



OPEN Machine learning algorithms for prediction of cerebrospinal fluid leakage after posterior surgery for thoracic ossification of the ligamentum flavum

Ruizhou Guo^{1,2,3}, Ben Liu⁵, Yunqi Wu^{1,2,3}, Yilu Zhang^{1,2,3}, Xiyang Wang^{1,2,3}, Dingyu Jiang^{1,2,3,6}✉ & Zheng Liu^{4,6}✉

To develop and validate a machine-learning (ML) model that pre-operatively predicts cerebrospinal-fluid leakage (CSFL) after posterior decompression for thoracic ossification of the ligamentum flavum (TOLF), and to elucidate the key risk factors driving model decisions. Electronic medical-record and imaging data of 318 consecutive TOLF patients who underwent laminectomy between January 2009 and June 2023 were retrospectively analysed (CSFL = 101, 31.8%). The cohort was randomly split 4:1 into training (n = 254) and testing (n = 64) sets. Class imbalance was addressed with two synthetic oversampling techniques, SMOTE and ADASYN. A baseline logistic-regression (LR) model and four ML algorithms—XGBoost, Random Forest, LightGBM and Support Vector Machine (SVM)—were tuned via Bayesian optimisation. Primary endpoints were F1-score and recall; secondary metrics included AUC, accuracy, calibration curves and Brier scores. Probabilities were recalibrated with Platt Scaling and Isotonic Regression, and model interpretability was assessed with SHAP and LIME. Under SMOTE, SVM achieved the best overall performance (F1 = 0.889, recall = 0.881); its Brier score improved to 0.103 after Isotonic Regression. Feature-attribution analyses consistently identified multi-segment involvement, residual spinal-canal area (RrSCA) and related diametric ratios (RrPD, RrDCM), operative time, and intra-operative blood loss as the strongest predictors of postoperative CSFL. A SMOTE-enhanced, isotonic-calibrated SVM provides accurate and reliable CSFL risk estimation in TOLF patients and is freely available as an online tool (https://github.com/DebtVC2022/CSFL_predict). The model supports preoperative risk stratification, patient counselling, and peri-operative management, yet requires prospective, multicentre validation to establish broad clinical utility.

Keywords Machine learning, Logistic regression, Orthopaedics surgery, Ossification of the ligamentum flavum, Cerebrospinal fluid leakage

Ossification of the ligamentum flavum (OLF) is a pathological process characterized by ectopic ossification, which can lead to spinal canal stenosis and spinal cord compression [1], with the thoracic spine being the most frequently affected region [2]. When symptoms related to thoracic spinal cord involvement manifest—such as bilateral lower limb numbness and weakness, thoracic and back pain or discomfort, a sensation of tightness around the chest and abdomen, intermittent claudication, and disturbances in urination and defecation—conservative treatment often proves ineffective [3]. Posterior laminectomy is the main therapeutic approach for

¹Department of Spine Surgery and Orthopaedics, Xiangya Hospital of Central South University, Changsha 410008, People's Republic of China. ²Hunan Engineering Laboratory of Advanced Artificial Osteo-Materials, Xiangya Hospital of Central South University, Changsha 410008, People's Republic of China. ³National Clinical Research Center for Geriatric Diseases, Xiangya Hospital of Central South University, Changsha 410008, People's Republic of China. ⁴Department of Orthopedic, Hunan Provincial Key Laboratory of Pediatric Orthopedics, The Affiliated Children's Hospital of Xiangya School of Medicine, Central South University (Hunan Children's Hospital), 86# Ziyuan Road, Changsha 410007, Hunan, People's Republic of China. ⁵School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, People's Republic of China. ⁶Dingyu Jiang and Zheng Liu have contributed equally to the work. ✉email: 13687345560@163.com; liuzheng_0629@sina.com

symptomatic thoracic ossification of the ligamentum flavum (TOLF) [4]. Nevertheless, this surgical intervention is relatively complex and is associated with a spectrum of intraoperative complications, including dural tears and the cerebrospinal fluid leakage (CSFL) [5]. According to previous studies, the incidence of CSF leakage was higher in patients undergoing thoracic or lumbar surgery (2.65%~13%) [6]. CSFL not only hinders wound healing but may also lead to the formation of pseudomeningoceles and increase the risk of retrograde central nervous system infections [7]. Identifying a subset of patients at heightened risk for CSFL prior to surgical intervention is crucial for informed treatment decisions and effective expectation management. Recent studies have highlighted numerous patient factors associated with this complication, although most of these studies are limited to logistic regression (LR) models [6, 8, 9]. According to a meta-analysis reported by Jin et.al [9], risk factors such as smoking, spinal stenosis, and multiple surgical levels are all associated with CSFL.

In contrast, machine learning (ML) algorithms offer significant advantages in identifying risk factors and have been extensively applied in developing predictive models for various diseases and complications [10, 11]. ML can analyze complex interactions among multiple factors within the data and explore intricate linear or non-linear associations, thereby enhancing the accuracy of potential risk factor identification. Previous research has successfully demonstrated the application of ML models in predicting postoperative complications within the surgical domain [12, 13]. However, to date, there have been few studies utilizing machine learning models specifically to predict CSFL, especially in patients with thoracic ossification of the ligamentum flavum. As such, the purpose of this study was to (1) use machine learning algorithms to develop a model for predicting cerebrospinal fluid leakage complications in patients with ossification of the ligamentum flavum after posterior thoracic surgery and (2) deploy these predictive algorithms as an open-access digital application and we have open-sourced the code and models (https://github.com/DebtVC2022/CSFL_predict/tree/main).

Methods

Guidelines

The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines was followed for this study [14].

Study cohort

This study was limited to a retrospective examination of electronic health records at a prominent academic medical center in China. Based on previous studies [15, 16], the following inclusion and exclusion criteria were included. The inclusion criteria were as follows: 1) all patients had a definitive diagnosis of thoracic ossification of the ligamentum flavum, with conspicuous manifestations of spinal cord compression; 2) adequate clinical and imaging data must be available; and 3) all patients underwent posterior thoracic spine surgery (laminectomy) between January 2009 and June 2023.

The exclusion criteria were as follows: 1) combined surgeries, such as thoracolumbar combined procedures. 2) concomitant thoracic spinal trauma, deformities, tuberculosis, and neoplasms; and 3) previous history of thoracic spinal surgery.

From January 2009 and June 2023, there were 533 patients with thoracic ossification of the ligamentum flavum who underwent posterior thoracic surgery in our institution. Due to failure to meet the aforementioned criteria, 215 patients were excluded in this study. Finally, 318 eligible patients were enrolled in the study among whom 101 patients developed postoperative CSFL.

Data collection and definition

The primary outcome was the postoperative CSFL. Ascertainment of it postoperatively relied on an intricate scrutiny of the patient's electronic medical records. A postoperative CSFL was defined in this study as a cerebrospinal fluid discharge from the operative wound or cerebrospinal fluid collection by wound drain. In the cases where the nature of the fluid discharge was unclear, B-2 transferrin assays are used to distinguish cerebrospinal fluid from serous exudate. The following categories of features or variables were deemed as potential predictors for the risks of CSFL. Variables collected for the patients included 1) Demographics, 2) Radiographic parameters and classification, and 3) Surgical procedure (Table 1). Demographic variables included the patients' sex, age, body mass index (BMI), duration of symptoms (DOS), smoking and drinking history (SH and DH), and history of diabetes and hypertension. Smoking history was defined as smoking at least 20 cigarettes a day and abstinence for less than 3 months. A history of alcohol consumption was defined as a daily consumption of 50 g or more, or a frequency of drinking greater than 3 times a week and abstinence for less than 3 months.

The radiographic parameters and classification were evaluated by using preoperative thoracic CT scans. Radiographic variables included the Sato classification of OLF, involved segment of OLF (monosegment or multisegment(≥ 2)) (MoS/MuS), dural ossification (DO) and the indicators of spinal canal encroachment. The Sato classification of OLF included the lateral, extended, enlarged, fused, and tuberous types. In general, the degree of spinal canal stenosis progresses from lateral type to tuberous type and becomes progressively more severe.

The indicators of spinal canal encroachment include: the residual rate of spinal canal area (RrSCA), which is the ratio of the cross-sectional area at the most stenotic level of the spinal canal on axial CT images to the area of the spinal canal at the normal pedicle level. The residual rate of diameter of the canal of on the midline (RrDCM), which is the ratio of the vertical distance from the midpoint of the posterior vertebral margin to the ossified lesion at the most stenotic level on axial CT images to the normal anteroposterior diameter at the midline of the spinal canal. The residual rate of transverse and anteroposterior diameter (RrTAD), which is the ratio of the vertical distance from the posterior vertebral margin to the ossified lesion at the lateral boundary on axial CT images to the normal anteroposterior diameter at the lateral boundary. The residual rate of paramedian diameter (RrPD), which is the ratio of the vertical distance from the midpoint between the midline and the

Variable names and abbreviations	Type	Mean standard deviation or frequency
Age	Continuous (years)	52.5 (9.5)
Sex	Binary (Female/Male)	161 (50.6)/157 (49.4)
BMI	Continuous (kg/m ²)	24.5 (2.9)
Duration of symptoms	Continuous (months)	5.7 (4.5)
Smoking history (SH)	Binary (Yes/No)	129 (40.6)
Drinking history (DH)	Binary (Yes/No)	106 (33.3)
Diabetes	Binary (Yes/No)	55 (17.3)
Hypertension	Binary (Yes/No)	104 (32.7)
Radiographic parameters		
Mono-segment (MoS)	Binary (Yes/No)	89 (28.0)
Multi-segment (≥ 2) (MuS)	Binary (Yes/No)	229 (72.0)
Dural ossification (DO)	Binary (Yes/No)	91 (28.6)
Indicators of spinal canal encroachment		
The residual rate of spinal canal area (RrSCA)	Continuous (%)	56.0 (15.9)
The residual rate of diameter of the canal of on The midline (RrDCM)	Continuous (%)	60.5 (15.3)
The residual rate of transverse and anteroposterior diameter (RrTAD)	Continuous (%)	52.8 (8.0)
The residual rate of paramedian diameter (RrPD)	Continuous (%)	46.8 (9.6)
The residual rate of the sagittal diameter (RrSD)	Continuous (%)	43.8 (10)
Sato classification of OLF		
Lateral type (LaT)	Binary (Yes/No)	38 (11.9)
Extended type (ExT)	Binary (Yes/No)	54 (17.0)
Enlarged type (EnT)	Binary (Yes/No)	59 (18.6)
Fused type (FuT)	Binary (Yes/No)	68 (21.4)
Tuberous type (TuT)	Binary (Yes/No)	99 (31.1)
Surgical procedure		
Decompression instrument		
Conventional bone chisels	Binary (Yes/No)	121 (38.1)
High speed drill	Binary (Yes/No)	102 (32.1)
Piezosurgery	Binary (Yes/No)	95 (29.8)
Intraoperative blood loss, IBL	Continuous (mL)	517.5 (220.0)
Operation time, OT	Continuous (mins)	222.8 (61.7)

Table 1. Basic clinical data of the patient cohort. BMI: body mass index; OLF: Ossification of the Ligamentum Flavum. Continuous variables are presented as mean \pm standard deviation, while categorical variables are represented by their frequency and proportion (in percentage) within the patient cohort.

lateral boundary to the ossified lesion to the normal anteroposterior diameter at the paramedian line. The residual rate of the sagittal diameter (RrSD), which is the ratio of the most prominent distance of the ossified lesion from the posterior vertebral margin on sagittal CT images to the normal anteroposterior diameter of the spinal canal. The normal anteroposterior diameter and area of the spinal canal are obtained by averaging the sum of the normal anteroposterior diameters and spinal canal areas of the adjacent upper and lower segments. Those measurements are visually to present in eFigure 1 (Supplement material). For patients with multilevel ossification of the ligamentum flavum, the level with the lowest residual ratio of spinal canal area is selected for analysis.

