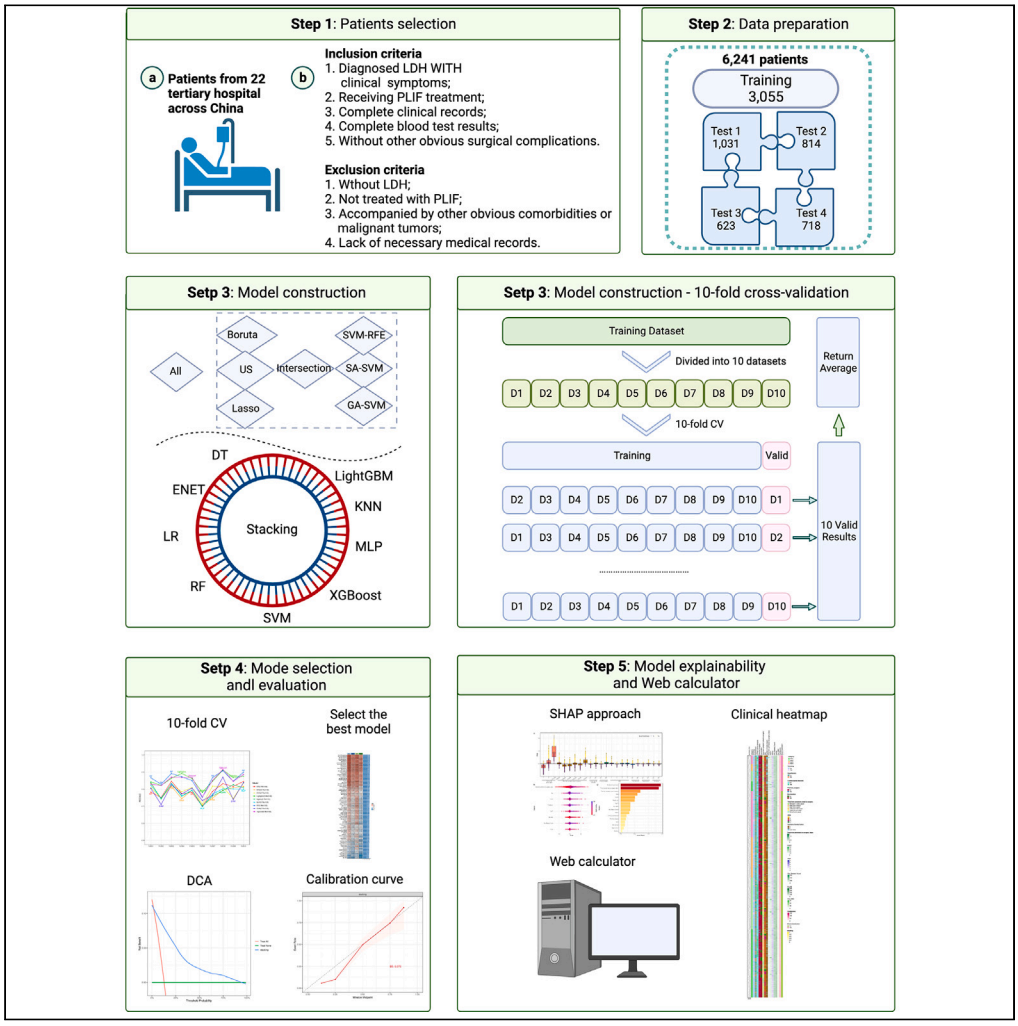
iScience

CellPress
OPEN ACCESS

Article

Development and validation of machine learning models for intraoperative blood transfusion prediction in severe lumbar disc herniation



Qiang Liu, An-Tian Chen, Runmin Li, ..., Zhaohui Hu, Baicheng Qiu, Wenle Li

329870376@qq.com (Z.H.)
qbc1314@126.com (B.Q.)
drlee0910@163.com (W.L.)

Highlights

A nationwide registered multicenter study with over 6000 patients from 22 hospitals

Comprehensive machine learning approaches to predict blood transfusion requirements

Construction of a web calculator to aid the clinical decision-making process

Proactively manage resources and optimize patient outcomes with the novel model

Liu et al., iScience 27, 111106
November 15, 2024
https://doi.org/10.1016/j.isci.2024.111106

iScience

CellPress
OPEN ACCESS

Article

Development and validation of machine learning models for intraoperative blood transfusion prediction in severe lumbar disc herniation

Qiang Liu,[1,35] An-Tian Chen,[2,3,35] Runmin Li,[4] Liang Yan,[5] Xubin Quan,[6] Xiaozhu Liu,[7,8] Yang Zhang,[9,10] Tianyu Xiang,[11] Yingang Zhang,[12] Anfa Chen,[13] Hao Jiang,[14] Xuewen Hou,[15] Qizhong Xu,[16] Weiheng He,[17] Liang Chen,[18] Xin Zhou,[19] Qiang Zhang,[20] Wei Huang,[21] Haopeng Luan,[22] Xinghua Song,[22] Xiaolin Yu,[23] Xiangdong Xi,[24] Kai Wang,[25] Shi-Nan Wu,[26] Wencai Liu,[27] Yusi Zhang,[28,29,30] Jialiang Zheng,[31] Haizhen Ding,[32] Chan Xu,[32] Chengliang Yin,[33] Zhaohui Hu,[6,*] Baicheng Qiu,[34,*] and Wenle Li[6,25,32,36,*]

SUMMARY

Lumbar disc herniation (LDH) is a common cause of lower back pain and sciatica, and posterior lumbar interbody fusion (PLIF) is always employed. This multicenter retrospective study investigates predicting intraoperative blood transfusion for LDH patients undergoing PLIF in China. The research includes 6,241 patients from 22 medical centers and employs 8 feature selection methods and 10 machine learning models, including an integrated stacking model. The optimal predictive model was selected based on the receiver operating characteristic area under the curve, clinical applicability, and computational efficiency. Among the evaluated combinations, the simulated annealing support vector machine recursive + stacking model achieved the highest performance with an area under the curve of 0.884, supported by robust calibration and decision curve analyses. A publicly accessible web calculator was developed to assist clinicians in decision-making. This work significantly enhances intraoperative transfusion predictions, providing valuable tools for improving patient management.

INTRODUCTION

Lumbar disc herniation (LDH) is a frequent cause of lower back pain and sciatica. Both nonoperative and operative treatments have been employed. Common approaches include patient education, physical therapy, alternative medicine, and pharmacotherapy.[1] To avoid painful suffering, a combination of discectomy and posterior lumbar interbody fusion (PLIF) is often employed. PLIF, introduced in the 1940s, is now regarded as the gold standard for spinal arthrodesis.[2] It is widely used for treating lumbar spondylolisthesis.[3] Although traditionally regarded

[1]Department of Orthopedics, Xianyang Central Hospital, Xianyang, Shannxi, China
[2]Department of Cardiology, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College/National Center for Cardiovascular Diseases, Beijing, China
[3]Department of Computer Science, University of Texas at Austin, Austin, TX, USA
[4]Department of Foot and Ankle Surgery, Honghui Hospital, Xi'an Jiaotong University, Xi'an, Shaanxi Province, China
[5]Department of Spinal Surgery, Honghui Hospital of Xi'an Jiaotong University, Xi'an, China
[6]Department of Spinal Surgery, Guangxi Medical University Affiliated Liuzhou People's Hospital, Liuzhou, China
[7]Department of Critical Care Medicine, Beijing Shijitan Hospital, Capital Medical University, Beijing, China
[8]Department of Cardiology, The Second Affiliated Hospital of Chongqing Medical University, Chongqing, China
[9]College of Medical Informatics, Chongqing Medical University, Chongqing, China
[10]Medical Data Science Academy, Chongqing Medical University, Chongqing, China
[11]Information Center, The University-Town Hospital of Chongqing Medical University, Chongqing, China
[12]Department of Orthopaedics of the First Affiliated Hospital, Medical School, Xi' an Jiaotong University, Xi'an, China
[13]Department of Orthopedics, Jiangxi Province Hospital of Integrated Chinese & Western Medicine, Nanchang, China
[14]Spine Tumor Center, Changzheng Hospital, Second Military Medical University, 415 Feng Yang Road, Shanghai 200003, China
[15]Department of Radiology, The First Dongguan Affiliated Hospital, Guangdong Medical University, Dongguan, China
[16]Department of Radiology, The First Affiliated Hospital of Shenzhen University, Shenzhen Second People's Hospital, Shenzhen, China
[17]Department of Radiology, People's Hospital of Ningxia Hui Autonomous Region, Yinchuan, China
[18]Department of Radiology, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, China
[19]Third Hospital of Shanxi Medical University, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, Taiyuan 030032, China
[20]First Department of Orthopaedics, Xi 'an Central Hospital, Xi'an, Shaanxi Province, China
[21]Department of Orthopedics, The Sixth Affiliated Hospital, School of Medicine, South China University of Technology, Foshan, Guangdong 528000, China
[22]Department of Spine Surgery, The Six Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, China
[23]Department of Orthopedics, Affiliated Hospital of Guizhou Medical University, Guiyang, China
[24]Department of Joint Surgery, No.215 Hospital of Shaanxi Nuclear Industry, Xianyang, Shannxi, China

Continued

as a thorough surgical approach, PLIF is less commonly employed for simple lumbar disc herniation today but may offer advantages such as sufficient access for thorough decompression of neural elements and restored segmental lordosis.[4] While traditional open PLIF effectively manages lumbar spinal instability, there are concerns about prolonged hospitalization, blood loss, and potential postoperative complications.[4] While minimally invasive (MIS) PLIF reduces blood loss, open PLIF remains necessary in specific cases like recurrent LDH or failed MIS attempts.

Intraoperative blood loss is a significant problem for surgeons and multimodal and multidisciplinary strategies were proposed.[5] Blood transfusion is often effective and required when intraoperative blood loss occurs to an extent. In a study involving 5803 patients in 126 European centers, the intraoperative transfusion rate was 1.8% in general.[6] The mean and median transfused packed red blood cells were 2.5 and 2 units respectively.[6] Hospitals that frequently perform blood transfusions to patients experiencing considerable surgical blood loss tend to have lower adjusted 30-day mortality rates, indicating that intraoperative blood transfusion in hospitals could serve as a potential indicator of surgical quality.[7]

However, research on blood transfusion and PLIF is limited. In this multicenter retrospective study, we aim to utilize a comprehensive machine learning approach to predict intraoperative blood transfusion in open PLIF fusion surgery of LDH patients (Figure 1).

