Statistical significance was defined as $P < 0.05$.

3. Results

3.1. Characteristics of subjects

According to the inclusion and exclusion criteria, 3092 patients with gastric cancer were enrolled in this study, of whom 105 developed VTE during hospitalization. Regarding the sociodemographic characteristics of the patients: the male to female ratio was about 7:3; Among all age groups, the proportion of 55–70 years old was the most (53.43%), and the incidence of VTE was the highest in the >70-year-old group. Patients with KPS <70, Angiocardiopathy, and previous blood transfusion had a significantly higher incidence of VTE than other patients. Similarly, patients with TNM stage IV, CVC, and surgery also had a higher incidence of VTE ($P < 0.05$). Among biomarkers, leukocyte count <11, hemoglobin <100, albumin concentration <30, serum creatinine >133, FDP >5, D-dimer >0.5 had a greater proportion of VTE occurrences. The detailed results are shown in [Table 1](#).

3.2. Model performance

The evaluation, calibration, and identification of model performance, as well as the future application of the model, are mainly based on the model's performance on the validation set. During model training, according to the youden's-index maximum principle, find the cut-off value of each model: when the prediction probability is greater than the cut-off, VTE occurs. This cut-off value is applied to the validation set to obtain a confusion matrix for each model on the validation set, as shown in [Table 2](#). Each model's accuracy, sensitivity, and specificity are calculated according to the confusion matrix, as shown in [Table 3](#).

[Table 3](#) and [Figs. 2 and 3](#) show the performance of the different models on the validation set in detail. The SVM model outperforms all other models in terms of AUC. Although RFS were slightly more accurate than SVM, they were much less sensitive and predicted fewer patients with VTE, contrary to our purpose. On the other hand, although XGB has a high sensitivity, its low specificity means that some patients without VTE are incorrectly predicted to have VTE. Therefore, we recommend the SVM model as the best classifier for this VTE prediction task.

In addition to the model performance, we ranked the features in the model training stage according to the feature importance score generated by the RF and XGB models, as shown in [Figs. 4 and 5](#). We combine the scores of the two models and conclude that the five variables of clinical stage, blood transfusion history, D-dimer, age, and FDP contribute the most.

4. Discussion

Currently, there is no VTE prediction model designed explicitly for gastric cancer patients, so it is urgent to establish a VTE clinical prediction model for gastric cancer patients, which has good application value in clinical identification and decision-making. In this study, five simple and practical ML models were developed and validated, all of which incorporated 11 variables of age, angiocardiopathy, blood transfusion history, clinical stage, CVC, operation, leukocyte count, albumin concentration, serum creatinine, FDP, and D-dimer as predictors. These variables are screened from three aspects: patient-related factors, cancer-related factors, and laboratory biomarkers, and can be easily collected in clinical practice.

The performance of the five ML models and the traditional LR model in terms of AUC, accuracy, sensitivity, and specificity was comprehensively considered, and the SVM model was determined as the best prediction model. SVM is of great value in clinical diagnosis, and the probability value of the disease can be calculated according to the relevant information of the patient. Zhou et al. used two machine learning algorithms, k-nearest neighbor (KNN) and support vector machine (SVM), to construct a prognostic prediction model for small samples of patients with advanced schistosomiasis, and recommended SVM as the best model by comparing AUC [16].

SVM is essentially a mathematical model whose basic idea is to look for the optimal separation hyperplane, that is, the support vector, to separate the two types of data [17]. The basic model of the SVM is the linear classifier defined on the feature space with the largest interval, and the learning strategy is interval maximization. The maximum interval distance method is often used to find

Table 2
Validation set confusion matrix.

Model	Actual situation	Predict VTE	Predict NO VTE
Logistic Regression	VTE	23	8
	No VTE	237	659
Random Forest	VTE	14	17
	No VTE	143	753
BP-Network	VTE	24	7
	No VTE	303	593
SVM	VTE	22	9
	No VTE	177	719
XGBoost	VTE	18	13
	No VTE	219	677
Naive Bayes	VTE	22	9
	No VTE	242	654

Table 3
Model validation set evaluation metrics.