All patients were received posterior thoracic laminectomy. Three decompression instruments (DI), including traditional bone chisels, high speed drill and piezosurgery (DI-1, DI-2, and DI-3) were used to remove the ossified ligamentum flavum in the spinal canal. Unfortunately, it was difficult to randomly select a surgical decompression method in clinical practice. Therefore, in our study, all patients treated with high speed drill and piezosurgery were collected more recently, while the patients treated with conventional bone chisels were collected earlier. In addition, surgical data included operation time (OT) and intraoperative blood loss (IBL).

Data analysis

The original dataset was randomly divided into a training set and a test set according to a ratio of 4:1. This resulted in a training set that contained data from 255 patients and a testing set that contained data from 64 patients. When splitting the training set and the test set in a 4:1 ratio, we ensured that the proportion of CSFL patients and non-CSFL patients in both the training set and the test set was consistent with the original dataset. Therefore, the number of individuals with CSFL in the training set is 81, and the number in the test set is 20. Furthermore, to eliminate the impact of the units of measure, all continuous variables were normalized ((x-mean)/variance). No preprocessing was performed for categorical variables.

Feature engineering is required for traditional machine learning models, which use data preprocessing techniques to handle continuous and discrete variables. Specifically, for this dataset, continuous variables such

as BMI, the indicators of spinal canal encroachment, duration of operation, and intraoperative blood loss are normalized and binned. For discrete variables such as Sex, we apply one-hot encoding.

Multicollinearity control and feature selection strategy

To address the potential issue of multicollinearity—particularly among one-hot encoded, mutually exclusive categorical features such as segment involvement, decompression instrument type (DI-1, DI-2, DI-3), and ossification morphology (LaT, ExT, EnT, FuT, TuT)—we computed the variance inflation factor (VIF) for each variable (Table 2). Several of these variables exhibited extremely high VIF values (some exceeding 300), indicating strong structural collinearity due to their categorical exclusivity.

Rather than removing variables, we applied L2 regularization (ridge penalty) across all models to reduce the impact of correlated features by shrinking their coefficients. This approach has been shown to stabilize models in the presence of multicollinearity effectively. Importantly, we did not observe any signs of performance degradation (e.g., instability in AUC, F1, or recall) and thus opted to retain all clinically relevant variables in the final models. This strategy ensures both model robustness and comprehensive utilization of domain-specific information.

Model training and testing

To develop predictive models for postoperative CSFL, we implemented five classifiers: XGBoost [17], Random Forest (RF) [18], LightGBM [19], Support Vector Machine (SVM) [20], and a baseline LR model. All models were trained on the same training dataset using stratified cross-validation to preserve class proportions and improve generalizability. Hyperparameter tuning was conducted via Bayesian optimization, with the F1 score used as the objective metric due to its relevance in imbalanced medical prediction tasks.

We employed five widely used metrics from the confusion matrix to evaluate classification performance under imbalanced conditions. In this context, TP (True Positives) refers to correctly predicted CSFL cases, TN (True Negatives) refers to correctly predicted non-CSFL cases, FP (False Positives) refers to incorrectly predicted CSFL when no leak occurred, and FN (False Negatives) refers to missed actual CSFL cases.

Based on these definitions, the metrics are calculated as follows:

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

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1 score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$\text{AUC} = \int_0^1 \frac{TP}{TP + FN}(t) \frac{FP}{FP + TN}(t) dt.$$

In clinical decision-making, recall and F1-score are particularly critical. Recall measures the proportion of true CSFL cases correctly identified, which is essential to minimize the risk of missed diagnoses that could lead to complications such as infection or reoperation. However, excessive recall may come at the cost of increased false positives. We emphasized the F1 score to balance this trade-off, harmonizing recall and precision into a single metric. AUC was additionally used to assess the model's discrimination capacity across different probability thresholds, offering a threshold-independent summary of classification quality. Accuracy, though included, was interpreted cautiously due to its sensitivity to class imbalance.

Two synthetic oversampling techniques were employed to address the pronounced class imbalance (CSFL prevalence: 31.8%): SMOTE [21] and ADASYN [22]. These methods generated additional minority class samples to improve model sensitivity. Specifically, SMOTE creates new examples through interpolation between nearest neighbours, while ADASYN adaptively generates more samples in harder-to-learn regions. Both methods were applied separately to evaluate their effects across all classifiers.

To ensure the reliability of predicted probabilities for clinical application, we applied two post hoc calibration methods—platt scaling [23] and isotonic regression [24]. For each trained model, probability calibration was

Feature	Age	Sex	BMI	DOS	SH	DH	Diabetes	Hypertension	MoS
VIF	1.4947	1.8353	1.3516	1.1046	1.4971	1.3612	1.3005	1.2914	68.0658
Feature	MuS	DO	LaT	ExT	EnT	FuT	TuT	RrSCA	RrDCM
VIF	68.3879	1.2397	36.7899	47.4002	53.6855	58.8639	74.1363	8.5221	4.9457
Feature	RrTAD	RrPD	RrSD	DI-1	DI-2	DI-3	OT	IBL	
VIF	1.4771	4.6008	4.6269	338.2821	310.3377	294.3985	2.0078	1.9343	

Table 2. Variance inflation factor (VIF) values.

assessed through calibration curves and Brier scores. Isotonic regression generally yielded more substantial improvements, particularly in models trained with SMOTE.

Feature importance analysis was also conducted to interpret model decisions. We used gain or Gini-based rankings for tree-based models, while coefficient magnitudes were used for linear models such as LR and SVM. In addition, SHAP (Shapley Additive explanations) values were computed to assess global feature contributions [25], and LIME (Local Interpretable Model-agnostic Explanations) was used to examine local interpretability on a per-sample basis [26]. While decision tree structures for RF, XGBoost, and LightGBM were not visualized, attention was focused on feature attribution patterns to aid clinical interpretability.

This study employed an integrated modelling pipeline involving advanced resampling strategies, probabilistic calibration, and multi-level interpretability tools to build and evaluate machine learning models for CSFL prediction.

Results

Cohort characteristics

The baseline characteristics of the patient cohort are presented in Table 1. To investigate the interrelationships among various clinical variables, we developed a visual representation of the correlation coefficients, as depicted in Fig. 1. The study population comprised 101 patients diagnosed with postoperative CSFL, representing 31.8% of the total cohort. The cohort had a mean age of 52.5 years (9.5) and a mean body mass index (BMI) of 24.5 (2.9). Among the 318 participants, 157 (49.4%) were male. The average duration of symptoms prior to intervention was 5.7 months (4.5). The cohort included 129 (40.6%) patients with a history of smoking and 106 (33.3%) with a history of alcohol consumption. Comorbid conditions included diabetes in 55 patients (17.3%) and hypertension in 104 patients (32.7%).

Based on surgical characteristics, 89 patients (28.0%) were classified into the Mono-segment group, while the remaining 229 patients (72.0%) were assigned to the Multi-segment (≥ 2) group. Dural ossification concomitant with ossification of the ligamentum flavum (OLF) was identified in 91 patients (28.6%).

Spatial canal morphology was quantitatively assessed through multiple parameters. The mean residual rates were as follows: RrSCA = 56.0% (SD = 15.9), RrDCM = 60.5% (SD = 15.3), RrTAD = 52.8% (SD = 8.0), RrPD = 46.8% (SD = 9.6), and RrSD = 43.8% (SD = 10.0). According to the Sato classification system for OLF, the patient distribution across types was as follows: lateral (38 patients, 11.9%), extended (54 patients, 17.0%), enlarged (59 patients, 18.6%), fused (68 patients, 21.4%), and tuberous (99 patients, 31.1%).

All patients underwent posterior thoracic laminectomy, with three distinct decompression instruments utilized in 121 (38.1%), 102 (32.1%), and 95 (29.8%) cases, respectively. The surgical procedures were characterized by an average intraoperative blood loss of 517.5 mL (220.0) and a mean operation time of 222.8 min (61.7).

Algorithm performance

The LR, XGBoost, RF, LightGBM, and SVM models were trained using a training dataset of 254 patients for CSFL prediction, employing two class-imbalanced data sampling methods, SMOTE and ADASYN. Tables 3 and 4 show the performance of the models under the two data sampling methods, including AUC, accuracy, recall, precision, F1 score, platt scaling score, and isotonic regression score. In this study, the LR model was used as the reference model, and its results in both tables showed relatively poor predictive performance, with recall rates of only 0.7857 and 0.6579, respectively. However, the AUCs remained above 0.8 (Tables 3, 4, and Fig. 1).

As mentioned in the previous methods section, F1 score and recall are the most crucial indicators in the context of medical decision-making. This is because, despite the application of SMOTE and ADASYN to mitigate class imbalance, the underlying clinical distribution of postoperative CSFL remains highly skewed, and the cost of false negatives (i.e., missing an actual CSFL) can be clinically catastrophic, potentially leading to infection, pseudo meningocele, or the need for reoperation. Therefore, maximizing recall ensures the highest capture rate of true positive cases. However, recall alone may lead to overcalling, introducing unnecessary surgical interventions. The F1 score, which balances precision and recall, becomes especially relevant in this trade-off. Meanwhile, accuracy and AUC are less informative in this context, as they can be inflated by correct predictions of the majority class or overly optimistic probability estimates.