RESULTS

Characteristics of patients

As indicated in Table 1, a total of 6,241 patients were involved in the study, with a nearly balanced distribution of males (51%) and females (49%). The gender distribution was slightly different between the transfusion (51% female, 49% male) and no transfusion groups (48% female, 52% male), with no statistically significant difference ($p = 0.118$). Median age was higher in the transfusion group (64 years) compared to the no-transfusion group (57 years). The height and weight of patients were similar across groups, but the body mass index (BMI) was slightly higher in the transfusion group and was statistically significantly different. 81% of the total participants were non-smokers, with a higher percentage of non-smokers in the transfusion group (86%) compared to the no transfusion group (80%). This difference was statistically significant ($p < 0.001$).

As for underlying conditions, significant differences were observed in the prevalence of hypertension, diabetes, coronary heart disease, ischemic stroke, lumbar spinal stenosis, lumbar spondylolisthesis, and lumbar scoliosis between the two groups. Significant differences were also found in the history of previous surgery and pre-surgery function. The study recorded the time from symptom onset to surgery, ranging from less than a week to more than 3 years. Significant differences were observed between the groups in this aspect. Information on the number of levels of fusion in surgery was provided. The majority (60%) had a single level fusion, with significant differences in the distribution between the transfusion and no transfusion groups.

Several preoperative lab measures including hematocrit, APTT, PT, fibrinogen, platelet count, hemoglobin, WBC, and albumin were compared. Significant differences were spotted in hematocrit, APTT, platelet count, hemoglobin, and albumin levels between groups.

Feature selection

The dataset initially contained 35 features. Different feature selection methods resulted in different subsets (Figures S1, 2A, and 2B). SVM-RFE selected 26 features, Boruta and GA-SVM each selected 25, while US and SA-SVM selected 18 and 16 features, respectively. Lasso identified only 5 features. Interestingly, the intersection of all methods contained only 5 features, indicating that while each method identifies distinct features, certain features were consistently chosen.

Basic and stacking machine learning models construction

Besides LR, basic models were first tuned and guided by AUC-ROC (Figure S2). Both the tuning process and LR (Figure S2) can be found in supplement materials. 10-fold cross-validation was applied for all nine fine-tuned models to test their performance on the training dataset (Figure S3A). It turned out that the LightGBM and RF had the best performance, with the AUC-ROC reaching 0.86. XGBoost laid behind with an AUC-ROC of 0.85. Other models failed to reach 0.80 though.

[25]Key Laboratory of Neurological Diseases, The Second Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China
[26]Eye Institute of Xiamen University, School of Medicine, Xiamen University, Xiamen, Fujian, China
[27]Department of Orthopedics, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200233, China
[28]Cancer Center, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
[29]Precision Medicine Center, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
[30]Department of Medical Oncology, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
[31]Cancer Research Center, School of Medicine, Xiamen University, Xiamen, China
[32]State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics & Center for Molecular Imaging and Translational Medicine, School of Public Health, Xiamen University, Xiamen, China
[33]Faculty of Medicine, Macau University of Science and Technology, Macau, China
[34]Medical Cosmetic Department, Foresea Life Insurance Guangxi Hospital, Nanning, China
[35]These authors contributed equally
[36]Lead contact
*Correspondence: 329870376@qq.com (Z.H.), qbc1314@126.com (B.Q.), drlee0910@163.com (W.L.)
https://doi.org/10.1016/j.isci.2024.111106

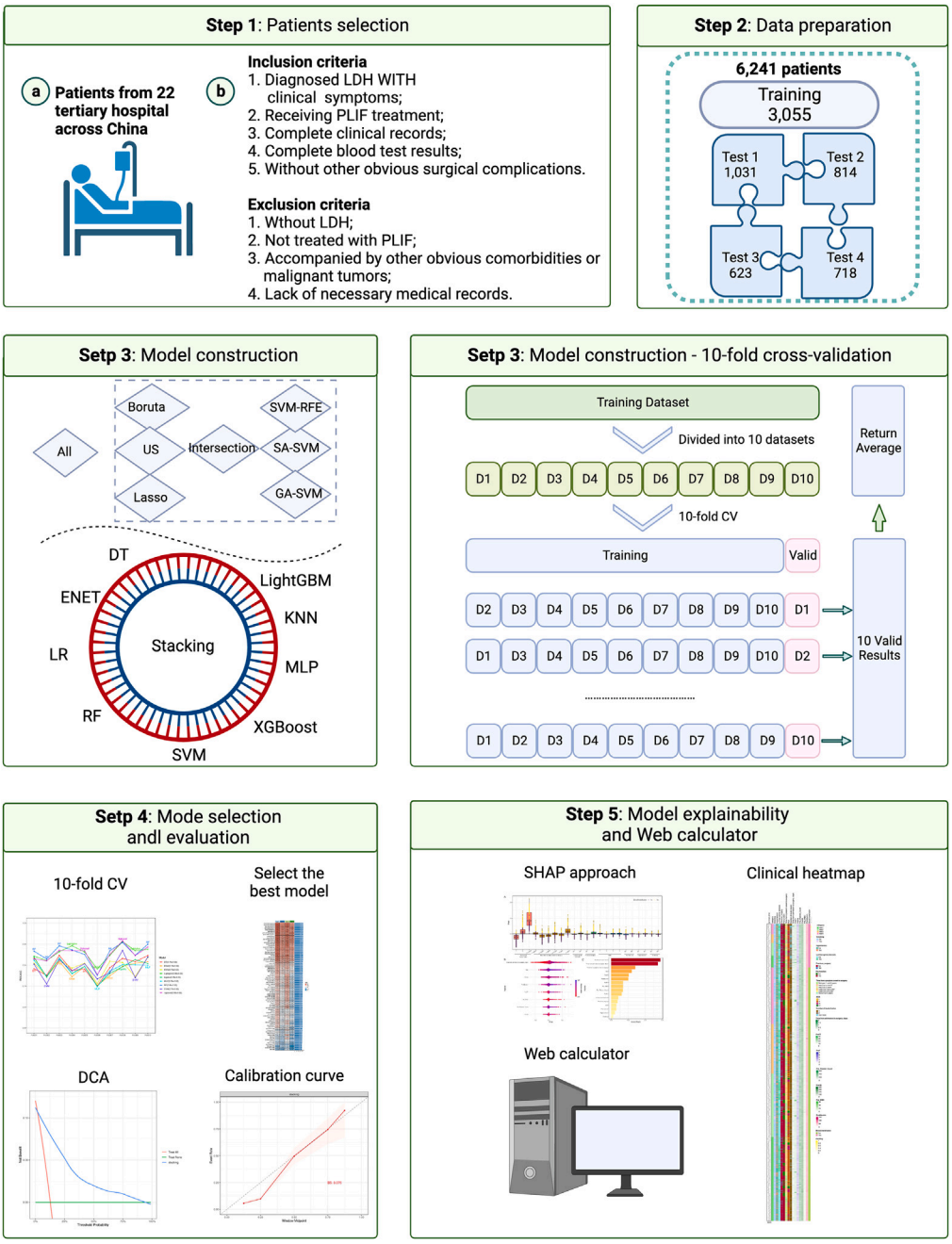


Figure 1. Comprehensive flowchart of the study

Patients were selected from 22 hospitals across China based on the inclusion and exclusion criteria. 6,241 patients were finally involved and divided into training and test groups. 8 feature selection approaches, and 10 machine learning models (9 basic and 1 integrated stacking model) were adopted. After choosing the best model, DCA and calibration curve were applied for evaluation. Model explainability was guaranteed by SHAP approach. A publicly accessible web calculation was established for clinical usage. CV, cross-validation, SVM-RFE, support vector machine recursive feature elimination; GA-SVM, genetic algorithm support vector machine recursive; US, univariate screening; SA-SVM, simulated annealing support vector machine; Lasso, least absolute shrinkage and selection operator; LightGBM, light gradient boosting machine; RF, random forest; DT, decision tree; ENet, efficient neural network; KNN, k-nearest neighbors; LR, logistic regression; MLP, multilayer perceptron; SVM, support vector machine; XGBoost, extreme gradient boosting; DCA, decision curve analysis.

Table 1. Basic characteristics of involved patients

Variables	Total (*n* = 6241)	No transfusion (*n* = 5315)	Transfusion (*n* = 926)	*p* value
Gender, n (%)				0.118
female	3050 (49)	2575 (48)	475 (51)	–
male	3191 (51)	2740 (52)	451 (49)	–
Smoking, n (%)				<0.001
No	5025 (81)	4228 (80)	797 (86)	–
Yes	1216 (19)	1087 (20)	129 (14)	–
Hypertension, n (%)				<0.001
No	4412 (71)	3846 (72)	566 (61)	–
Yes	1829 (29)	1469 (28)	360 (39)	–
Diabetes, n (%)				<0.001
No	5602 (90)	4811 (91)	791 (85)	–
Yes	639 (10)	504 (9)	135 (15)	–
Coronary heart disease, n (%)				0.011
No	5949 (95)	5082 (96)	867 (94)	–
Yes	292 (5)	233 (4)	59 (6)	–
Chronic heart failure, n (%)				0.561
No	6202 (99)	5280 (99)	922 (100)	–
Yes	39 (1)	35 (1)	4 (0)	–
Hemorrhagic stroke, n (%)				0.445
No	6135 (98)	5228 (98)	907 (98)	–
Yes	106 (2)	87 (2)	19 (2)	–
Ischemic stroke, n (%)				<0.001
No	6088 (98)	5223 (98)	865 (93)	–
Yes	153 (2)	92 (2)	61 (7)	–
Lumbar spinal stenosis, n (%)				<0.001
No	4336 (69)	3786 (71)	550 (59)	–
Yes	1905 (31)	1529 (29)	376 (41)	–
Lumbar spondylolisthesis, n (%)				<0.001
No	5723 (92)	4910 (92)	813 (88)	–
Yes	518 (8)	405 (8)	113 (12)	–
Sciatica, n (%)				0.178
No	6114 (98)	5201 (98)	913 (99)	–
Yes	127 (2)	114 (2)	13 (1)	–
Osteoporosis, n (%)				0.117
No	6088 (98)	5192 (98)	896 (97)	–
Yes	153 (2)	123 (2)	30 (3)	–
Lumbar spine fracture, n (%)				0.356
No	6204 (99)	5281 (99)	923 (100)	–
Yes	37 (1)	34 (1)	3 (0)	–
Lumbar scoliosis, n (%)				<0.001
No	6184 (99)	5280 (99)	904 (98)	–
Yes	57 (1)	35 (1)	22 (2)	–
Previous transfusion, n (%)				0.461
No	6153 (99)	5243 (99)	910 (98)	–
Yes	88 (1)	72 (1)	16 (2)	–