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.736(0.706–0.764)	0.742(0.551–0.875)	0.735(0.705–0.764)	0.816(0.757–0.875)
Random Forest	0.827(0.802–0.851)	0.451(0.278–0.637)	0.840(0.814–0.863)	0.784(0.726–0.843)
BP-Network	0.666(0.634–0.696)	0.774(0.585–0.897)	0.662(0.630–0.693)	0.779(0.701–0.858)
SVM	0.799(0.772–0.825)	0.710(0.518–0.851)	0.802(0.775–0.828)	0.825(0.770–0.881)
XGBoost	0.750(0.721–0.777)	0.581(0.393–0.749)	0.756(0.726–0.783)	0.756(0.685–0.827)
Naive Bayes	0.729(0.699–0.758)	0.710(0.518–0.851)	0.730(0.699–0.758)	0.803(0.736–0.870)

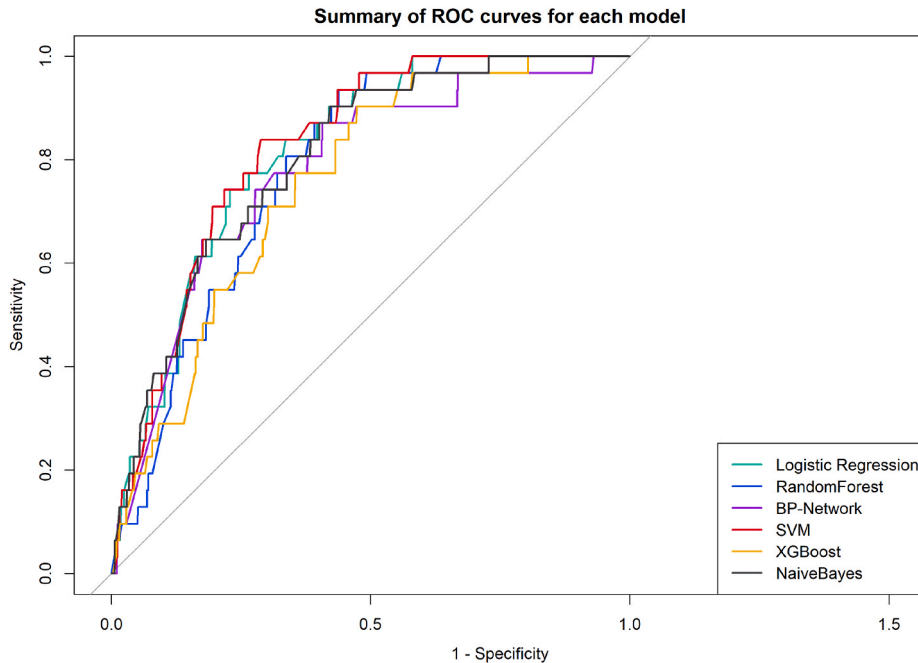


Fig. 2. Summary plot of ROC curves for each model validation set.

support vectors to maximize the interval. The final decision function of the SVM is determined only by a small number of support vectors, and the complexity of the calculation depends on the number of support vectors, not the dimensions of the sample space, which avoids the “dimensionality disaster.” A small number of support vectors determine the final result, which not only helps us grasp key samples and eliminate a large number of redundant samples but also dooms the method to be algorithmically simple and has good robustness [18]. SVM learning problems can be represented as convex optimization problems, so the global minimum of the objective function can be discovered using a known valid algorithm. Other classification methods, such as rule-based classifiers and artificial neural networks, use a strategy based on greedy learning to search for hypothetical spaces, which generally only obtain local optimal solutions [19]. In addition, SVMs can also get better results than other algorithms on small sample training sets, with excellent generalization ability. In other cases, SVM models have good clinical significance and practical value for accurate diagnosis and prognosis of acquired brain injury (ABI) [20], prediction of Hodgkin lymphoma prognosis [21], and prediction of mortality after radical cystectomy for bladder cancer [22].