Across both sampling strategies, the SVM model consistently delivered the best balance of sensitivity and precision (highest F1), making it our preferred algorithm for CSFL prediction. Under SMOTE, SVM achieved an F1 of 0.889, whereas ADASYN's F1 dropped slightly to 0.878—both still outperforming other models. Moreover, every model's key metrics (accuracy, AUC, recall, precision, F1) were uniformly higher when trained on SMOTE-augmented data versus ADASYN-augmented data. This suggests that, for our dataset, SMOTE generates synthetic minority-class examples that better capture the true decision boundary than ADASYN's density-based sampling, yielding stronger overall discrimination and calibration.

The sampling mechanisms of SMOTE and ADASYN can further explain the observed difference. Given a minority-class instance x and one of its k -nearest majority neighbours $x_{(i)}$, SMOTE generates synthetic samples using:

$$\tilde{x} = x + \lambda (x_{(i)} - x), \lambda \sim U(0, 1).$$

This method evenly interpolates between minority-class neighbours, preserving geometric structure.

In contrast, ADASYN emphasizes instances with greater class overlap. For each minority sample, the local imbalance ratio is defined as:

$$r_i = \frac{d_i}{k}, \text{ where } d_i = \# \text{ of majority - class neighbors of } x_{(i)}.$$

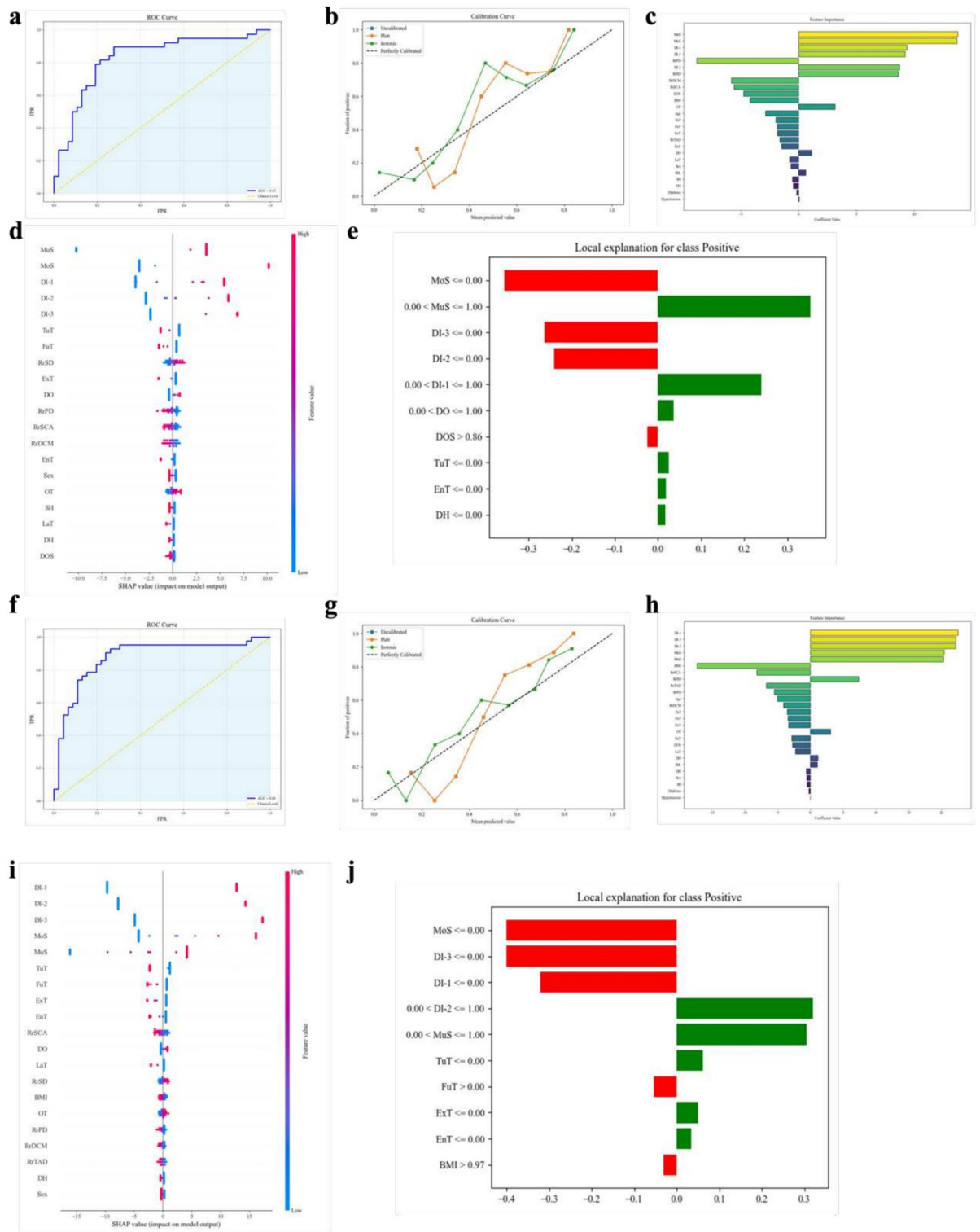


Fig. 1. Subfigure (a) and (f): The ROC curve of LR model under ADASYN or SMOTE. (b) and (g) Calibration curve of LR model under ADASYN or SMOTE. (c) and (h) The importance of features in the LR model under ADASYN or SMOTE. (d) and (i) The SHAP value of LR model under ADASYN or SMOTE. (e) and (j) The LIME value of LR model under ADASYN or SMOTE.

Model	Logistic regression	XGBoost	Random forest	LightGBM	Support vector machine
Accuracy	0.7412	0.8471	0.8471	0.8353	0.8824
Recall	0.6579	0.8421	0.8421	0.8421	0.9474
Precision	0.7353	0.8205	0.8205	0.8000	0.8182
F1 score	0.6944	0.8312	0.8312	0.8205	0.8780
AUC	0.8259	0.8863	0.9059	0.8919	0.8763
Platt scale	0.1744	0.1298	0.1124	0.1204	0.1380
Isotonic regression	0.1644	0.1375	0.1097	0.1216	0.1188
Uncalibrated	0.1742	0.1315	0.1124	0.1228	0.1371

Table 3. Comparison of the average resampled AUCs, accuracies, recalls, precisions, and F1-score of all machine learning and logistic regression models when applied to the testing set under ADASYN. Outcome: the occurrence of postoperative CSFL.

Model	Logistic regression	XGBoost	Random forest	LightGBM	Support Vector Machine
Accuracy	0.8182	0.8523	0.8864	0.8636	0.8864
Recall	0.7857	0.8571	0.8571	0.8333	0.881
Precision	0.825	0.8372	0.9	0.875	0.881
F1 score	0.8049	0.8471	0.878	0.8537	0.8889
AUC	0.8773	0.9348	0.9462	0.9161	0.9301
Platt Scale	0.1548	0.0981	0.0995	0.1063	0.1177
Isotonic Regression	0.1362	0.0971	0.0937	0.1017	0.1030
Uncalibrated	0.1548	0.0946	0.0995	0.0978	0.1071

Table 4. Comparison of the average resampled AUCs, accuracies, recalls, precisions, and F1-score of all machine learning and logistic regression models when applied to the testing set under SMOTE. Outcome: the occurrence of postoperative CSFL.

The number of synthetic samples for each instance is then:

$$g_i = \frac{r_i}{\sum_j r_i} \cdot G,$$

Where G is the total number of synthetic examples. ADASYN thus places greater emphasis on borderline or sparsely distributed minority cases.

While ADASYN can improve recall for hard-to-learn minority regions, its tendency to generate samples near decision boundaries may introduce noise and compromise precision. In contrast, SMOTE's balanced interpolation produces smoother margins that benefit tree-based models (e.g., RF and XGBoost) and margin-based classifiers like SVM. This helps explain why all five models, especially SVM, achieved better F1 and recall under SMOTE.

In practical terms, SVM under SMOTE is ideal when the cost of missing a CSFL case is unacceptable, as it yields the highest recall and F1 score. RF under SMOTE may be preferred when avoiding false positives is prioritized due to higher precision. XGBoost under SMOTE, which achieved the second-highest AUC after RF, may be optimal when probability ranking or clinical risk stratification is the goal (Fig. 2). Though ADASYN generally underperforms in this dataset, it remains useful when minority cases lie in highly sparse feature regions or represent subtle subphenotypes requiring targeted detection.

These results emphasize the importance of choosing evaluation metrics and resampling strategies aligned with clinical priorities. Recall and F1 remain the most informative metrics despite synthetic balancing in imbalanced, high-risk clinical prediction tasks like postoperative CSFL.

Calibration analysis

Calibrating a prediction algorithm is as important as its discrimination because accurate probability estimates underlie sound clinical decision-making. Calibration measures the alignment between predicted probabilities and observed outcome frequencies: the closer the calibration curve is to the 45° diagonal, the more reliable the model's risk estimates.

We applied Platt scaling and isotonic regression to each model under SMOTE and ADASYN sampling to enrich our calibration analysis. We compared the resulting Brier scores to the original, uncalibrated values plotted in Fig. 3. Under SMOTE, isotonic regression yielded the most considerable improvements—reducing the Brier score from 0.1548 to 0.1362 for LR, from 0.1071 to 0.1030 for SVM, and from 0.0995 to 0.0937 for Random Forest. In contrast, platt scaling offered no meaningful benefit or even slightly worsened calibration for most models. Under ADASYN, isotonic regression again delivered the most significant gains—lowering