(*Continued on next page*)

Table 1. *Continued*

Variables	Total (*n* = 6241)	No transfusion (*n* = 5315)	Transfusion (*n* = 926)	*p* value
Previous surgery, n (%)				<0.001
No	5289 (85)	4580 (86)	709 (77)	–
Yes	952 (15)	735 (14)	217 (23)	–
Pre function, n (%)				<0.001
No	1783 (29)	1600 (30)	183 (20)	–
Yes	4458 (71)	3715 (70)	743 (80)	–
Bone graft, n (%)				<0.001
No	322 (5)	300 (6)	22 (2)	–
Yes	5919 (95)	5015 (94)	904 (98)	–
Time from symptom onset to surgery, n (%)				<0.001
1 to 3 years	673 (11)	545 (10)	128 (14)	–
<1 month	1527 (24)	1379 (26)	148 (16)	–
<1 week	1141 (18)	1057 (20)	84 (9)	–
< half a year	1246 (20)	1020 (19)	226 (24)	–
<1 year	578 (9)	482 (9)	96 (10)	–
>3 years	1076 (17)	832 (16)	244 (26)	–
ASA, n (%)				<0.001
1	332 (5)	288 (5)	44 (5)	–
2	4419 (71)	3826 (72)	593 (64)	–
3	1083 (17)	845 (16)	238 (26)	–
4	407 (7)	356 (7)	51 (6)	–
Total comorbidities, n (%)				<0.001
0	3724 (60)	3289 (62)	435 (47)	–
1	1677 (27)	1375 (26)	302 (33)	–
2	692 (11)	546 (10)	146 (16)	–
3 or more	148 (2)	105 (2)	43 (5)	–
Number of levels fusion, n (%)				<0.001
1	3775 (60)	3558 (67)	217 (23)	–
2	1892 (30)	1464 (28)	428 (46)	–
3 or more	574 (9)	293 (6)	281 (30)	–
Age (yrs), Median (Q1, Q3)	58 (49, 67)	57 (48, 66)	64 (55.25, 70)	<0.001
Height (m), Median (Q1, Q3)	1.65 (1.62, 1.7)	1.65 (1.62, 1.7)	1.65 (1.62, 1.7)	<0.001
Weight (kg), Median (Q1, Q3)	67.12 (63, 72.6)	67.03 (63, 72.48)	67.98 (63, 73)	0.099
BMI (kg/m^2), Median (Q1, Q3)	24.03 (22.14, 26.2)	24 (22.1, 26.12)	24.49 (22.6, 26.79)	<0.001
Time from admission to surgery days, Median (Q1, Q3)	4.38 (3.43, 6.21)	4.19 (3.31, 6)	5.46 (4.39, 7.88)	<0.001
Pre Hct (%), Median (Q1, Q3)	41.1 (37.7, 44.2)	41.3 (38, 44.31)	39.7 (36.3, 43.49)	<0.001
Pre APTT (s), Median (Q1, Q3)	29.8 (27.4, 32.8)	29.89 (27.6, 33)	29.2 (26.7, 31.2)	<0.001
Pre PT (s), Median (Q1, Q3)	11.9 (11.3, 12.7)	11.9 (11.3, 12.7)	11.9 (11.1, 12.7)	0.186
Pre Fib (g/L), Median (Q1,Q3)	2.83 (2.46, 3.24)	2.83 (2.46, 3.26)	2.82 (2.45, 3.2)	0.462
Pre PLT (10^9/L), Median (Q1, Q3)	220.99 (187, 252)	221.59 (189, 252.19)	212.35 (175, 250)	<0.001

(*Continued on next page*)

Table 1. *Continued*

Variables	Total (*n* = 6241)	No transfusion (*n* = 5315)	Transfusion (*n* = 926)	*p* value
Pre HGB (g/L), Median (Q1, Q3)	135 (125.82, 146.01)	135.37 (126, 147)	132.35 (122, 144)	<0.001
Pre WBC (10^9/L), Median (Q1, Q3)	6.33 (5.3, 7.41)	6.34 (5.3, 7.4)	6.22 (5.2, 7.46)	0.123
Pre albumin (g/L), Median (Q1, Q3)	41.7 (39.36, 44.6)	41.8 (39.4, 44.7)	41.2 (38.77, 43.9)	<0.001

APTT, activated partial thromboplastin time; Fib, fibrinogen; Hct, hematocrit; HGB, hemoglobin; PLT, platelet count; Pre, preoperative; PT, prothrombin time; WBC, white blood cell.

A stacking model was built based on the combination of DT, ENet, KNN, LightGBM, LR, MLP RF, SVM, and XGBoost models. We adopted Lasso regression as the meta-classifier for our comprehensive stacking model. After training, the stacking model was found to have higher accuracy than any basic model mentioned above (Figure 3A).

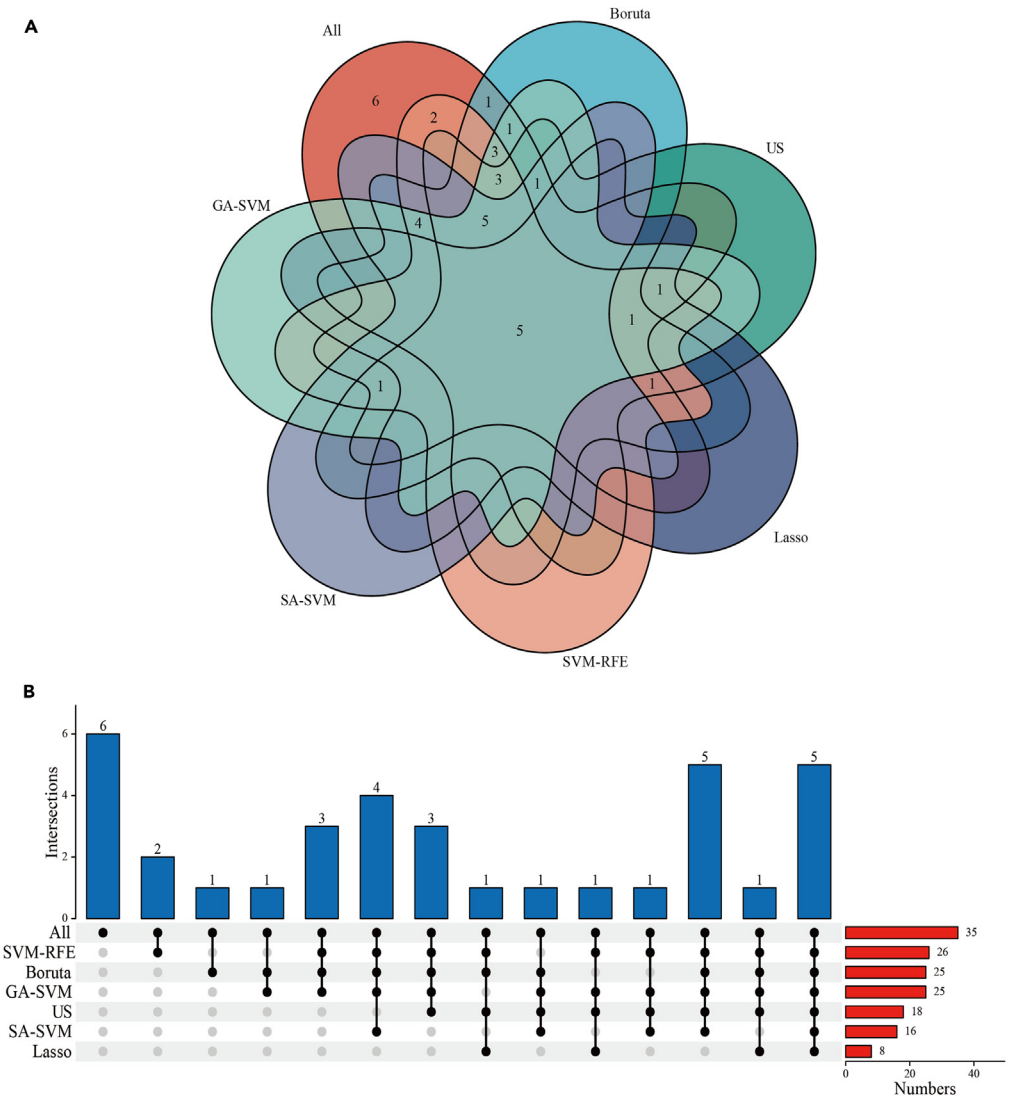


Figure 2. Results of feature selection

(A) Venn diagram of feature intersections among six feature selection models and original features. Region 5 reflects the intersection of all methods.

(B) Details of features selected by each model. There are 35 features in total. Abbreviations: SVM-RFE, support vector machine recursive feature elimination; GA-SVM, genetic algorithm support vector machine recursive; US, univariate screening; SA-SVM, simulated annealing support vector machine; Lasso, least absolute shrinkage and selection operator.