We found that clinical stage, blood transfusion history, D-dimer, age, and FDP were the top 5 predictors in the model we built. Evidence from extensive cohort studies suggests that the cancer stage is a risk factor for VTE. Yang et al. showed that advanced clinical stage (III-IV) was an independent risk factor for VTE in patients with cancer, which was 4.5 times higher than that in patients with early stage [23]. Meanwhile, Bezan et al. believe that higher clinical stages, particularly stages III-IV, strongly predict VTE in cancer patients [24]. Advanced cancer patients may more likely develop VTE due to increased combination therapy and reduced activity as their disease worsens. The history of blood transfusion is an easily overlooked factor. Chen et al. incorporated the blood transfusion history of patients into the VTE prediction algorithm for colorectal cancer patients. They achieved an excellent prediction effect with an AUC of 0.825 (95%CI: 0.721–0.930), which is a valuable indicator for predicting the risk of VTE [25]. Some studies have shown that a history of blood transfusion can be regarded as an independent predictor of VTE in cancer patients after surgery [26]. Nonetheless, Baumann et al. found no association between red blood cell (RBC) infusion and increased risk of VTE in 657412 hospitalized patients [OR = 1.0, 95%CI 0.96–1.05] [27]. So far, there is no apparent reason to explain this phenomenon. D-dimer, a degradation product of cross-linked fibrin, is rapidly increased in acute thrombosis. In the study of Osaki et al., D-dimer concentration was found to be an