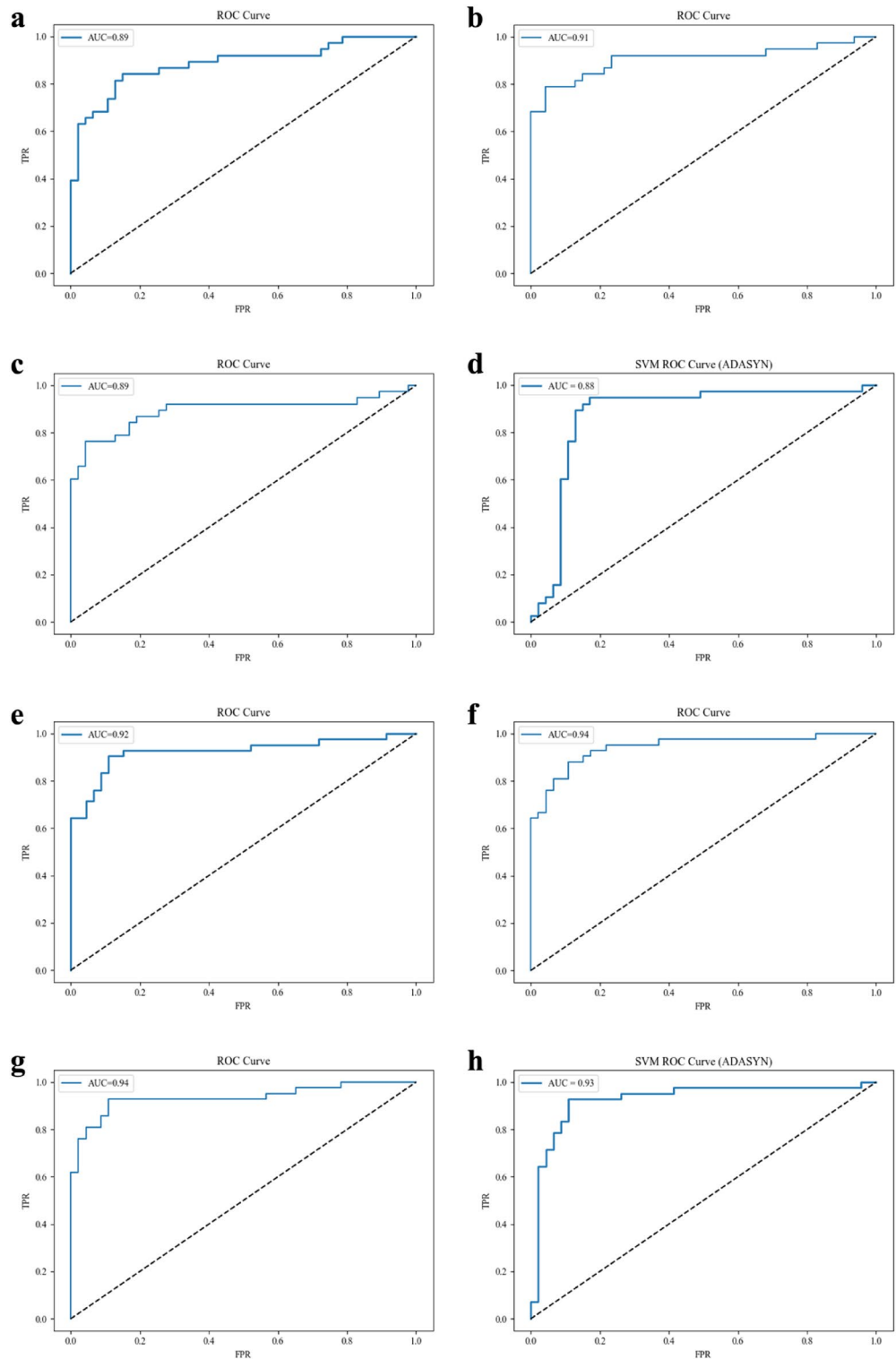


Fig. 2. The ROC curves and their AUC values for the four machine learning algorithms on the test set under ADASYN or SMOTE (**a–d**: XGBoost, RF, LightGBM, Support Vector Machine for ADASYN. **e–h**: XGBoost, RF, LightGBM, Support Vector Machine for SMOTE.). The RF model showed the highest AUC value, which was 0.95.

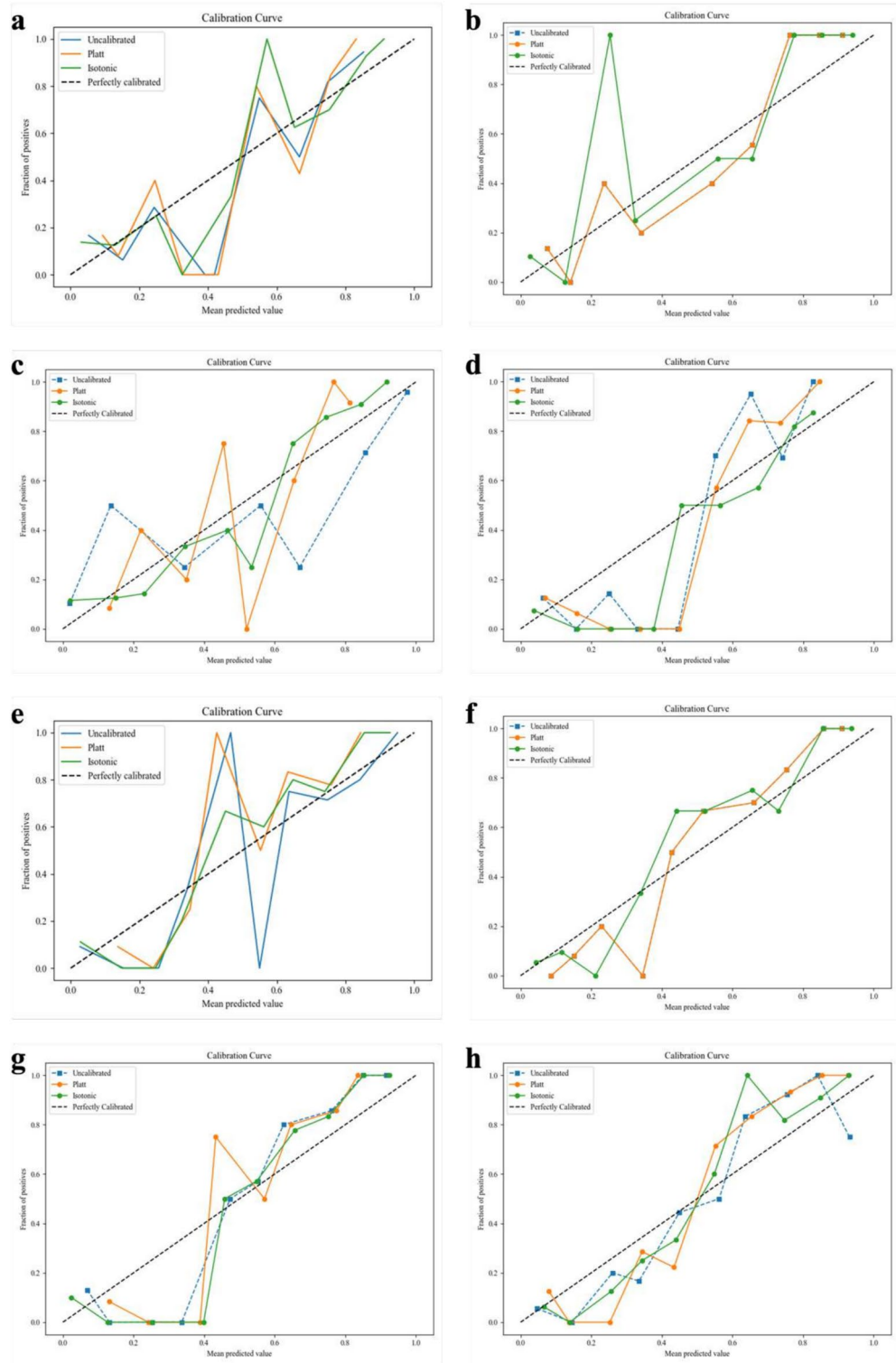


Fig. 3. The Calibration curve of four machine learning algorithms on the test set under ADASYN or SMOTE (a–d: XGBoost, RF, LightGBM, Support Vector Machine for ADASYN. e–h: XGBoost, RF, LightGBM, Support Vector Machine for SMOTE.)

Brier from 0.1742 to 0.1644 for LR, from 0.1371 to 0.1188 for SVM, and from 0.1124 to 0.1097 for Random Forest—while platt scaling provided only modest improvements for XGBoost (0.1315 → 0.1298) and LightGBM (0.1228 → 0.1204).

Visual inspection of the post-calibration curves (Fig. 3) confirms these findings: isotonic regression realigns predicted probabilities more closely to the ideal diagonal—especially for SVM and Random Forest—whereas

platt scaling often leaves residual over- or under-confidence in XGBoost and LightGBM. This pattern holds across both SMOTE and ADASYN, indicating that isotonic regression is the preferred calibration method for our dataset, particularly when used in conjunction with SMOTE's more evenly distributed synthetic samples.

Properly calibrated probabilities are indispensable in medical contexts: over-confident models risk unnecessary interventions, while under-confidence may fail to flag high-risk patients. Incorporating isotonic regression post-processing will, therefore, enhance the trustworthiness of our CSFL risk estimates. However, practitioners should remain aware that the choice of sampling strategy (SMOTE vs. ADASYN) and model class materially affect calibration quality.

Feature importance

Calculate the importance of variables to determine the most important predictors in the final machine learning model. Important variable screening helps identify the features that contribute most to the prediction results and highlights the degree of association between these features and the target. Figures 4, 5, and 6 show the important features sorted based on model coefficients, SHAP values, and LIME values under two class imbalance treatment methods. In these figures, the features are sorted from large to small according to their contribution, and the features at the top contribute more to the model prediction. In addition, the SHAP value breaks down the global prediction of the model into the “marginal contribution” of each feature to the prediction result. A positive value indicates that the feature pushes the prediction in the positive direction (such as higher risk), while a negative value indicates that it pushes in the negative direction. The absolute value reflects the strength of the influence. The LIME value estimates the contribution weight of each feature near the point by perturbing the original model in a local area and fitting a simple interpretable model (usually a linear model). A positive value indicates that the feature increases the probability of the positive class in the local area. In contrast, a negative value indicates that the probability of the positive class is reduced. The size indicates the influence of the feature in this local linear approximation model (Fig. 7).

Figure 7c shows that the importance of the basic feature results in LR model prioritizing parameter values based on regression results. In addition, the calculation results of the two feature contribution degrees, SHAP value and LIME value, are further supplemented. Combining the results of the three evaluation methods, we found that MuS, MoS, DI-1, DI-2, and DI-3 features ranked in the top five in regression coefficients, LIME local explanations, and SHAP global contributions under both resampling strategies (ADASYN and SMOTE), and can be regarded as the most robust predictors of CSFL risk. Among them, MuS/MoS ranked first with the highest weight under ADASYN, indicating that multi-segment or single-segment lesions significantly increase the risk of postoperative leakage. While under SMOTE, the coefficients and explanation values of the three instruments (DI-1, DI-2, and DI-3) are fully leading, which means that the resection method is particularly critical to the risk of CSFL. Although auxiliary features such as RrPD, RrSD, and BMI occasionally appear in the regression coefficient diagram, their contributions in LIME and SHAP are second-level objects of concern. In summary, clinical practice should focus on the number of lesion segments and instrument selection. These two factors are unanimously recognized under various balance and interpretation frameworks, providing the most reliable decision-making basis for preoperative risk assessment and optimization of surgical plans.

The importance of the baseline feature of XGBoost is determined based on the calculated gain value [17]. The higher the gain value, the higher the importance. In addition, the SHAP and LIME values are also used to characterize the feature's contribution to this model. As shown in Figs. 4a, 5a, and 6a, many important predictors of postoperative CSFL exist. Regardless of whether ADASYN or SMOTE resampling is used, the XGBoost model regards intraoperative blood loss and operation time as the most critical risk drivers. Under ADASYN, the RrSCA follows closely, while dural ossification is promoted to third place under SMOTE. Heavy bleeding and long operation time are the most robust factors for predicting postoperative CSFL. For the ADASYN method, it is necessary to focus on the degree of spinal canal stenosis, while for the SMOTE method, dural ossification should be particularly vigilant. Others, such as RrPD, BMI and drinking history, contributed to some evaluations, but their importance was relatively minor.