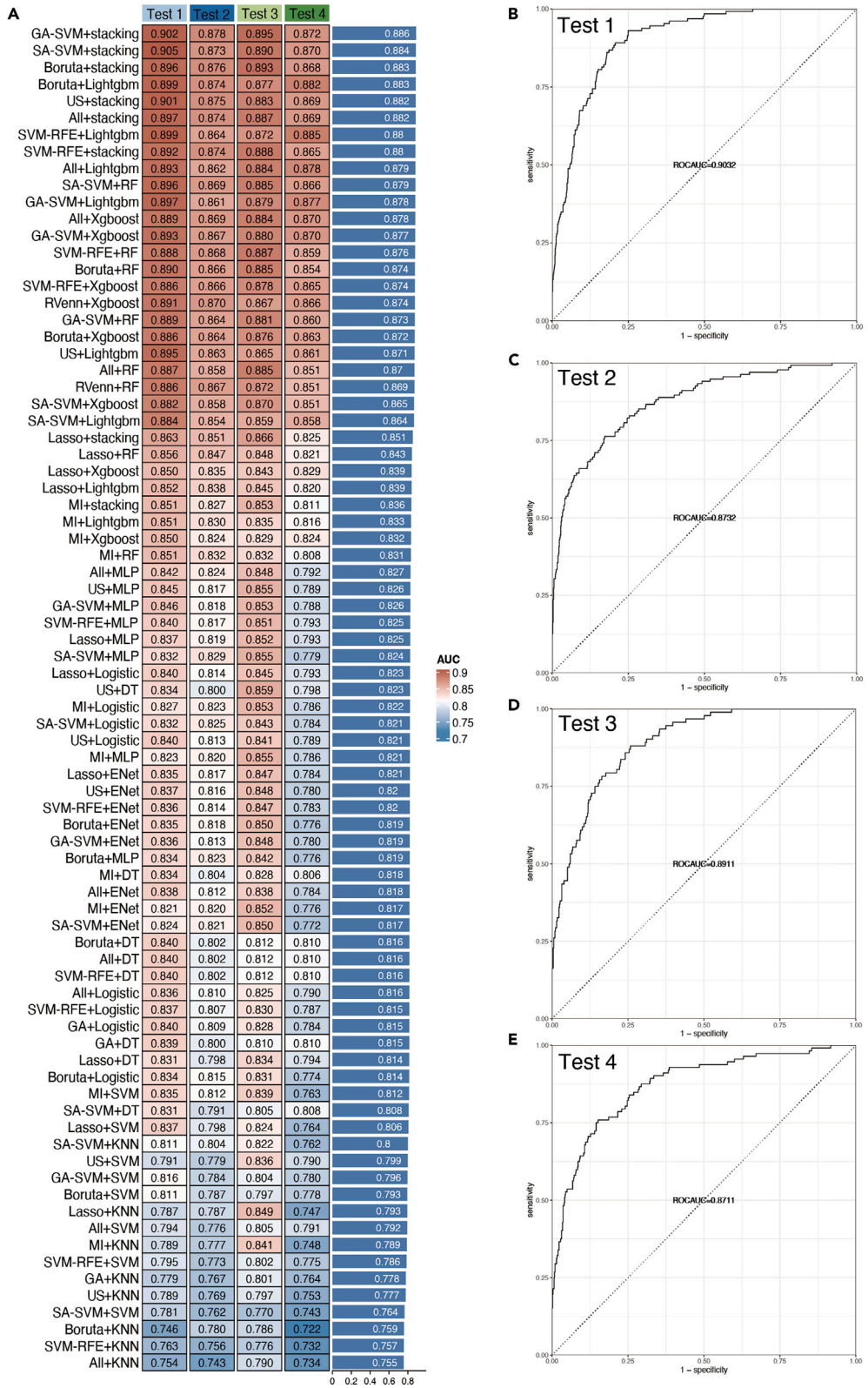


Figure 3. AUC-ROC of models

(A) AUC-ROC of all possible combinations between feature selection group and machine learning models.

(B) AUC-ROC of SA-SVM + stacking on Test 1.

Figure 3. *Continued*
(C) AUC-ROC of SA-SVM + stacking on Test 2.
(D) AUC-ROC of SA-SVM + stacking on Test 3.
(E) AUC-ROC of SA-SVM + stacking on Test 4. Though GA-SVM + stacking showed the highest AUC-ROC, SA-SVM + stacking was regarded as the best model since it achieved an AUC-ROC of only 0.02 behind with two fewer features, which made it easier to adopt in clinical settings. US, univariate screening; Lasso, least absolute shrinkage and selection operator; SVM-RFE, support vector machine recursive feature elimination, SA-SVM, simulated annealing support vector machine recursive; GA-SVM, genetic algorithm support vector machine recursive. DT, decision tree; ENet, efficient neural network; KNN, k-nearest neighbors; LightGBM, light gradient boosting machine; LR, logistic regression; MLP, multilayer perceptron; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

Model evaluation

Up to now, we acquired 8 feature groups and 10 machine learning models. By combining each feature group with a machine learning model, 80 combinations were reached. We first evaluated all possible joints by AUC-ROC (Figure 3A). Four test datasets were all evaluated, and the mean AUC-ROC was obtained for model evaluation. Of all combinations, GA-SVM + stacking showed the highest 0.886 AUC-ROC. It is followed by SA+SVM + stacking with an AUC-ROC of 0.884. Simply pouring all features without selection into the KNN model led to the worst AUC-ROC, which was only 0.755. However, when figuring out the best combination, AUC-ROC was not the only concern. Though GA-SVM + stacking led SA+SVM + stacking by 0.002, it required more features (18 versus 16 features) as illustrated in Figure 2. Taking the clinal convenience of usage into consideration, SA-SVM + stacking was regarded as the best combination since it achieved an almost optimized AUC-ROC with few features. Performance the best model measured by AUC-ROC could be found through Figures 3B–3E.

For further analysis, the calibration curve and DCA were evaluated for the SA-SVM + stacking model in all 4 test datasets. DCA aimed to evaluate the clinical benefit of the chosen prediction models (Figure 4). In DCA, the x axis refers to the range of threshold probabilities from 0% to 100%, indicating where a clinician would decide that the benefit of treatment outweighs the harm. The y axis represents the net benefit reflecting the benefit of predicting true positives minus the harm of predicting false negatives. In all 4 test datasets, the curve of the predictive SA-SVM + stacking model was spotted to be almost entirely above the 'Treat All' and 'Treat None' lines with a threshold of more than 90%, indicating that the model could provide better guidance in decision-making than either extreme approach.

The calibration curve measures how well the probabilities of the event correspond to the actual outcome in real cases. It is represented with the predicted probability midpoints on the x axis and the observed event rates on the y axis. In our analysis of all 4 test datasets, the line of the model deviated from the line of perfect calibration when predicted probabilities were lower, suggesting underestimations of the event risk (Figure 5). However, as the predicted probabilities increased, the model's calibration line approached closer and nearly coincided with perfect calibration, implying a more accurate prediction for a higher risk of blood transfusion after PLIF. Also, the Brier score was lower than 0.01 in all test datasets, indicating that the prediction of the SA-SVM + stacking model was overall accurate.

Explainable model interpretation

The Shapley additive explanations (SHAP) approach was adopted to make the SA-SVM + stacking model explainable by quantifying the impact of each feature on the prediction. As illustrated in Figure 6, both category variables (Figure 6A) and continuous variables (Figure 6B) were analyzed using the SHAP approach. For category variables, patients with a greater number of fusions, more than 3 years from symptom onset to surgery, presence of lumbar stenosis, history of previous surgery, dysfunction in body movement before surgery, hypertension, and smoking-free were more prone to blood transfusion in PLIF. Also, patients with a longer time from admission to surgery, aged, lower hemoglobin, higher albumin, higher PT, lower platelet count, and lower fibrinogen were linked to a higher probability of blood transfusion.

The study further interpreted the model by ordering the features in a descending manner based on their contribution to the established predictive model. As shown in Figure 6C, the top 2 leading features were the number of levels of fusion and days from admission to surgery. The following 4 contributing features included the time from symptom onset to surgery, age, hemoglobin, and albumin level before surgery. As illustrated in Figure 6D, the clinical heatmap of the constructed model revealed both the distribution and influence of involved clinical characteristics within the established model, enhancing the model's interpretability and explainability.

Public-access web-based calculator

An interactive calculator has been developed to predict the likelihood of intraoperative blood transfusion in patients undergoing PLIF for LDH. This user-friendly web-based calculator is publicly accessible and can be accessed at https://nicolazhang.shinyapps.io/TransFusion_LDH_PLIF/ (Figure 7). Limited by the capacity of the website, ***it may take minutes to access the calculator***. After that, users can input and select relevant variables within the "LDH patients awaiting PLIF" category, and the predictions regarding the need for intraoperative blood transfusion will be promptly displayed in the "Should intraoperative blood transfusion be prepared?" section.

DISCUSSION

As the world population steps into aging, LDH will negatively affect more populations due to its degenerative nature. PLIF acts as a significant approach in dealing with LDH, which can be further divided into open PLIF and MIS PLIF. Though MIS PLIF may have some advantages over

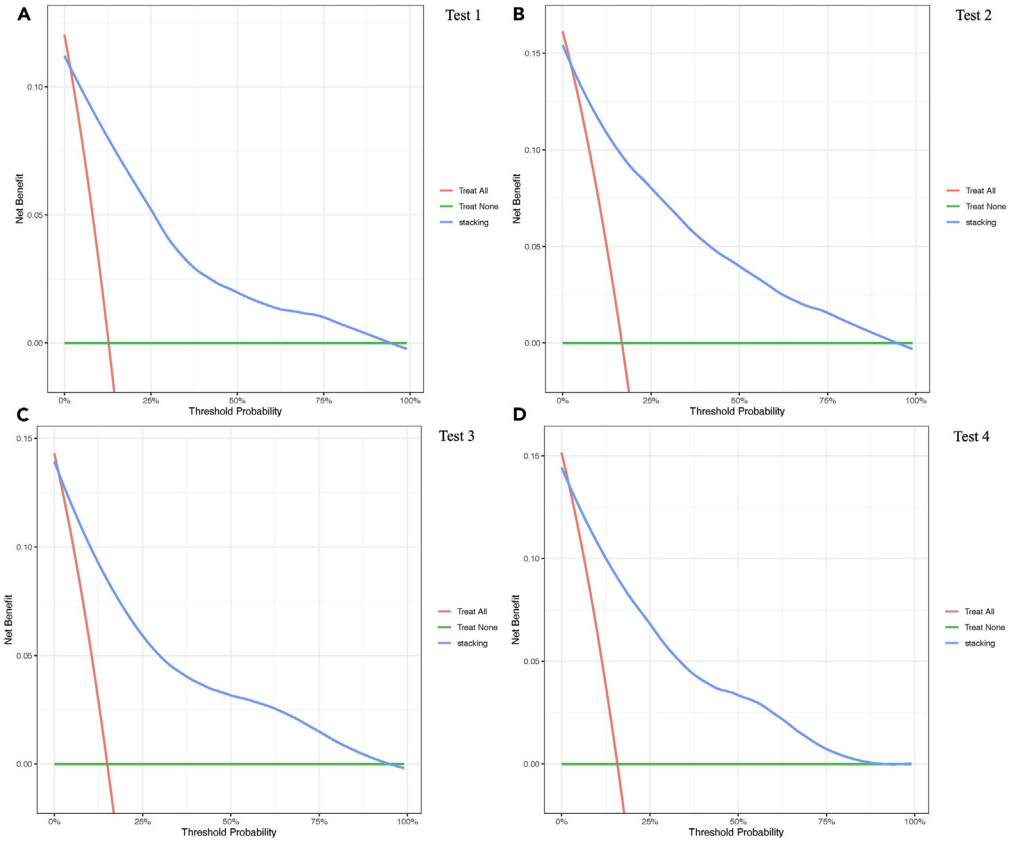


Figure 4. Decision curve analysis of the SA-SVM + stacking model on 4 test datasets

In all test datasets, the curve of the predictive model was above the other two extreme lines with a threshold of more than 90%, suggesting that the SA-SVM + stacking model provided better guidance in decision-making than either extreme approach.