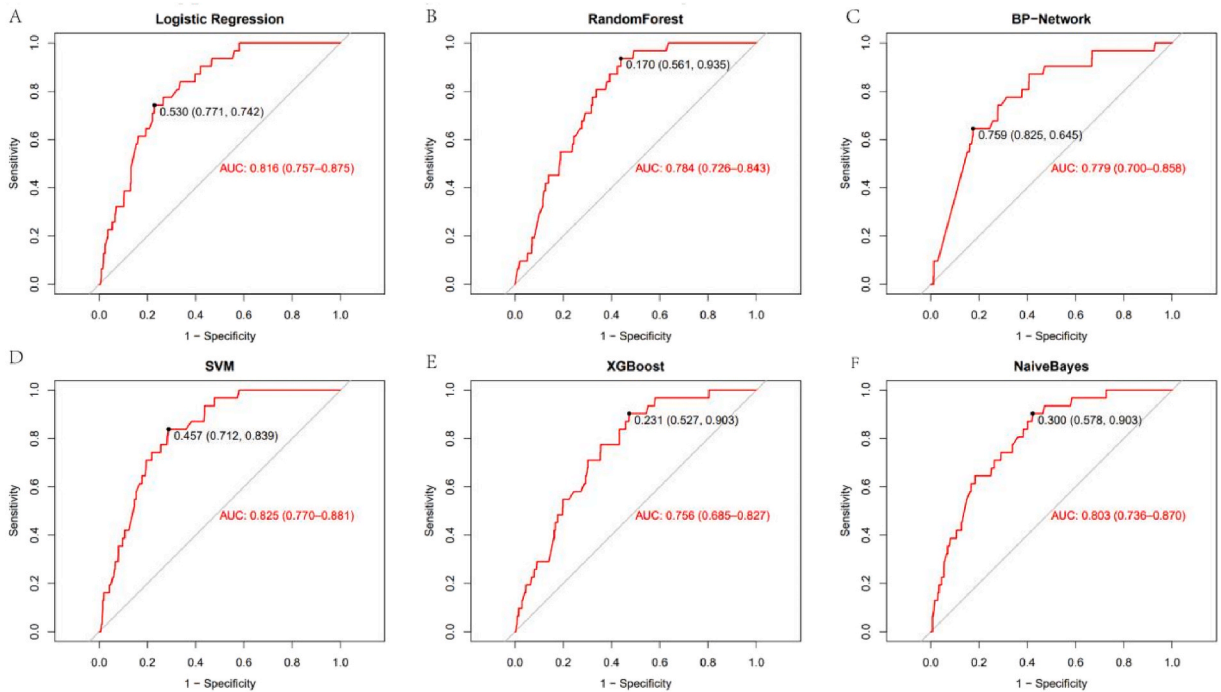


Fig. 3. Validation set ROC curve for each model(A:Logistic Regression, B:RandomForest, C:BP-Network, D:SVM, E:XGBoost, F: NaiveBayes).

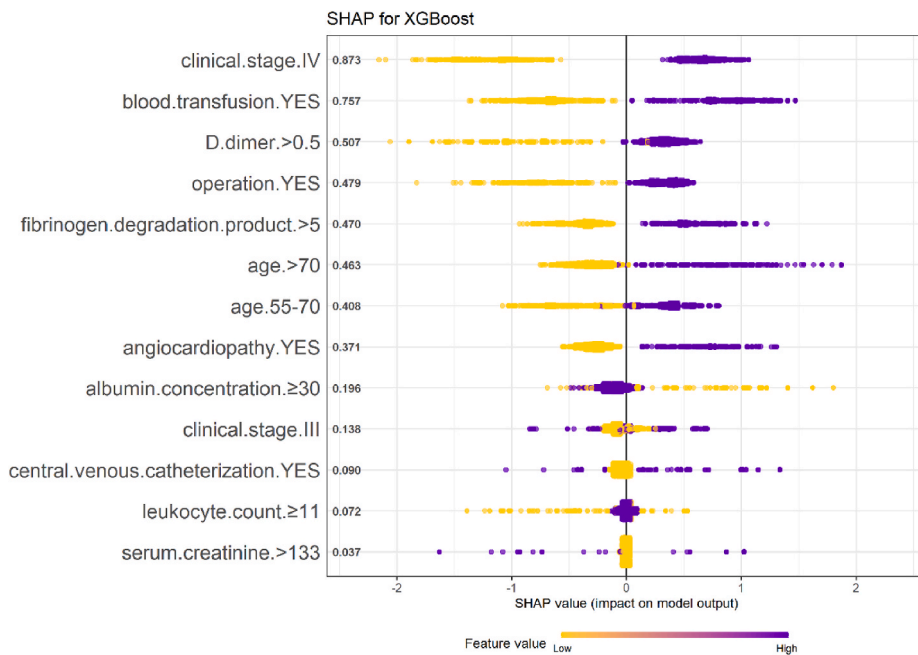


Fig. 4. SHAP values for each variable in the XGBoost model.

independent risk factor for preoperative VTE in gastric cancer patients [28]. In addition, some studies have confirmed that D-dimer has a high negative predictive value (96–100%) when detecting the occurrence of VTE in cancer patients, which can exclude VTE more accurately [29]. In a prospective study, Park et al. found that D-dimer level was the only marginally significant risk factor associated with VTE development (hazard ratio = 1.32, 95% CI: 1.00–1.75), suggesting that D-dimer could be used as a predictive marker for VTE development in patients with advanced gastric cancer [30]. The relationship between age and VTE is questionable. In their analysis of risk factors associated with VTE, Song et al. pointed out that age was the only factor related to VTE [31]. Lee believed that advanced