Furthermore, as shown in Figs. 4b, 5b, and 6b, through the comprehensive evaluation of the three perspectives of Gini coefficient ranking [18], SHAP global contribution, and LIME local interpretation, it can be seen that the RF model regards intraoperative blood loss, operation time, and the RrSCA as the most robust risk drivers under both ADASYN and SMOTE resampling strategies. The RrPD and RrSD follow closely behind. Although dural ossification and BMI appear in some methods, they are not listed simultaneously in the three explanation methods, so they are secondary objects of concern. This shows that no matter which category balance method is used, intraoperative bleeding control, operation time optimization, and spinal canal stenosis assessment are all key links in reducing the risk of postoperative CSFL.

As shown in Figs. 4c, 5c and 6c, the comprehensive LightGBM model under the two balanced strategies of ADASYN and SMOTE is based on information gain (or Gini reduction) [19], SHAP global contribution and LIME local explanation. It can be seen that intraoperative blood loss (IBL), operation time and the RrSCA are consistently ranked in the top three, which can be regarded as the most robust risk drivers. BMI and dural ossification are also ranked in the top 5 in both strategies, indicating that the patient's body mass index and combined DO status significantly impact the risk of CSFL. The rest, such as RrTAD, RrPD, RrDCM and duration of symptoms, have made outstanding contributions in evaluating some methods. However, they failed to be listed simultaneously in the three explanation methods and can be regarded as secondary objects of concern. This suggests that the LightGBM model believes that controlling intraoperative blood loss, compressing surgical time, and evaluating the degree of spinal canal stenosis while considering the patient's BMI and dural ossification status are the keys to optimizing postoperative CSFL risk.

Finally, the importance features calculated based on the dual coefficient [27], SHAP value, and LIME value of SVM are shown in Figs. 4d, 5d and 6d. As shown in the figure, ADASYN resampling makes multi-segment

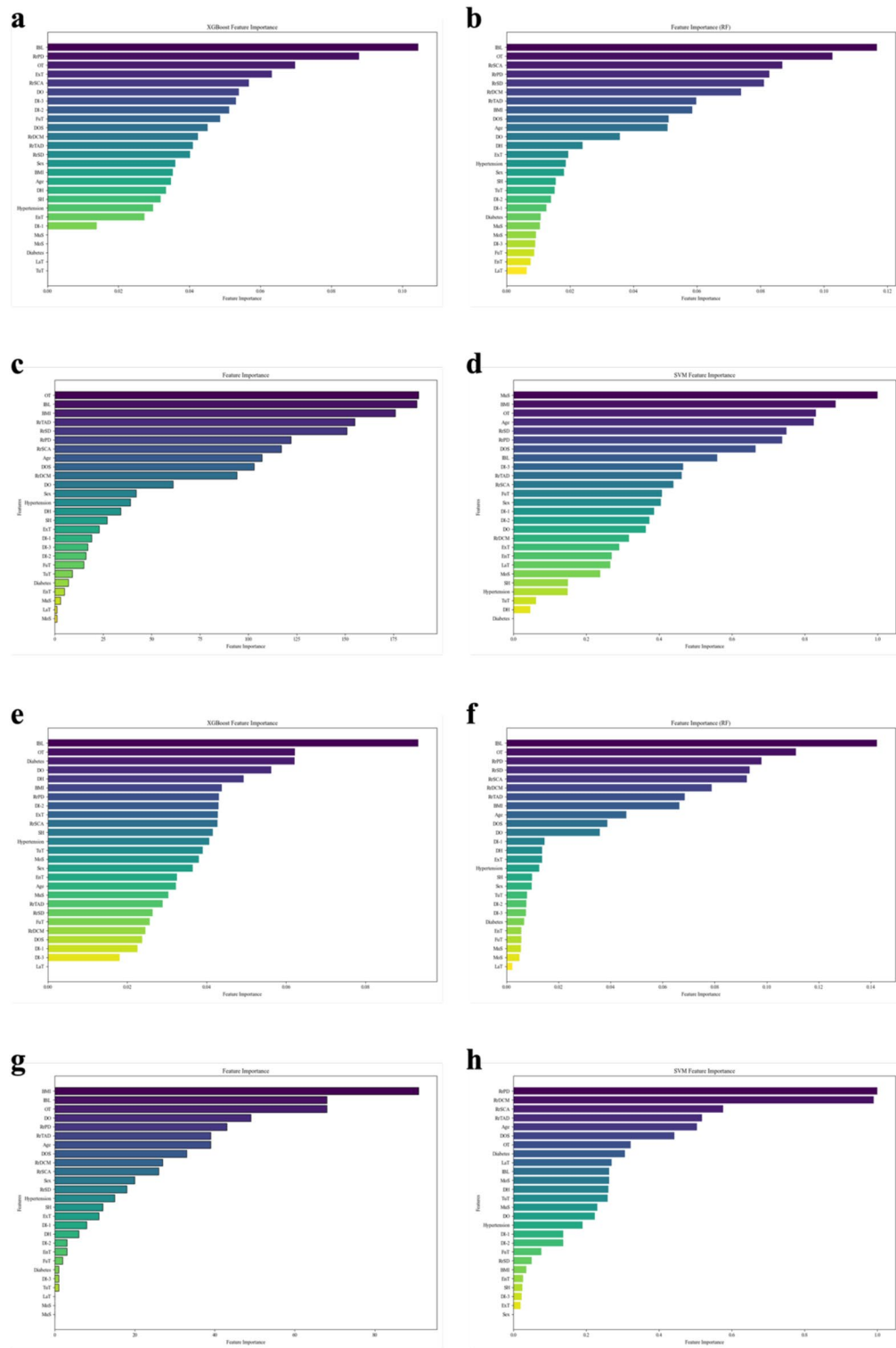


Fig. 4. Subfigures (a), (b), (c) and (d) in this figure represent the importance of features in the XGBoost, RF, LightGBM and Support Vector Machine under ADASYN, respectively. Subfigures (e), (f), (g) and (h) in this figure represent the importance of features in the XGBoost, RF, LightGBM and Support Vector Machine under SMOTE, respectively. Higher bars indicate that features are more important and contribute more to the predictive power of the model.

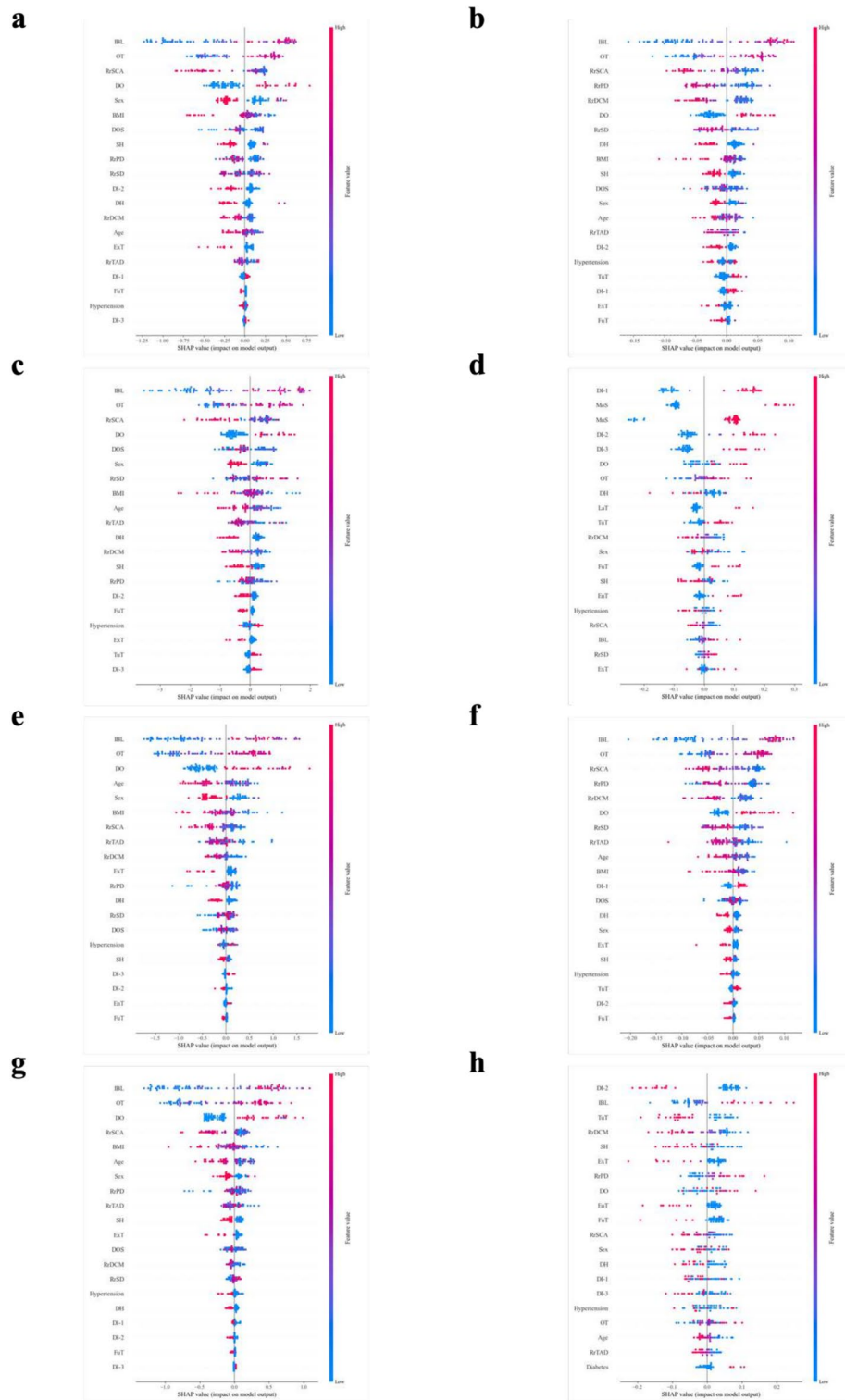


Fig. 5. Subfigures (a), (b), (c) and (d) in this figure represent the SHAP value in the XGBoost, RF, LightGBM and Support Vector Machine under ADASYN, respectively. Subfigures (e), (f), (g) and (h) in this figure represent the SHAP value in the XGBoost, RF, LightGBM and Support Vector Machine under SMOTE, respectively. The wider the point cloud, the more important the feature is and the greater its contribution to the model's prediction ability. The further the point cloud extends to the left, the greater the impact of the feature on the model's prediction of 0, and vice versa.