(A–D) represents test datasets 1 to 4 respectively.

open PLIF, there are conditions under which open PLIF is compulsory, especially in recurrent LDH or patients who cannot undergo MIS PLIF. Complex LDH cases are more prone to experience open PLIF, its complexity may lead to intraoperative blood loss and require transfusion. Predicting the need for blood transfusion during surgical procedures is critical. Previous studies combining machine learning and blood transfusion focused on single-center data, or only adopt a single machine learning algorithm, lacking multi-center evidence and advanced machine learning techniques.[8,9] This study addresses this gap by developing a machine-learning model to predict intraoperative transfusion during PLIF surgery, offering an approach to surgical planning.

Machine learning approaches have gained increasing popularity and have been widely used in the medical field, especially in the construction of predictive models.[10–13] Also, machine learning has made revolutions in medicine, assisting medical diagnosis and drug development.[14–16] Advantages of machine learning over traditional biostatistical methods include the ability to deal with scaled data and flexibility in handling diverse data types comprehensively.[17] Despite a bunch of advantages, machine learning is also challenging in data preprocessing, model construction, model explainability, and fit into clinical settings. Thus, in this multicenter retrospective study, we attempted to develop a comprehensive machine-learning model for predicting intraoperative blood transfusion in LDH patients during PLIF. This approach reflects the growing trend of using ensemble methods in clinical decision support systems, as stacking models outperform single algorithm models in various medical research. Besides providing model interoperability to solve the black-box problem of machine learning, we stepped further by selecting the most optimizing model and enabled the research-to-bedside application by constructing a publicly accessible web-based calculator.

The registered multicenter study collected data from 22 centers across China to enhance the representation. To optimize the performance and potential in clinical application, we adopted an integrated method for model construction. 8 feature selections were utilized to ensure that possible feature groups were included. Eight feature groups included those selected by Boruta, US, Lasso, SVM-RFE, SA-SVM, and GA-SVM algorithm, raw features without selection, and the intersection of all feature groups. 9 basic machine learning models, which were DT, ENet, KNN, LightGBM, logistic, MLP, RF, SVM, and XGBoost, and a stacking model were tested. To guarantee a high-quality comprehensive machine learning model, a cross-grouping between feature selection methods and machine learning algorithms was performed, resulting in a total of eighty combinations.

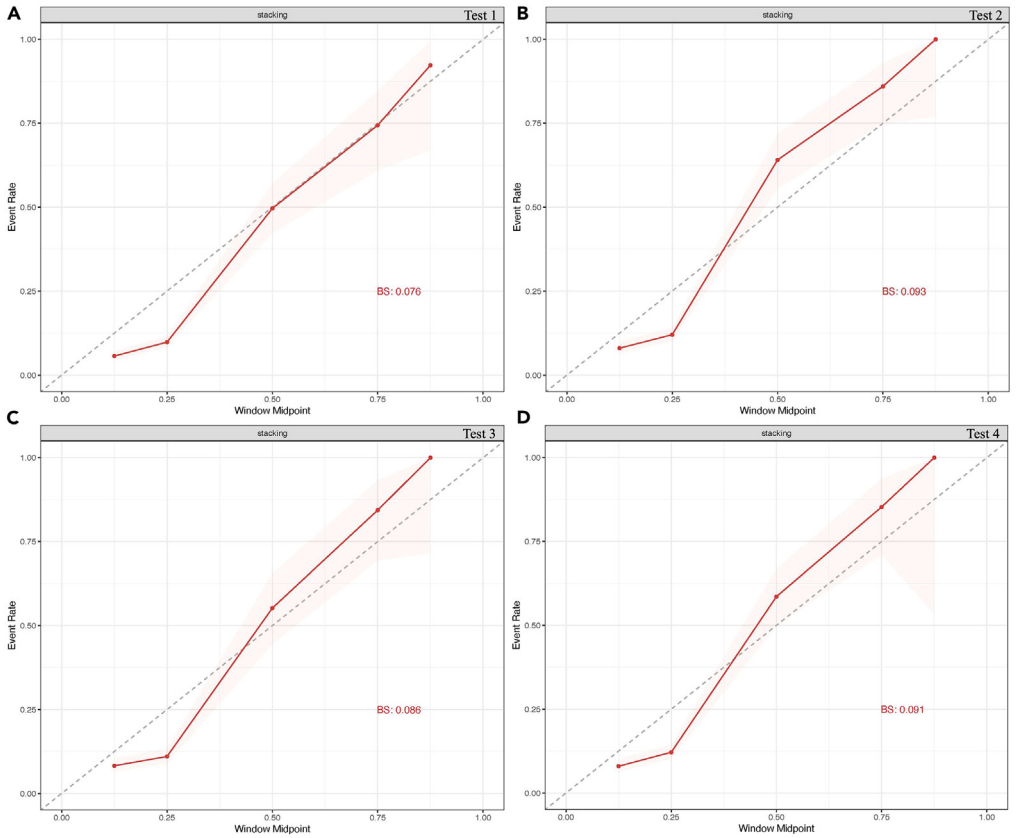


Figure 5. Calibration curve of the SA-SVM + stacking model on 4 test datasets

The red line indicates the calibration of the SA-SVM + stacking model, with each point on the line depicting the observed event rate for the corresponding predicted probability interval. The dotted line refers to perfect calibration, where predicted probabilities exactly match the observed event rates. The proximity of the evaluated model's calibration curve to this line is indicative of the model's performance. The shaded area surrounding the calibration curve provides the confidence interval.

(A–D) represents test datasets 1 to 4 respectively.

The study recognized that simply containing all features may not contribute to the model, but even be harmful to model performance. Applying feature selection approaches did help in improving the robustness of machine learning models. Among 80 combinations between feature groups and machine learning algorithms, the integrated stacking performed better than other basic models. The stacking ensemble model is an integrated machine learning model composed of base learners or individual classifiers to form a meta-learner.[18] In several previous studies, the robustness of stacking ensemble methods has been demonstrated to surpass that of individual basic machine learning learners.[19–21] The stacking model is gaining popularity and is even utilized for compound-protein binding affinity prediction.[22]

In general, we aimed to build a model that is balanced between performance and computing cost, which performance is not the only aspect we cared about. Computation cost was also taken into account, which could be represented as number of features in models in this study. Although the GA-SVM + stacking has the highest AUC-ROC, SA-SVM was regarded as the optimized one for fewer features and slightly compromised performance by 0.002. Developed in 2008 for parameter determination and feature selection in the SVM, the goal of SA-SVM was to optimize parameter values while simultaneously identifying a subset of features that maintained the accuracy of SVM[23,24]. The procedure of the SA-SVM approach could be accessed in the prior literature, in which the accuracy for classification was also found to be ideal.[25]

For the evaluation in terms of DCA and calibration, our ensembled SA-SVM + stacking model performed well. The DCA curve of the model mostly ran above the extreme cases with a threshold of more than 90%, revealing the model's ability to guide clinical decision-making. The observed calibration in test datasets suggested that while the SA-SVM + stacking model may not fit perfectly when the probability of blood transfusion was low, it demonstrated reliable performance for higher probability predictions of events. The model's reliability at higher probabilities is valuable in clinical settings since patients with higher event risk may benefit more. Also, the overall low Brier score validated the model's utility, though the noted calibration imperfections shall be considered when adopting the predictive model into real-world settings.

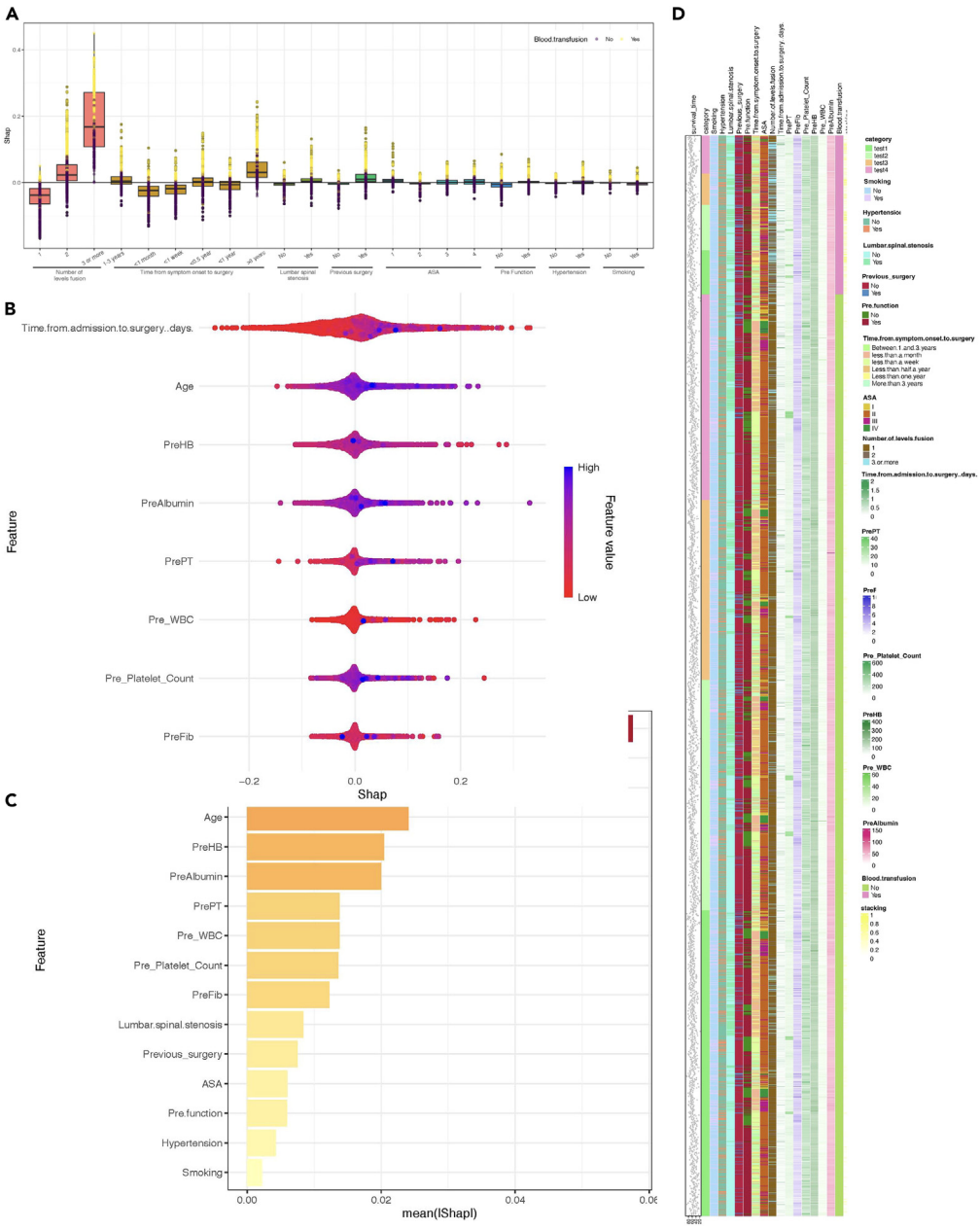


Figure 6. SHAP approach for model explainability and feature importance

(A) SHAP approach for category variables.