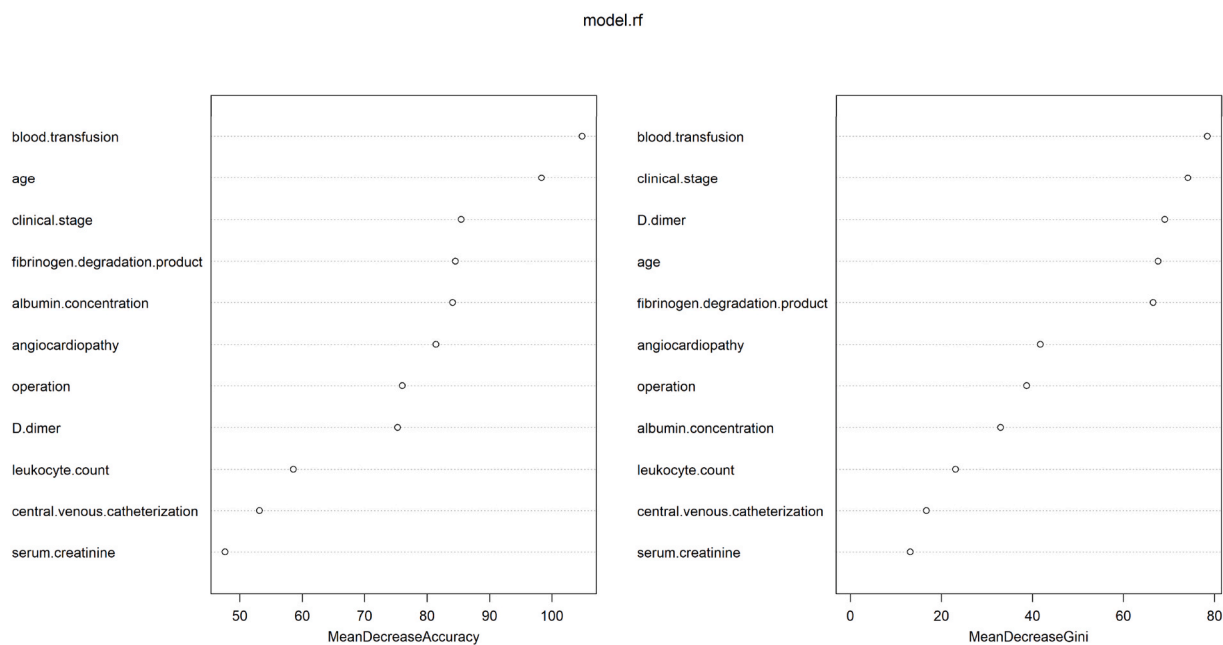


Fig. 5. The importance score of each variable in the RF.

age (>70 years) was an independent risk factor for VTE in gastric cancer patients, and the hazard ratio was 3.6 [32]. And Abdel-Razeq points out that age does not affect the occurrence of VTE in gastric cancer patients [33]. In terms of FDP, Zhou et al. found that FDP was significantly increased in patients with malignant tumors compared with healthy controls. Logistic regression analysis showed that FDP was correlated with VTE (OR = 1.022) [34]. Tsuji et al. also found that the FDP level of patients with VTE was significantly higher than that of patients without VTE, and the AUC value reached 0.933 when using FDP to diagnose VTE [35]. In the study of Hasegawa et al. the FDP level can be used to diagnose acute VTE [36].

The study used standard protocols and instruments, and all participants underwent a complete health examination. To ensure high-quality data collection, a rigorous personnel training process was established. All these are advantages, but the limitations also need to note. First, the study was cross-sectional, not longitudinal, so some confounding factors could not be eliminated. Second, discretization cut-off values for continuous variables based on clinical experience and literature may not be optimal.

5. Conclusion

In conclusion, based on SVM, we developed and validated a new clinical prediction model for VTE in gastric cancer patients. In the validation set, the AUC of the model was 0.825, the accuracy was 0.799, the sensitivity was 0.710, and the specificity was 0.802. The model has good performance and high application value in clinical practice, which can identify high-risk groups and prevent VTE in gastric cancer patients. In the future, external data from other centers can be considered for further validation of the model to test its generalization ability.

Declarations

Data availability statement

The authors have made the raw data supporting the conclusion of this article available to all qualified researchers without undue reservation.

Ethics statement

In our research, we followed the Declaration of Helsinki's ethical principles concerning the use of human subjects in medical research. Chongqing University Cancer Hospital's Ethics Committee reviewed and approved research studies.