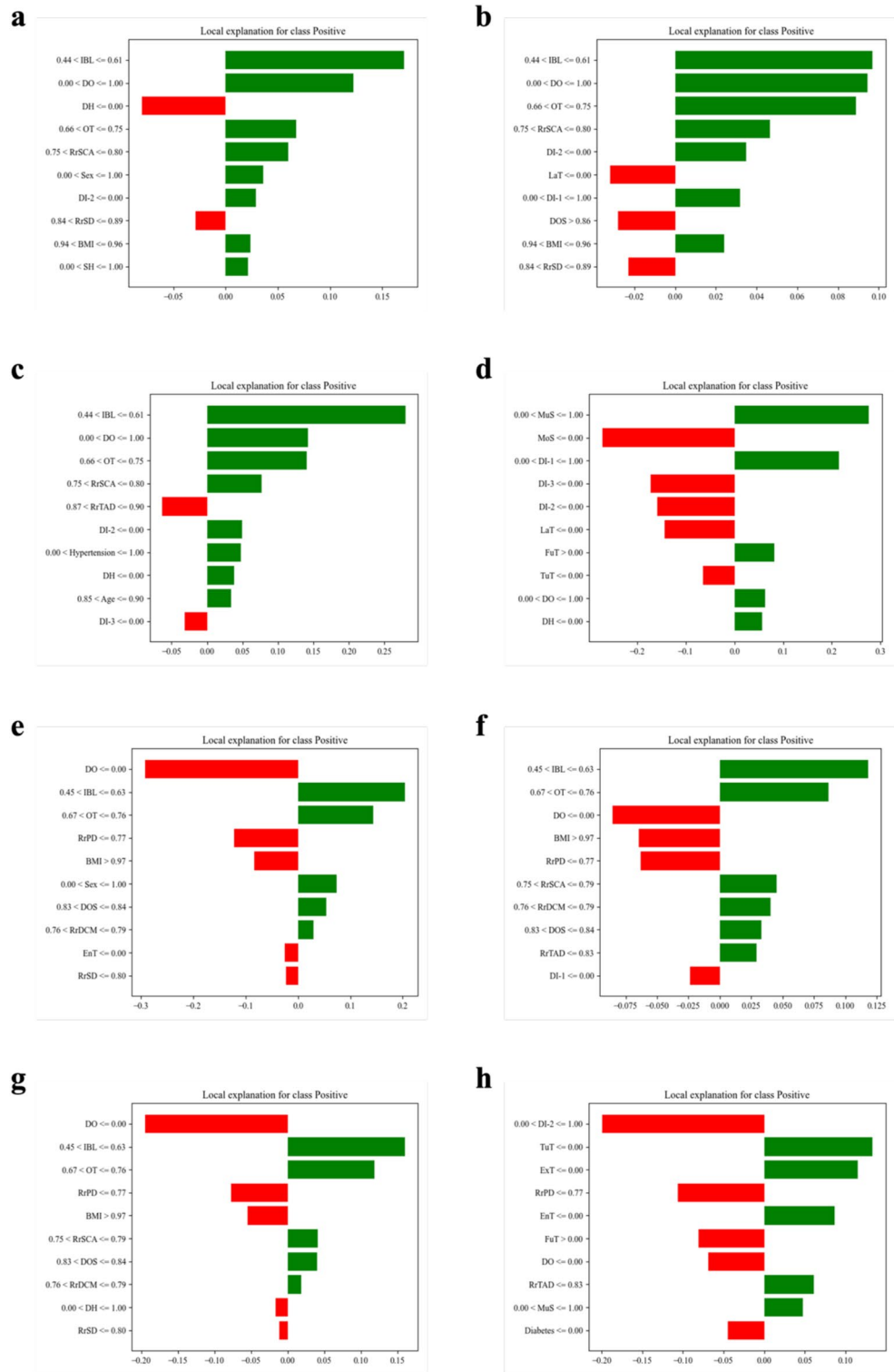


Fig. 6. Subfigures (a), (b), (c) and (d) in this figure represent the LIME value in the XGBoost, RF, LightGBM and Support Vector Machine under ADASYN, respectively. Subfigures (e), (f), (g) and (h) in this figure represent the LIME value in the XGBoost, RF, LightGBM and Support Vector Machine under SMOTE, respectively. The higher the bar, the more important the feature is and the greater its contribution to the model's predictive ability. The further the bar extends to the left, the greater the impact of the feature on the model's prediction of 0, and vice versa, the greater the impact of the feature on the model's prediction of 1.

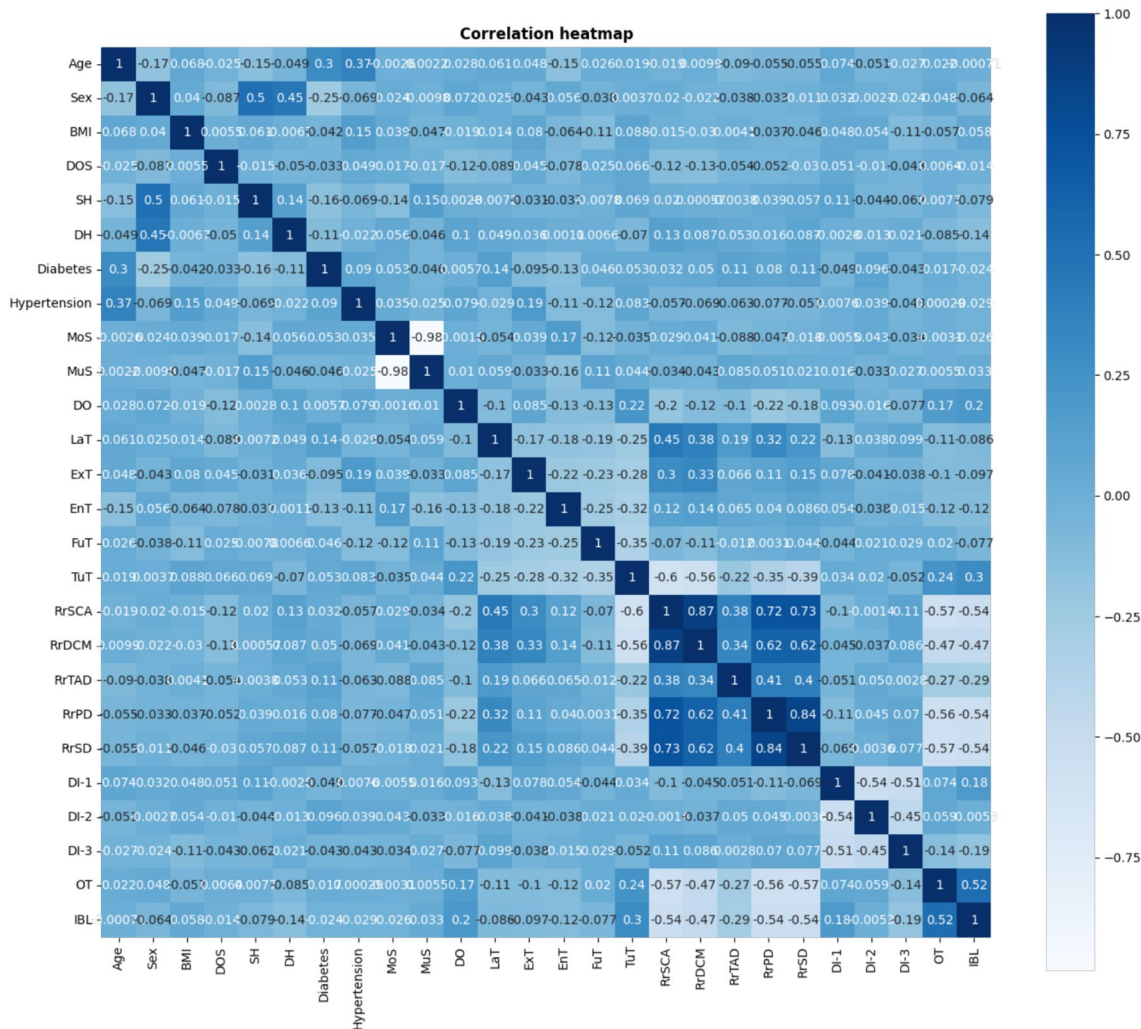


Fig. 7. The correlation between the different characterization factors. Darker colors indicate a stronger positive correlation between two feature factors and lighter colors indicate a stronger negative correlation between two feature factors.

lesions take the lead with the most significant global weight, followed by BMI, operation time, and age, which is verified in the dual coefficient of SVM, SHAP summary, and LIME local interpretation. LIME also named DI-1 (piezosurgery) for its significant positive impact on single-point prediction. In contrast, under SMOTE resampling, SVM's primary focus shifted to radiographic stenosis indicators: the dual coefficient plot ranked the RrPD, RrDCM, and RrSCA in the top three. The results of the SHAP value further highlighted the importance of DI-2 (high-speed drill) and intraoperative blood loss. The LIME value highlights the role of the Tuberos type (TuT) and Extended type (ExT) in promoting the optimistic prediction of this sample. Overall, ADASYN strengthens the discrimination of lesion segmentation and patient characteristics, while SMOTE makes imaging stenosis and operation details dominant in the decision boundary.

Discussion Analysis of importance features

The treatment of symptomatic ossification of the thoracic ligamentum flavum with posterior laminectomy is very complicated and accompanied by many surgical complications such as dural tear and CSFL, which significantly increases the difficulty of surgery and the risk and greatly affect the patient's prognosis and patient satisfaction with surgery. Machine learning models leveraged diverse data modalities—demographics, radiographic parameters, and surgical metrics—to predict CSFL. To capture both global and local drivers of risk, we applied three complementary interpretability techniques across two resampling strategies (SMOTE and ADASYN): Model coefficients (linear weights for LR and SVM; gain or impurity reduction for tree ensembles), SHAP values, quantifying each feature's average contribution to model output, and LIME explanations, revealing how small perturbations in individual cases shift predicted CSFL probability.

Across all methods and sampling schemes, four features emerged as the most robust predictors:

Multi-segment involvement: Consistently highest in coefficient and SHAP rankings, reflecting that ossification spanning ≥ 2 levels demand larger bony resections and more extensive dura manipulation, mechanically elevating tear risk. **Intraoperative blood loss (IBL):** Large positive coefficients and SHAP contributions indicate that heavy bleeding degrades visualization and increases tissue stress, corroborated by LIME's local weights showing that even moderate increases in IBL substantially raise CSFL probability. **Operation time:** Prolonged surgeries similarly degrade operative conditions; OT ranks among the top three features in both global and local explanations. Although surgical time and have been explicitly studied in relation to complications after spinal decompression [8]. **Spinal canal encroachment ratios (especially RrSCA, RrPD, RrDCM):** Cases with residual canal area below 50% or paramedian residual diameter under $\sim 45\%$ consistently pushed model outputs toward higher CSFL risk, as evidenced by pronounced SHAP shifts in both SMOTE and ADASYN models.