(B) SHAP approach for continuous features.

(C) Feature importance in descending order.

(D) Clinical heatmap of the integrated model. A clinical heatmap is a graphical representation tool used extensively in medical research and clinical data analysis It depicted the spread and influence of diverse clinical characteristics within the established comprehensive model, which aids in comprehending the interconnections and contributions of various factors to the model's predictive outcomes.

The SHAP approach enabled our final SA-SVM + stacking model to be explainable. The most clinically relevant findings indicated that intraoperative blood transfusion during PLIF for LDH patients was positively associated with the following factors: greater number of fusions, delayed surgery (more than 3 years from symptom onset to surgery), longer time from admission to surgery, presence of lumbar stenosis, history of previous surgery, preoperative dysfunction in body movement, older age, lower hemoglobin, higher albumin, and abnormal coagulation parameters (higher PT, lower platelet count, and lower fibrinogen).

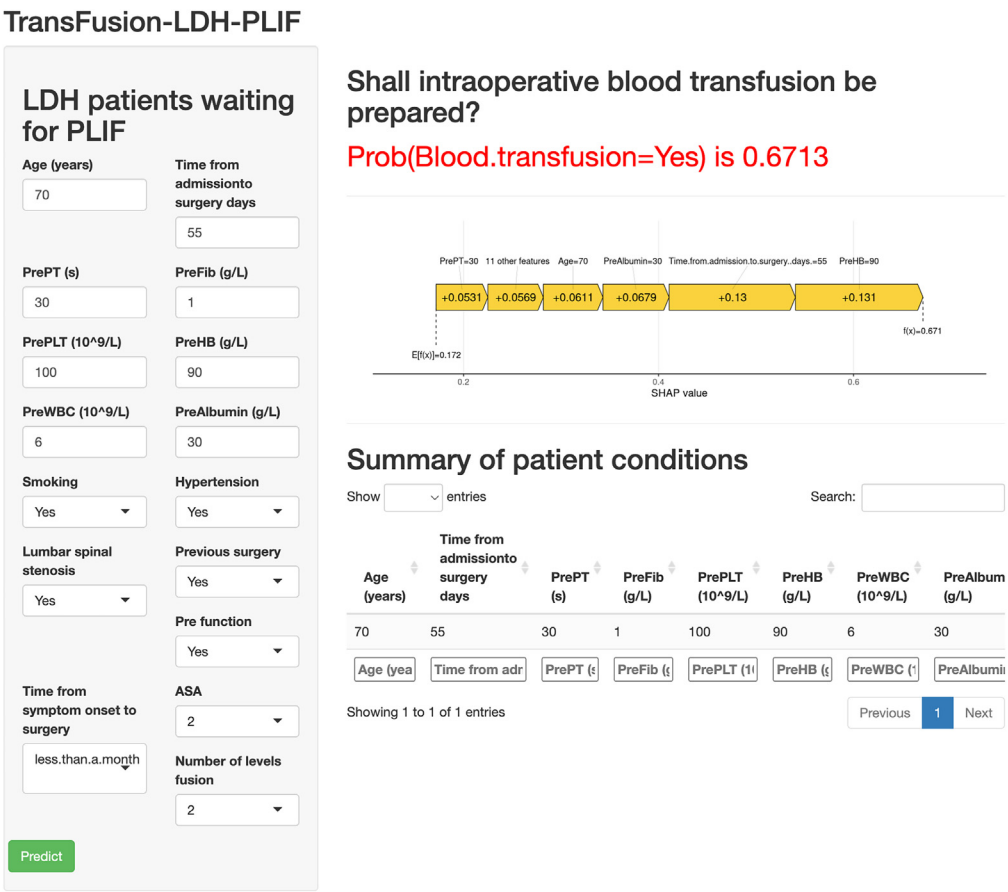


Figure 7. Prediction made by the online calculator

The probability of intraoperative blood transfusion of the patient is 0.6713 from the comprehensive machine-learning model. Surgeons shall be prepared for blood transfusion during surgery.

These findings underscore the complexity of LDH, with the number of fusions, lumbar stenosis, previous surgery, and preoperative dysfunction in body movement being indicative of surgical challenges that can lead to increased blood loss and require intraoperative blood transfusion. For instance, complex adult deformity correction through multilevel spine fusion surgery is often linked with considerable blood loss.[26] A study found that lumbar spinal stenosis was associated with higher hidden blood loss in patients with rheumatoid arthritis.[27] This study suggests that performing surgery earlier may reduce the probability of requiring intraoperative blood transfusions, as delays can complicate the disease, and more extensive surgical intervention would be needed.

Age was found to be an indicator. Older were more likely to require blood transfusion during PLIF, which aligns with previous research.[28] However, another study recognized older age as a predictor for postoperative but not intraoperative blood transfusion.[29] In older patients, intraoperative transfusion was linked to a higher risk of postoperative delirium.[30] Besides lower hemoglobin, preoperative coagulation status is crucial, higher PT, lower platelet count, and lower fibrinogen correlate with an increased possibility of intraoperative blood transfusion. Insufficient clotting ability would inevitably lead to increased blood loss, elevating the need for transfusion.

Albumin level reflects the nutrition status of individuals. Surprisingly, a higher albumin was associated with an increased probability of transfusion in our study. A previous study in the cardiac field reported lower pre-operative albumin as a significant predictor of intraoperative blood transfusion.[31] Low albumin levels have been linked to worse outcomes in PLIF, with hypoalbuminemia associated with extended hospital stays.[32] Intraoperative blood loss independently predicted the need for albumin infusion following PLIF.[33] In our study, low albumin levels led to less intraoperative blood transfusion, which may result from preoperative preparation for patients in worse overall conditions. More research is needed to further understand the behind-screen mechanism.

Though smoking did not emerge as a contributing factor to intraoperative blood transfusion, it is linked to greater blood loss and increased transfusion needs in lumbar spinal surgery.[34] Hypertension patients were also found to be more prone to transfusion during PLIF, negatively impacting recovery rates.[35] Moreover, a meta-analysis revealed that a history of smoking or hypertension was associated with adjacent segment degeneration (ASD).[36] Though sometimes intraoperative blood transfusion indicates complex surgical procedures

and poor patient conditions, it can be beneficial for patient outcomes. Patients receiving transfusion had better outcomes when their hemoglobin were maintained within the range of 7.5–11.5 g/dL compared to extreme values.[37]

In this study, we employed machine learning techniques to predict intraoperative blood transfusion during PLIF surgery, ensuring model explainability through SHAP and optimizing for clinical use by developing a web-based calculator. Future research could explore validating this model in different populations or regions, since patient characteristics and medical practices may vary. Additionally, other metrics beyond blood transfusion could be explored using machine learning approaches to make full use of the clinical utility of predictive models.

Conclusion

In our study, we successfully constructed a comprehensive machine learning model for intraoperative blood transfusion prediction for LDH patients in open PLIF surgery through a registered multicenter-based experience from China. The model was evaluated on a variety of test datasets and showed robust performance. The SHAP approach guarantees the established model with explainability and could prompt usage in real-world conditions. A publicly accessible web calculator was set up to enable clinical usage.

Limitations of the study

The loading speed of the online calculator is slow due to the complexity of the integrated machine learning model. This may require preloading action or will lead to delays in clinical settings. Future work should aim to optimize the model for quicker performance and explore advanced approaches to improve clinical utility.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, W.L. (drlee0910@163.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The codes used to generate the model are available in the following repository and publicly accessible: https://doi.org/10.5281/zenodo.13854235
- The datasets used and/or analyzed during the current study are available from the lead contact upon reasonable request.
- Any additional information required to reanalyze the data reported in this work article is available from the lead contact upon request.

ACKNOWLEDGMENTS

The graphical abstract and Figure 1 was created using BioRender.com.

Funding: Shaanxi Provincial Health and Health Research Fund Project (2022E006), and Liuzhou 2023 Central Guide Local Science and Technology Development Fund project (2023YRQO101).

AUTHOR CONTRIBUTIONS

Conceiving the study and design: W.L., Z.H., Q.L., and B.Q. Data collection: R.L., L.Y., X.Q., X.L., Yang Zhang, T.X., Yingang Zhang, A.C., H.J., X.H., Q.X., W.H., L.C., X.Z., Q.Z., W.H., H.L., X.S., X.Y., and X.X. Developed the machine learning model and data analysis: W.L. and A.-T.C. Drafting of the manuscript: Q.L., and A.-T.C.. Suggestions and check for the manuscript: K.W., S.-N.W., W.L., Y.Z., J.Z., H.D., C.X., C.Y. All authors have read and supervised the final manuscript and the process of research.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Patients selection
  - Basic and clinical characteristics
  - Data preprocessing and variable selection
  - Machine learning model construction and evaluation
  - Explainable machine learning approaches
  - Ethics approval and consent to participate
- METHOD DETAILS
  - Boruta algorithm
  - US
  - SA-SVM
  - GA-SVM
  - Lasso

- SVM-RFE
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111106.