Author contributions

Q Xu and H Lei performed the experiments. X Li, F Li analyzed and interpreted the data. H Shi, Q Xu wrote the paper. X Li, H Shi and H Lei contributed reagents, materials, analysis tools or data. G Wang, A Sun, Y Wang and B Peng conceived and designed the

experiments.

Funding

Support for this work was provided by the Chongqing Performance Incentive and Guidance Project for Scientific Research Institutions (cstc2020xj1130016).

Declaration of competing interest

There are no potential conflicts of interest among the authors.

References

- [1] K. Yang, J.K. Hu, Gastric cancer treatment: similarity and difference between China and Korea, *Transl Gastroenterol Hepatol* 2 (2017) 36.
- [2] N.B. Abdol Razak, et al., Cancer-associated thrombosis: an overview of mechanisms, risk factors, and treatment, *Cancers* 10 (10) (2018).
- [3] C.J. Fernandes, et al., Cancer-associated thrombosis: the when, how and why, *Eur. Respir. Rev.* 28 (151) (2019).
- [4] A.R. Yan, et al., Risk factors and prediction models for venous thromboembolism in ambulatory patients with lung cancer, *Healthcare (Basel)* 9 (6) (2021).
- [5] A. Zaheer, et al., Prediction models for venous thromboembolism in ambulatory adults with pancreatic and gastro-oesophageal cancer: protocol for systematic review and meta-analysis, *BMJ Open* 12 (3) (2022) e056431.
- [6] D. Farge, et al., International clinical practice guidelines including guidance for direct oral anticoagulants in the treatment and prophylaxis of venous thromboembolism in patients with cancer, *Lancet Oncol.* 17 (10) (2016) e452–e466.
- [7] J.A. Heit, F.A. Spencer, R.H. White, The epidemiology of venous thromboembolism, *J. Thromb. Thrombolysis* 41 (1) (2016) 3–14.
- [8] A. Labianca, et al., Risk prediction and new prophylaxis strategies for thromboembolism in cancer, *Cancers* 12 (8) (2020).
- [9] E.D. Tetzlaff, et al., Significance of thromboembolic phenomena occurring before and during chemoradiotherapy for localized carcinoma of the esophagus and gastroesophageal junction, *Dis. Esophagus* 21 (7) (2008) 575–581.
- [10] H.E. Fuentes, et al., Venous thromboembolism is an independent predictor of mortality among patients with gastric cancer, *J. Gastrointest. Cancer* 49 (4) (2018) 415–421.
- [11] E.D. Tetzlaff, et al., The impact on survival of thromboembolic phenomena occurring before and during protocol chemotherapy in patients with advanced gastroesophageal adenocarcinoma, *Cancer* 109 (10) (2007) 1989–1995.
- [12] K. Jayasakoon, et al., Gynecologic malignancy-associated venous thromboembolism and predictive tool at Thammasat university hospital, *Asian Pac. J. Cancer Prev. APJCP* 23 (6) (2022) 2113–2118.
- [13] L. Gervaso, H. Dave, A.A. Khorana, Venous and arterial thromboembolism in patients with cancer: JACC: CardioOncology state-of-the-art review, *JACC CardioOncol* 3 (2) (2021) 173–190.
- [14] X. Cheng, et al., Construction and verification of risk predicting models to evaluate the possibility of venous thromboembolism after robot-assisted radical prostatectomy, *Ann. Surg Oncol.* 29 (8) (2022) 5297–5306.
- [15] J. Nudel, et al., Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database, *Surg. Endosc.* 35 (1) (2021) 182–191.
- [16] X. Zhou, et al., Application of kNN and SVM to predict the prognosis of advanced schistosomiasis, *Parasitol. Res.* 121 (8) (2022) 2457–2460.
- [17] O. Mehrpour, et al., Utility of support vector machine and decision tree to identify the prognosis of metformin poisoning in the United States: analysis of National Poisoning Data System, *BMC Pharmacol Toxicol* 23 (1) (2022) 49.