Beyond these primary drivers, several secondary factors showed moderate but nontrivial importance:

Duration of symptoms and diabetes history: Although prior studies did not flag these as CSFL predictors, our models—especially under ADASYN—assigned positive weights to longer duration of symptoms and diabetic status, suggesting that prolonged cord compression and microvascular changes may subtly increase dural fragility [28]. According to a multivariate regression analysis by Ahmet Kinaci's team [29], younger age, male, higher body mass index, smoking history were associated with increased incisional CSFL risk. However, duration of symptoms and diabetes have not been shown to be important factors predictive of postoperative CSFL. One possible explanation is that these variables are actually meaningful, and that the ML algorithm captured a complex, nonlinear association between these features that was not detected by previous works [30]. In addition, a longer duration of symptoms often means a longer period of spinal cord compression, and the progression of ossification often makes the risk of CSFL greater [31]. **Dural ossification:** Ranked lower in aggregate importance but featured among the top three predictors in XGBoost and LightGBM under SMOTE, highlighting model-specific sensitivity to combined ossification patterns. Notably, some studies [32, 33] have shown DO to have an impact on postoperative outcomes, especially when DO and ligament ossification are present at the same time, which significantly increases the risk of dural tear and CSFL, greatly affecting the patient's prognosis. **Decompression instrument (DI-1, DI-2, DI-3):** Linear models and tree gains ranked traditional bone chisels and high-speed drill above piezosurgery, a novel finding that warrants further clinical investigation. The traditional bone chisels are associated with high labor intensity and prolonged decompression time, which may exacerbate mechanical compression-induced injury to the spinal cord within the canal [34]. In contrast, the high-speed drill significantly enhances operative efficiency while reducing operator workload. However, its use requires advanced technical proficiency, as improper handling may lead to direct trauma to the spinal cord and dura mater [35]. Additionally, the high-speed drill carries risks of thermal injury to neural structures [36] and potential damage to adjacent soft tissues. In recent years, piezosurgery has gained widespread adoption in spinal surgical procedures. This innovative technique utilizes high-frequency micro-vibrations of the cutting blade to achieve precise and safe osteotomy. The technology offers several theoretical advantages, including exceptional cutting precision, superior tissue selectivity, minimal neural tissue trauma, reduced operative duration, and decreased intraoperative blood loss.

These insights suggest tailored strategies at each perioperative stage:

Preoperative planning: Patients with MuS, deep stenosis (e.g. RrSCA $< 50\%$), long symptom duration or diabetes should receive more detailed preoperative risk stratification. Furthermore, those patients may benefit from piezosurgery decompression to minimize risk of dural tear. **Intraoperative monitoring:** Real-time tracking of cumulative blood loss and elapsed time against model-derived thresholds can trigger dural safeguard protocols—such as staged decompression or early dural inspection—when key metrics cross critical cutoffs. **Postoperative risk stratification:** Incorporating calibrated probability estimates (via isotonic regression under SMOTE) will refine individualized CSFL risk scores, supporting shared decision-making and targeted follow-up.

By uniting traditional coefficient analysis with SHAP's game-theoretic attributions and LIME's local fidelity, this integrated assessment delivers a nuanced, actionable feature hierarchy that aligns mechanistic understanding with data-driven risk prediction.

Analysis of the causes of model performance

In this study, we analyze theoretical and practical factors shaping model behaviour to understand the differing performances observed across models and sampling strategies. The inherent class imbalance in the dataset—only 31.8% of cases exhibited CSFL—necessitated the use of oversampling strategies to ensure adequate representation of the minority class during training. We applied two well-established techniques, SMOTE and ADASYN, and evaluated their impact across five classifiers.

Although oversampling nominally balances class distributions, the clinical context of CSFL prediction dictates that evaluation metrics must go beyond simple accuracy or AUC. In real-world practice, the cost of a false negative (i.e., missing a true CSFL case) is far higher than a false positive. Missed leaks can lead to meningitis, wound healing failure, or reoperation, while false positives may only prompt additional intraoperative inspection. Therefore, recall (sensitivity) and the F1 score, which balances recall with precision, are the most meaningful indicators of model value in this high-risk setting.

The results showed that SVM under SMOTE yielded the best balance of sensitivity and precision (F1 = 0.8889, recall = 0.881), maintaining strong generalization with minimal drop under ADASYN. SVM leverages its margin-maximization principle:

$$\min_{\omega, b} \frac{1}{2} \omega^2 + C \sum_{i=1}^N \xi_i, \text{ subject to } y_i (\omega^T \phi(x_i) + b) \geq 1 - \xi_i,$$

where ω represents the weight vector, b is the bias term, ξ_i is the slack variable for handling misclassified samples, and C is a regularization parameter, enables it to separate classes while minimizing misclassification errors effectively [20]. In other words, this robustness of SVM can be attributed to its margin maximization principle, which inherently resists overfitting to noise and focuses on a sparse set of support vectors. When synthetic samples generated by SMOTE fill in the minority manifold uniformly, the SVM benefits from a clearer margin boundary, enhancing recall without sacrificing precision. In contrast, ADASYN concentrates synthetic data generation on minority instances near class boundaries. This approach can help when minority samples lie in sparse regions but may also inject instability into models like SVM, resulting in slightly degraded F1.

LR, by comparison, consistently underperformed across both sampling methods. Its linear decision boundary cannot model the nonlinear interactions among demographic, radiological, and surgical features. The issue is exacerbated with ADASYN, where synthetic examples in complex regions of feature space may lie outside LR's representational ability, leading to underfitting and poor recall.

Among ensemble models, RF achieved the highest AUC under SMOTE (0.9462), reflecting its strength in capturing complex variable interactions (Fig. 2). However, RF's reliance on bootstrapped decision trees makes it more susceptible to overfitting in small-sample, imbalanced settings. When synthetic samples are concentrated via ADASYN near noisy class boundaries, the model may overly adjust to spurious splits, thereby reducing precision and F1.

Similarly, XGBoost, a gradient-boosting framework, closely followed RF, optimizing:

$$\mathcal{L}(\theta) = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k),$$

Where $l(y_i, \hat{y}_i)$ is the loss function measuring classification error, and $\Omega(f_k)$ is a regularization term controlling model complexity [17]. XGBoost incrementally fits new trees to correct prior errors, performed strongly under SMOTE but declined under ADASYN (SMOTE F1=0.8537; ADASYN F1=0.8312). Its greedy optimization process can overcompensate in the presence of borderline synthetic points, leading to miscalibrated probability estimates. LightGBM, while efficient in large-scale data, was modestly outperformed by RF and SVM. Its leaf-wise splitting strategy may fail to generalize in small datasets, especially when positive-class signals are diluted by oversampling artefacts.

Crucially, these observations align with the underlying mathematical assumptions of SMOTE and ADASYN. SMOTE generates synthetic points by linear interpolation between minority-class neighbours, effectively regularizing the feature space and smoothing class boundaries. ADASYN, in contrast, assigns more synthetic points to “harder-to-learn” instances—those surrounded by majority-class neighbours. While this adaptive mechanism is beneficial in truly sparse minority regions, it may also amplify class overlap and introduce labelling ambiguity, which is particularly detrimental to variance-sensitive models like RF and XGBoost.

From a deployment perspective, these findings suggest that SVM combined with SMOTE sampling and isotonic calibration may be the most appropriate choice when minimizing missed CSFL cases is paramount. This combination demonstrated high recall, robust generalization, and well-calibrated probabilities for scenarios prioritizing precision—such as when overtreatment carries a high clinical or economic cost—RF under SMOTE may be favoured. While ADASYN is less effective in this dataset, it may remain valuable when minority subclasses are poorly represented or exhibit atypical patterns.

In conclusion, a model's effectiveness in imbalanced medical prediction tasks depends not only on performance metrics but also on the interplay between data distribution, sampling strategy, and model structure. Our results emphasize the importance of aligning these factors to the clinical decision context and balancing sensitivity, interpretability, and reliability in high-stakes prediction.

Comparison with previous studies

Most previous studies predicting CSFL following thoracic decompression surgery have relied on traditional LR models, which offer limited flexibility in accounting for complex, non-linear relationships among patient characteristics, radiological findings, and intraoperative variables [7–9]. While applicable for exploratory analysis, such models may not adequately reflect the multifactorial nature of CSFL risk, particularly in patients with multi-segment disease or severe spinal canal stenosis.

In this study, we implemented a broader machine learning framework incorporating advanced sampling strategies to address the class imbalance, probability calibration to ensure clinical interpretability of risk scores, and model explainability tools to clarify the role of individual predictors. To our knowledge, this is the first investigation to systematically evaluate combinations of SMOTE and ADASYN with Platt scaling and isotonic regression, offering insight into how different modelling strategies affect predictive accuracy, reliability, and transparency—qualities critical to adoption in clinical settings.

Beyond technical performance, our findings have direct relevance to preoperative risk stratification. The best-performing model—SVM with SMOTE resampling and isotonic calibration—demonstrated high sensitivity and excellent discrimination, making it well-suited for identifying patients at elevated risk of CSFL. Importantly, the output of this model is not a binary classification but a calibrated probability score that can be readily integrated into clinical decision-making workflows. For example, patients identified as high-risk could be candidates for enhanced dural protection techniques, staged decompression, or intensified postoperative monitoring. Furthermore, the predictors identified as most influential—multi-segment involvement, spinal canal encroachment parameters, operative time, and blood loss—are measurable pre or intraoperatively and modifiable to some extent through surgical planning.

It is also worth noting that while SVM with SMOTE emerged as the optimal combination in our cohort, this may not be universally true across all institutional contexts. In clinical settings where rapid inference, system resource constraints, or interpretability requirements differ, other models such as random forest, XGBoost, or LightGBM may offer advantages. Our findings thus provide a flexible foundation for tailoring predictive solutions to the needs of different surgical teams, patient populations, or health system infrastructures.

Nevertheless, this work remains a single-centre, retrospective study, and the generalizability of our results should be confirmed through external, multi-institutional validation. Moreover, while we focused on structured clinical and imaging features, future studies should explore whether advanced imaging analysis, such as radiomic texture features or deep learning-based segmentation, can improve prediction accuracy and support surgical precision.

In conclusion, by integrating robust machine learning approaches with clinically relevant evaluation strategies, our study offers a practical and interpretable tool for improving perioperative risk assessment in patients undergoing thoracic decompression. This approach holds promise for improving surgical outcomes and enhancing personalized care through data-driven decision support.