Received: May 26, 2024
Revised: July 17, 2024
Accepted: October 1, 2024
Published: October 5, 2024

REFERENCES

1. Awad, J.N., and Moskovich, R. (2006). Lumbar disc herniations: surgical versus nonsurgical treatment. Clin. Orthop. Relat. Res. *443*, 183–197. https://doi.org/10.1097/01.blo.0000198724.54891.3a.
2. Fenton-White, H.A. (2021). Trailblazing: the historical development of the posterior lumbar interbody fusion (PLIF). Spine J. *21*, 1528–1541. https://doi.org/10.1016/j.spinee.2021.03.016.
3. de Kunder, S.L., van Kuijk, S.M.J., Rijkers, K., Caelers, I.J.M.H., van Hemert, W.L.W., de Bie, R.A., and van Santbrink, H. (2017). Transforaminal lumbar interbody fusion (TLIF) versus posterior lumbar interbody fusion (PLIF) in lumbar spondylolisthesis: a systematic review and meta-analysis. Spine J. *17*, 1712–1721. https://doi.org/10.1016/j.spinee.2017.06.018.
4. Sidhu, G.S., Henkelman, E., Vaccaro, A.R., Albert, T.J., Hilibrand, A., Anderson, D.G., and Rihn, J.A. (2014). Minimally invasive versus open posterior lumbar interbody fusion: a systematic review. Clin. Orthop. Relat. Res. *472*, 1792–1799. https://doi.org/10.1007/s11999-014-3619-5.
5. Shah, A., Palmer, A.J.R., and Klein, A.A. (2020). Strategies to minimize intraoperative blood loss during major surgery. Br. J. Surg. *107*, e26–e38. https://doi.org/10.1002/bjs.11393.
6. Meier, J., Filipescu, D., Kozek-Langenecker, S., Llau Pitarch, J., Mallett, S., Martus, P., and Matot, I.; ETPOS Collaborators (2016). Intraoperative transfusion practices in Europe. Br. J. Anaesth. *116*, 255–261. https://doi.org/10.1093/bja/aev456.
7. Wu, W.C., Trivedi, A., Friedmann, P.D., Henderson, W.G., Smith, T.S., Poses, R.M., Uttley, G., Vezeridis, M., Eaton, C.B., and Mor, V. (2012). Association between hospital intraoperative blood transfusion practices for surgical blood loss and hospital surgical mortality rates. Ann. Surg. *255*, 708–714. https://doi.org/10.1097/SLA.0b013e31824a55b9.
8. Lang, F.F., Liu, L.Y., and Wang, S.W. (2023). Predictive modeling of perioperative blood transfusion in lumbar posterior interbody fusion using machine learning. Front. Physiol. *14*, 1306453. https://doi.org/10.3389/fphys.2023.1306453.
9. Chen, C., Wang, Y., Yang, X., Zhang, M., He, J., Yang, L., Qin, L., Chen, B., Chen, B., and Wang, Q. (2024). Machine Learning-Based Prediction of Intraoperative Red Blood Cell Transfusion in Aortic Valve Replacement Surgery. Clin. Lab. *70*, 38623654. https://doi.org/10.7754/Clin.Lab.2023.230930.
10. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. N. Engl. J. Med. *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259.
11. Haug, C.J., and Drazen, J.M. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N. Engl. J. Med. *388*, 1201–1208. https://doi.org/10.1056/NEJMra2302038.
12. Pencina, M.J., Goldstein, B.A., and D'Agostino, R.B. (2020). Prediction Models — Development, Evaluation, and Clinical Application. N. Engl. J. Med. *382*, 1583–1586. https://doi.org/10.1056/NEJMp2000589.
13. Chen, A.-T., Zhang, Y., and Zhang, J. (2024). Explainable machine learning and online calculators to predict heart failure mortality in intensive care units. ESC Heart Fail. https://doi.org/10.1002/ehf2.15062.
14. Chen, Y., and Wu, M. (2023). Artificial Intelligence-enabled contactless sensing for medical diagnosis. Med. Rev. *3*, 195–197. https://doi.org/10.1515/mr-2023-0022.
15. Raj, G.M., Dananjayan, S., and Gudivada, K.K. (2024). Applications of artificial intelligence and machine learning in clinical medicine: What lies ahead? Med. Adv. *2*, 202–204. https://doi.org/10.1002/med4.62.
16. Xie, Z., and Li, H. (2023). Drug development accelerated by artificial intelligence. Med. Rev. *3*, 193–194. https://doi.org/10.1515/mr-2023-0024.
17. Ngiam, K.Y., and Khor, I.W. (2019). Big data and machine learning algorithms for health-care delivery. Lancet Oncol. *20*, e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4.
18. Kim, C., You, S.C., Reps, J.M., Cheong, J.Y., and Park, R.W. (2021). Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. J. Am. Med. Inform. Assoc. *28*, 1098–1107. https://doi.org/10.1093/jamia/ocaa277.
19. Kang, S., Cho, S., and Kang, P. (2015). Multi-class classification via heterogeneous ensemble of one-class classifiers. Eng. Appl. Artif. Intell. *43*, 35–43.
20. Zhai, B., and Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China. Sci. Total Environ. *635*, 644–658.
21. Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., and Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Appl. Soft Comput. *77*, 188–204.
22. Lei, C., Lu, Z., Wang, M., and Li, M. (2023). StackCPA: A stacking model for compound-protein binding affinity prediction based on pocket multi-scale features. Comput. Biol. Med. *164*, 107131. https://doi.org/10.1016/j.compbiomed.2023.107131.
23. Lin, S.-W., Lee, Z.-J., Chen, S.-C., and Tseng, T.-Y. (2008). Parameter determination of support vector machine and feature selection using simulated annealing approach. Appl. Soft Comput. *8*, 1505–1512. https://doi.org/10.1016/j.asoc.2007.10.012.
24. Chen, Y.-P., Li, Y., Wang, G., Zheng, Y.-F., Xu, Q., Fan, J.-H., and Cui, X.-T. (2017). A novel bacterial foraging optimization algorithm for feature selection. Expert Syst. Appl. *83*, 1–17. https://doi.org/10.1016/j.eswa.2017.04.019.
25. Medjahed, S.A., Ouali, M., Saadi, T.A., and Benyettou, A. (2015). An optimization-based framework for feature selection and parameters determination of svms. Int. J. Inf. Technol. Comput. Sci. *7*, 1–9.
26. Carabini, L.M., Moreland, N.C., Vealey, R.J., Bebawy, J.F., Koski, T.R., Koht, A., Gupta, D.K., and Avram, M.J.; Northwestern High Risk Spine Group (2018). A Randomized Controlled Trial of Low-Dose Tranexamic Acid versus Placebo to Reduce Red Blood Cell Transfusion During Complex Multilevel Spine Fusion Surgery. World Neurosurg. *110*, e572–e579. https://doi.org/10.1016/j.wneu.2017.11.070.
27. Xu, S., Liang, Y., Wang, J., Yu, G., Guo, C., Zhu, Z., and Liu, H. (2019). Blood Loss of Posterior Lumbar Interbody Fusion on Lumbar Stenosis in Patients With Rheumatoid Arthritis: A Case-Control Study. Spine *44*, E1045–E1052. https://doi.org/10.1097/brs.0000000000003037.
28. Kolbenschlag, J., Schneider, J., Harati, K., Goertz, O., Hernekamp, J.F., Kempf, R., Lehnhardt, M., and Hirsch, T. (2016). Predictors of Intraoperative Blood Transfusion in Free Tissue Transfer. J. Reconstr. Microsurg. *32*, 706–711. https://doi.org/10.1055/s-0036-1586255.
29. Ad, N., Massimiano, P.S., Burton, N.A., Halpin, L., Pritchard, G., Shuman, D.J., and Holmes, S.D. (2015). Effect of patient age on blood product transfusion after cardiac surgery. J. Thorac. Cardiovasc. Surg. *150*, 209–214. https://doi.org/10.1016/j.jtcvs.2015.03.022.
30. Chou, M.Y., Wang, Y.C., Peng, L.N., Liang, C.K., Chu, C.S., Liao, M.C., Lin, Y.T., Hsu, C.J., and Chen, L.K. (2019). Intraoperative blood transfusion predicts postoperative delirium among older patients undergoing elective orthopedic surgery: A prospective cohort study. Int. J. Geriatr. Psychiatr. *34*, 881–888. https://doi.org/10.1002/gps.5086.

31. Pennington, Z., Ehresman, J., Feghali, J., Schilling, A., Hersh, A., Hung, B., Lubelski, D., and Sciubba, D.M. (2021). A clinical calculator for predicting intraoperative blood loss and transfusion risk in spine tumor patients. Spine J. *21*, 302–311. https://doi.org/10.1016/j.spinee.2020.09.011.

32. Phan, K., Ranson, W., White, S.J.W., Cheung, Z.B., Kim, J., Shin, J.I., Ukogu, C., Lee, N.J., Kothari, P., and Cho, S.K. (2019). Thirty-Day Perioperative Complications, Prolonged Length of Stay, and Readmission Following Elective Posterior Lumbar Fusion Associated With Poor Nutritional Status. Global Spine J. *9*, 417–423. https://doi.org/10.1177/2192568218797089.

33. Liu, B., Pan, J., Zong, H., and Wang, Z. (2021). The risk factors and predictive nomogram of human albumin infusion during the perioperative period of posterior lumbar interbody fusion: a study based on 2015-2020 data from a local hospital. J. Orthop. Surg. Res. *16*, 654. https://doi.org/10.1186/s13018-021-02808-5.

34. McCunniff, P.T., Young, E.S., Ahmadinia, K., Ahn, U.M., and Ahn, N.U. (2016). Smoking is Associated with Increased Blood Loss and Transfusion Use After Lumbar Spinal Surgery. Clin. Orthop. Relat. Res. *474*, 1019–1025. https://doi.org/10.1007/s11999-015-4650-x.

35. Sakaura, H., Miwa, T., Yamashita, T., Kuroda, Y., and Ohwada, T. (2016). Lifestyle-Related Diseases Affect Surgical Outcomes after Posterior Lumbar Interbody Fusion. Global Spine J. *6*, 2–6. https://doi.org/10.1055/s-0035-1554774.

36. Wang, T., and Ding, W. (2020). Risk factors for adjacent segment degeneration after posterior lumbar fusion surgery in treatment for degenerative lumbar disorders: a meta-analysis. J. Orthop. Surg. Res. *15*, 582. https://doi.org/10.1186/s13018-020-02032-7.

37. Will, N.D., Kor, D.J., Frank, R.D., Passe, M.A., Weister, T.J., Zielinski, M.D., and Warner, M.A. (2019). Initial Postoperative Hemoglobin Values and Clinical Outcomes in Transfused Patients Undergoing Noncardiac Surgery. Anesth. Analg. *129*, 819–829. https://doi.org/10.1213/ane.0000000000004287.

38. Kursa, M.B., and Rudnicki, W.R. (2010). Feature selection with the Boruta package. J. Stat. Software *36*, 1–13.

39. Muthukrishnan, R., and Rohini, R. (2016). LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning (IEEE), pp. 18–20.

40. Feng, X., Hong, T., Liu, W., Xu, C., Li, W., Yang, B., Song, Y., Li, T., Li, W., Zhou, H., and Yin, C. (2022). Development and validation of a machine learning model to predict the risk of lymph node metastasis in renal carcinoma. Front. Endocrinol. *13*, 1054358. https://doi.org/10.3389/fendo.2022.1054358.