- [18] C. Anil Kumar, et al., Lung cancer prediction from text datasets using machine learning, *BioMed Res. Int.* 2022 (2022), 6254177.
- [19] Y. Ruan, et al., A convex model for support vector distance metric learning, *IEEE Transact. Neural Networks Learn. Syst.* 33 (8) (2022) 3533–3546.
- [20] X. Wu, et al., Intrinsic functional connectivity patterns predict consciousness level and recovery outcome in acquired brain injury, *J. Neurosci.* 35 (37) (2015) 12932–12946.
- [21] S. Parodi, et al., Logic Learning Machine and standard supervised methods for Hodgkin's lymphoma prognosis using gene expression data and clinical variables, *Health Inf. J.* 24 (1) (2018) 54–65.
- [22] G. Wang, et al., Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques, *Comput. Biol. Med.* 63 (2015) 124–132.
- [23] J. Yang, et al., A novel nomogram based on prognostic factors for predicting venous thrombosis risk in lymphoma patients, *Leuk. Lymphoma* 62 (10) (2021) 2383–2391.
- [24] A. Bezan, et al., Risk stratification for venous thromboembolism in patients with testicular germ cell tumors, *PLoS One* 12 (4) (2017) e0176283.
- [25] Y. Chen, et al., A risk of venous thromboembolism algorithm as a predictor of venous thromboembolism in patients with colorectal cancer, *Clin. Appl. Thromb. Hemost.* 27 (2021), 10760296211064900.
- [26] S. Kaida, et al., A prospective multicenter observational study of venous thromboembolism after gastric cancer surgery (SHISA-1601), *Eur. Surg. Res.* 62 (1) (2021) 10–17.
- [27] L. Baumann Kreuziger, et al., Red blood cell transfusion does not increase risk of venous or arterial thrombosis during hospitalization, *Am. J. Hematol.* 96 (2021) 218–225.
- [28] T. Osaki, et al., Risk and incidence of perioperative deep vein thrombosis in patients undergoing gastric cancer surgery, *Surg. Today* 48 (5) (2018) 525–533.
- [29] N. Fünfsinn, et al., Rapid D-dimer testing and pre-test clinical probability in the exclusion of deep venous thrombosis in symptomatic outpatients, *Blood Coagul. Fibrinolysis* 12 (3) (2001) 165–170.
- [30] K. Park, et al., Incidence of venous thromboembolism and the role of D-dimer as predictive marker in patients with advanced gastric cancer receiving chemotherapy: a prospective study, *World J. Gastrointest. Oncol.* 9 (4) (2017) 176–183.
- [31] K.Y. Song, et al., Optimal prophylactic method of venous thromboembolism for gastrectomy in Korean patients: an interim analysis of prospective randomized trial, *Ann. Surg Oncol.* 21 (13) (2014) 4232–4238.
- [32] K.W. Lee, et al., The incidence, risk factors and prognostic implications of venous thromboembolism in patients with gastric cancer, *J. Thromb. Haemostasis* 8 (3) (2010) 540–547.
- [33] H. Abdel-Razeq, et al., Patterns and predictors of thromboembolic events among patients with gastric cancer, *Sci. Rep.* 10 (1) (2020), 18516.
- [34] K. Zhou, et al., Diagnostic and prognostic value of TAT, PIC, TM, and t-PAIC in malignant tumor patients with venous thrombosis, *Clin. Appl. Thromb. Hemost.* 26 (2020), 1076029620971041.
- [35] A. Tsuji, et al., Elevated levels of soluble fibrin in patients with venous thromboembolism, *Int. J. Hematol.* 88 (4) (2008) 448–453.
- [36] M. Hasegawa, et al., The evaluation of fibrin-related markers for diagnosing or predicting acute or subclinical venous thromboembolism in patients undergoing major orthopedic surgery, *Clin. Appl. Thromb. Hemost.* 24 (1) (2018) 107–114.