Limitations

Several limitations of this study warrant consideration. First, although the results demonstrate promising predictive performance, the model was developed and validated using data from a single academic centre. As such, its generalizability to other institutions with different surgical techniques, patient demographics, or imaging protocols remains uncertain. External validation across multiple centres and prospective clinical studies will be essential before routine clinical implementation can be considered.

Second, while the study incorporated a range of structured variables—including demographics, imaging-derived measurements, and intraoperative details—certain predictors, such as residual canal diameters and blood loss, may not be readily or consistently available across all settings. In particular, the reliance on high-resolution preoperative CT scans and standardized intraoperative documentation could limit the model's scalability in resource-limited environments. Future iterations of this work may benefit from evaluating model performance using more universally accessible data inputs or automated imaging extraction techniques.

Third, although our machine learning framework outperformed traditional approaches, we did not include deep learning-based methods, such as convolutional neural networks or transformer-based architectures, which may be capable of capturing higher-order patterns in imaging or temporal surgical data. These methods may enhance predictive accuracy, particularly when combined with raw image inputs or intraoperative video data. However, their higher computational burden and limited interpretability pose challenges for near-term clinical use.

Fourth, the exclusion of 215 patients based on predefined criteria, while intended to ensure data consistency and quality, may have introduced selection bias. Differences in characteristics between included and excluded patients could lead to underestimation or overestimation of CSFL risk in certain subpopulations, thereby affecting the model's generalizability. Although we conducted internal comparisons of baseline characteristics and applied robust validation techniques such as cross-validation method to mitigate this risk, the potential for bias remains. Future studies should systematically evaluate the impact of exclusion criteria on predictive performance.

Fifth, as a retrospective study, the absence of prospective validation limits the applicability of our model in real-world clinical workflows. While current constraints prevent immediate prospective analysis, we have enhanced internal validation and made our model publicly available to encourage external validation by other institutions. Additionally, we are actively planning multi-center prospective studies to further assess and refine our predictive framework.

Sixth, the model was trained using oversampling techniques to address the class imbalance, but some misclassification of CSFL cases remained—particularly in less typical presentations. While we employed established techniques such as SMOTE and ADASYN, the recall of rare variants of CSFL remains suboptimal, and future work could explore more advanced strategies, such as cost-sensitive learning, federated augmentation, or generative adversarial networks to simulate rare events.

Finally, the current model represents a static snapshot based on historical data. In real-world clinical practice, predictive tools must evolve alongside changing surgical standards, imaging quality, and patient characteristics. Mechanisms for periodic model updating and recalibration will be essential to maintaining relevance and reliability over time. Furthermore, ethical considerations—such as transparency, explainability, and equity across different patient subgroups—should remain a central focus in future deployment efforts.

In light of these limitations, our future work will focus on multi-centre collaboration, including richer data modalities and developing clinically integrated tools that can be adapted to diverse healthcare environments while maintaining interpretability and safety.

Conclusion

Using retrospective data from 318 patients with thoracic ossification of the ligamentum flavum, we systematically compared logistic regression with four machine-learning algorithms for predicting postoperative cerebrospinal-fluid leakage (CSFL). After SMOTE oversampling and Bayesian hyper-parameter tuning, the Support Vector Machine (SVM) achieved the best overall performance on the test set (F1 = 0.889, recall = 0.881) and showed excellent calibration after isotonic regression (Brier = 0.103). Interpretability analyses consistently highlighted multi-segment involvement, residual spinal-canal area and diametric ratios, operative time, and intra-operative blood loss as the most influential predictors. We have released both the source code and an online demo, providing clinicians with an immediately available tool for individualized CSFL risk assessment.

The model can support pre-operative risk stratification, surgical-technique selection, peri-operative monitoring, and patient counselling, potentially reducing CSFL-related morbidity. Nevertheless, its single-centre, retrospective design, modest sample size, and possible selection bias limit generalisability; prospective, multicentre validation is required. Future work will focus on cross-institutional data sharing, automated imaging-feature extraction, and periodic model updating to facilitate clinical deployment and continual refinement.

Data availability

All data gathered can be requested from the corresponding author.

Received: 21 February 2025; Accepted: 16 June 2025

Published online: 03 July 2025

References

- Zhang, C., Chang, Y., Shuang, L. & Chen, Z. Pathogenesis of thoracic ossification of the ligamentum flavum. *Front Pharmacol* **15**, 1496297 (2024).
- Kägi, S., Ciureaand, A. & Micheroli, R. Ossification of the ligamentum flavum. *Rheumatology (Oxford)* **59**(7), 1616 (2020).
- Fan, T. et al. Clinical progression of ossification of the ligamentum flavum in thoracic spine: a 10- to 11-year follow-up study. *Eur Spine J* **32**(2), 495–504 (2023).
- Pan, Q. et al. Zoning laminectomy for the treatment of ossification of the thoracic ligamentum flavum. *Asian J Surg* **46**(2), 723–729 (2023).
- Zhao, Y. et al. Incidence and risk factors of dural ossification in patients with thoracic ossification of the ligamentum flavum. *J Neurosurg Spine* **38**(1), 131–138 (2023).
- Tang, J. et al. Risk factors and management strategies for cerebrospinal fluid leakage following lumbar posterior surgery. *BMC Surg* **22**(1), 30 (2022).
- Hu, P. P., Liu, X. G. & Yu, M. Cerebrospinal fluid leakage after thoracic decompression. *Chin Med J (Engl)* **129**(16), 1994–2000 (2016).
- Jiang, L. et al. Predictors of cerebrospinal fluid leak following dural repair in spinal intradural surgery. *Neurospine* **20**(3), 783–789 (2023).
- Jin, J. Y. et al. Risk factors for cerebrospinal fluid leakage after extradural spine surgery: A meta-analysis and systematic review. *World Neurosurg* **179**, e269–e280 (2023).
- Wijnberge, M. et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: The HYPE randomized clinical trial. *Jama* **323**(11), 1052–1060 (2020).
- Pruneski, J. A. et al. 3rd, Supervised machine learning and associated algorithms: Applications in orthopedic surgery. *Knee Surg Sports Traumatol Arthrosc* **31**(4), 1196–1202 (2023).
- Fatima, N., Zheng, H., Massaad, E., Hadzipsic, M. & G. M. Shankarand J. H. Shin., Development and validation of machine learning algorithms for predicting adverse events after surgery for lumbar degenerative spondylolisthesis. *World Neurosurg* **140**, 627–641 (2020).
- Karhade, A. V. et al. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. *Spine J* **21**(10), 1635–1642 (2021).
- Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* **13**, 1 (2015).
- Gao, A. et al. One-stage posterior surgery with intraoperative ultrasound assistance for thoracic myelopathy with simultaneous ossification of the posterior longitudinal ligament and ligamentum flavum at the same segment: A minimum 5-year follow-up study. *Spine J* **20**(9), 1430–1437 (2020).
- Zhai, J., Guo, S., Zhao, Y., Liand, C. & Niu, T. The role of cerebrospinal fluid cross-section area ratio in the prediction of dural ossification and clinical outcomes in patients with thoracic ossification of ligamentum flavum. *BMC Musculoskelet Disord* **22**(1), 701 (2021).
- Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
- Ke G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Yeand T.-Y. Liu. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. in *Neural Information Processing Systems*. 2017.
- Cortes, C. & Vapnik, V. N. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
- Chawla, N., Bowyer, K., Halland, L. & Kegelmeyer, W. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)* **16**, 321–357 (2002).
- Haibo H., Yang, B., Garcia, E. A., Shu tao, L. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008.
- Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**(3), 61–74 (2000).
- Zadrozny, B., & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, Association for Computing Machinery: Edmonton, Alberta, Canada. p. 694–699.
- Lundberg S. M. & Lee, S.-I. A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.
- Ribeiro M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 1135–1144.
- Scholkopf, B., Smola, A. J., & Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. 2018: The MIT Press.
- Horton, W. B. & Barrett, E. J. Microvascular dysfunction in diabetes mellitus and cardiometabolic disease. *Endocr Rev* **42**(1), 29–55 (2021).
- Kinaci, A. et al. Risk factors and management of incisional cerebrospinal fluid leakage after craniotomy: A retrospective international multicenter study. *Neurosurgery* **92**(6), 1177–1182 (2023).
- Luo, W. et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res* **18**(12), e323 (2016).
- Kalsi-Ryan, S., Clout, J., Rostami, P., Massicotteand, E. M. & Fehlings, M. G. Duration of symptoms in the quantification of upper limb disability and impairment for individuals with mild degenerative cervical myelopathy (DCM). *PLoS One* **14**(9), e0222134 (2019).

32. Zhao, Y. et al. Prevalence, diagnosis, and impact on clinical outcomes of dural ossification in the thoracic ossification of the ligamentum flavum: a systematic review. *Eur Spine J* **32**(4), 1245–1253 (2023).
33. Yu, L. et al. The relationship between dural ossification and spinal stenosis in thoracic ossification of the ligamentum flavum. *J Bone Joint Surg Am* **101**(7), 606–612 (2019).
34. Lin, J. D. et al. Quantitative and qualitative analyses of spinal canal encroachment during cervical laminectomy using the Kerrison rongeur versus High-Speed burr. *Br J Neurosurg* **33**(2), 131–134 (2019).
35. Huan, Y. et al. Application of piezosurgery osteotomy in cervical laminoplasty: Prospective, randomized, single-blind, clinical comparison study. *Clin Surg Res Commun* **4**, 32–38 (2020).
36. Takenaka, S., Hosono, N., Mukai, Y., Miwaand, T. & Fuji, T. The use of cooled saline during bone drilling to reduce the incidence of upper-limb palsy after cervical laminoplasty: clinical article. *J Neurosurg Spine* **19**(4), 420–427 (2013).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62306109), Health Research Project of Hunan Provincial Health Commission (No. W20243215), the special fund of the Hunan provincial key laboratory of pediatric orthopedics (No. 2023TP1019).

Author contributions

R.G.: Conceptualization, Data curation, Methodology, Statistics, Validation, Resources, Writing – original draft, writing – review, and editing. B.L.: Statistics. Y.W., Y.Z. and X.W.: Recruitment, data curation, Visualization, Investigation. D.J. and Z.L.: Supervision and Writing – Editing. All authors contributed to the article and approved the submitted version.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study protocol was approved by the Ethics Committee of Xiangya Hospital (No:202408121). We also followed the Declaration of Helsinki and its later amendments. The data is published with the written informed consent of all participants or their legal representatives.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-07430-7>.

Correspondence and requests for materials should be addressed to D.J. or Z.L.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2025