41. Li, W., Hong, T., Fang, J., Liu, W., Liu, Y., He, C., Li, X., Xu, C., Wang, B., Chen, Y., et al. (2022). Incorporation of a machine learning pathological diagnosis algorithm into the thyroid ultrasound imaging data improves the diagnosis risk of malignant thyroid nodules. Front. Oncol. *12*, 968784. https://doi.org/10.3389/fonc.2022.968784.

42. Ranstam, J., and Cook, J.A. (2018). LASSO regression. Br. J. Surg. *105*, 1348. https://doi.org/10.1002/bjs.10895.

43. Meiri, R., and Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications. Eur. J. Oper. Res. *171*, 842–858. https://doi.org/10.1016/j.ejor.2004.09.010.

44. Leardi, R., Boggia, R., and Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. J. Chemometr. *6*, 267–281. https://doi.org/10.1002/cem.1180060506.

45. (2018). Welcome to the SHAP documentation. https://shap.readthedocs.io/en/latest/.

46. Nohara, Y., Matsumoto, K., Soejima, H., and Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput. Methods Progr. Biomed. *214*, 106584.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Clinical trial		
Clinical trial identifier	ClinicalTrials.gov	ID: NCT05867732
Software and algorithms		
R (v4.3.2)	R CRAN	https://cran.r-project.org/
Codes and datasets	Zenodo	https://zenodo.org/records/13954463
caret R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/caret/index.html
randomForest R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/randomForest/index.html
nnet R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/nnet/index.html
DALEX R package (4.3.2)	R CRAN	https://cran.r-project.org/web/packages/DALEX/index.html
rms R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/rms/index.html
glmnet R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/glmnet/index.html
ggplot2 R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/ggplot2/index.html
fastshap R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/fastshap/index.html
tidymodels R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/tidymodels/index.html
stacks R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/tidymodels/index.html
bonsai R package (v4.3.2)	R CRAN	https://cran.r-project.org/web/packages/bonsai/index.html
ComplexHeatmap R package (v4.3.2)	Bioconductor	https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

A total of 6241 patients from 22 hospitals across China diagnosed with lumbar disc herniation and treated with PLIF were included. Among all participants, there were 3050 (49%) females and 3191 (51%) males. In the machine learning model construction, they were randomly divided into the training set (3,055) and 4 test sets, containing 1,031, 814, 623, and 718 participants respectively.

Patients selection

In this multi-center retrospective study registered with ClinicalTrials.gov (ID: NCT 05867732), inclusion criteria included: 1. Patients diagnosed with LDH by imaging and examinations and accompanied by clinical symptoms; 2. Patients who have received PLIF treatment; 3. Complete clinical records; 4. Blood test results prior to surgery, including blood routine, coagulation parameters, and liver and kidney function; and 5. Patients without other obvious surgical complications. Exclusion criteria included: 1. Patients without LDH; 2. Patients not treated with PLIF; 3. Accompanied by other obvious comorbidities or malignant tumors; 4. Lack of necessary medical records (>5% missing).

Basic and clinical characteristics

The basic characteristics, clinical conditions, and blood test results before surgery were collected. Basic characteristics included age, smoking, and hypertension. Clinical conditions included the number of levels of fusion, time from symptom onset to surgery, time from admission to surgery, lumbar spinal stenosis, function before surgery, previous surgery, American Society of Anesthesiologists (ASA) classification. Blood test results before surgery contained hemoglobin, albumin, activated partial thromboplastin time (APTT), prothrombin time (PT), white blood cell (WBC), platelet count, and fibrinogen.

Data preprocessing and variable selection

To address the issue of missing data, we employed the random forest approach to fill in the gaps. We then utilized multiple feature selection techniques, including Boruta, univariate screening (US), least absolute shrinkage and selection operator (Lasso), support vector machine recursive feature elimination (SVM-RFE), simulated annealing support vector machine recursive (SA-SVM), and genetic algorithm support vector machine recursive (GA-SVM). Furthermore, we tested all features without any selection as well as the intersection of features selected by the before-mentioned six methods.

The Boruta algorithm was chosen for variable selection based on its adaptability in handling missing values and noises in data. Boruta algorithm acts as a wrapper around the Random Forest classification algorithm and has been recognized for its robustness in dealing with

complex data from real-world scenarios.[38] Lasso regression, another feature selection technique used in predictive modeling for machine learning, has been widely adopted for variable selection.[39–41] Its objective is to identify the factors and their associated regression weights that result in a model optimized for minimum prediction error.[42]

The US was implemented following a cross-validation based on the SVM. Initially, an SVM model was developed and then trained using 5-fold cross-validation. The feature selection of the 5 SVM models were selected through the US. RFE, SA, and GA were based on SVM as well. RFE is a wrapper-type feature selection algorithm. It operates by including all features from the training dataset first and eliminating them successively until the desired number of features are left. SA, a leading stochastic search method,[43] has been reported to be beneficial in determining parameters and selecting features.[23] GA was created as a strategy for optimization and has been shown to be more efficient than traditional feature selection methods.[44]

Machine learning model construction and evaluation

The dataset was divided into a training set and four test sets based on different regions. Basic machine learning models were constructed using 10-fold cross-validation, We developed ten machine learning algorithms, including decision tree (DT), efficient neural network (ENet), k-nearest neighbors (KNN), light gradient boosting machine (LightGBM), logistic regression (LR), multilayer perceptron (MLP), random forest (RF), SVM, extreme gradient boosting (XGBoost), and Lasso meta-model based stacked ensemble model. The nine individual models were constructed using the tidymodels R package.

The comprehensive stacking model was built using the stack R package. First, the training set is divided into several groups (usually 5 or 10). A basic model, such as DT, is then fitted on 4/5 or 9/10 and used to predict the left group. The process is repeated for each group. Subsequently, the basic model is applied to the entire training dataset and predictions are made on the test set. This process was repeated for all other basic models. The predictions from all basic models were used to create an ensemble model, which was then applied to the test set.

We combined 8 different feature groups, including 6 groups using different feature selection methods, all original features, and the intersection of features. We used 10 machine learning models and tested the performance on 4 external evaluation datasets. Performance was evaluated comprehensively by the area under the curve of the receiver operating characteristic (AUC-ROC), calibration curve, and decision curve analysis (DCA). AUC-ROC indicated the accuracy of prediction, the calibration curve could reflect calibration ability, and DCA provided insights into the decision boundary in real-life settings. Finally, the model with the best balance of performance and computational efficiency was selected to create a web-based calculator for predicting blood transfusion after PLIF.

Explainable machine learning approaches

Explainability was an issue for machine learning models since the working mechanism and the correlation between the input and output were unclear to users. To avoid the black-box nature of machine learning models, we applied the Shapley Additive Explanations (SHAP) approach to make the final optimized model interpretable through the SHAP package in R. To be specific, SHAP is a game theoretic approach to explain the output of machine learning models.[45] SHAP interpretation was found to be in accordance with existing methods and applied to real data in hospitals.[46]

Ethics approval and consent to participate

Our study was approved by the institutional review board of all participating institutions (main institution ethical number: 2022-IRB-04).

METHOD DETAILS

The Boruta algorithm was chosen for variable selection based on its adaptability in handling missing values and noises in data. Boruta algorithm acts as a wrapper around the Random Forest classification algorithm and has been recognized for its robustness in dealing with complex data from real-world scenarios. Lasso regression, another feature selection technique used in predictive modeling for machine learning, has been widely adopted for variable selection. Its objective is to identify the factors and their associated regression weights that result in a model optimized for minimum prediction error.

The US was implemented following a cross-validation based on the SVM. Initially, an SVM model was developed and then trained using 5-fold cross-validation. The feature selection of the 5 SVM models were selected through the US. RFE, SA, and GA were based on SVM as well. RFE is a wrapper-type feature selection algorithm. It operates by including all features from the training dataset first and eliminating them successively until the desired number of features are left. SA, a leading stochastic search method, has been reported to be beneficial in determining parameters and selecting features. GA was created as a strategy for optimization and has been shown to be more efficient than traditional feature selection methods.

Boruta algorithm

Boruta algorithm creates copies of all features in the dataset called shadow features and randomizes them to remove any associations with the target variable. It applies a random forest classifier to the dataset, extended with these shadow features. The algorithm measures the importance of each original feature compared to the best of the shadow features, usually based on the Z score of feature importance. Features that are statistically more important than the best shadow feature are deemed important. The process is iteratively repeated until all features are either confirmed or rejected as important or the algorithm reaches a specified limit of iterations.

US

In the US feature selection, each feature is tested individually to assess its impact on the target variable. The selection process often involves statistical tests like t-tests for continuous data or chi-squared tests for categorical data to determine the significance of each feature. Features that meet a certain statistical significance threshold are selected, while others are discarded.

SA-SVM

Initialization starts with a random set of features. The cost function evaluation is performed by evaluating the performance of the SVM with the current feature set using a cost function. After that, it generates a "neighbor" solution by slightly altering the feature set, adding or removing features for instance. Evaluation is again performed on this new set using the same cost function. Referring to the acceptance criterion, it decides whether to move to the new set based on the cost and a probabilistic function. Repeat the above process until convergence criteria are met.

GA-SVM

The GA-SVM algorithm starts with a population of random feature sets. Each feature set is evaluated using the SVM model. It selects feature sets for reproduction based on accuracy. Crossover and mutation are performed by applying genetic operators like crossover (mixing features of two sets) and mutation (randomly altering features) to create a new generation. The evaluation, selection, and genetic operation steps are repeated until a stopping criterion is met.

Lasso

Lasso adds a regularization term to the regression model that penalizes the absolute size of the regression coefficients. During the training process, less important features' coefficients are shrunk toward zero. Features whose coefficients become exactly zero are eliminated from the model, effectively performing feature selection.

SVM-RFE

The SVM-RFE algorithm first trains an SVM model on the dataset. It ranks features based on the absolute value of their coefficients in the SVM model and removes the least important features, which are those with the smallest coefficients. Re-train the SVM model with the reduced set of features and repeat the ranking and elimination steps. Continue this process until the desired number of features is reached or another stopping criterion is met.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis in this study was conducted using R (version 4.3.2). Statistically significant was considered as a p-value less than 0.05. The tidymodels R package was utilized for individual model construction and the stack R package was used for the comprehensive stacking model.

ADDITIONAL RESOURCES

This was registered with ClinicalTrials.gov (ID: NCT05867732).