BioMedical Engineering OnLine

RESEARCH

Open Access

Development of a deep learning radiomics model combining lumbar CT, multi-sequence MRI, and clinical data to predict high-risk cage subsidence after lumbar fusion: a retrospective multicenter study

Zou Congying[4†], Chen Ruiyuan[4†], Wang Baodong[4], Fei Qi[2], Song Hongxing[3] and Zang Lei[1,4*]

†Congying Zou and Ruiyuan Chen have contributed equally to this work.

*Correspondence: zanglei@ccmu.edu.cn

[1] Department of Orthopedic Surgery, Beijing Chao-Yang Hospital, Capital Medical University, 8 Gong Ti Nan Lu, Chaoyang District, Beijing 100020, China
[2] Department of Orthopedics, Beijing Friendship Hospital, Capital Medical University, No 95, Yong'an Road, Xicheng District, Beijing 100050, China
[3] Department of Orthopedics, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, China
[4] Department of Orthopedics, Beijing Chaoyang Hospital, Capital Medical University, 5 JingYuan Road, Shijingshan District, Beijing 100043, China

Abstract

Background: To develop and validate a model that integrates clinical data, deep learning radiomics, and radiomic features to predict high-risk patients for cage subsidence (CS) after lumbar fusion.

Methods: This study analyzed preoperative CT and MRI data from 305 patients undergoing lumbar fusion surgery from three centers. Using a deep learning model based on 3D vision transformations, the data were divided the dataset into training ($n=214$), validation ($n=61$), and test ($n=30$) groups. Feature selection was performed using LASSO regression, followed by the development of a logistic regression model. The predictive ability of the model was assessed using various machine learning algorithms, and a combined clinical model was also established.

Results: Ultimately, 11 traditional radiomic features, 5 deep learning radiomic features, and 1 clinical feature were selected. The combined model demonstrated strong predictive performance, with area under the curve (AUC) values of 0.941, 0.832, and 0.935 for the training, validation, and test groups, respectively. Notably, our model outperformed predictions made by two experienced surgeons.

Conclusions: This study developed a robust predictive model that integrates clinical features and imaging data to identify high-risk patients for CS following lumbar fusion. This model has the potential to improve clinical decision-making and reduce the need for revision surgeries, easing the burden on healthcare systems.

Keywords: Posterior lumbar interbody fusion, Cage subsidence, Magnetic resonance imaging, Computed tomography, Deep learning radiomics, Predictive model

Introduction

Lumbar degenerative disease, a prevalent spinal disorder, manifests as chronic low back pain and progressive mobility restriction with heterogeneous clinical presentations. In advanced stages, progressive neurological deterioration and biomechanical instability

may profoundly compromise functional capacity, frequently resulting in substantial impairment of activities of daily living [1]. Lumbar fusion surgery has become a standard treatment option when conservative treatment fails. Interbody fusion cages play a crucial role in lumbar interbody fusion surgery for treating degenerative lumbar diseases. They achieve interbody fusion as a bone graft carrier, enabling unstable segments to regain stability and ultimately alleviating low back and leg pain [2]. However, cage subsidence (CS) reduces the intervertebral height, weakens the anterior column support, decreases the local lumbar lordosis, and deteriorates the posterior soft-tissue tension, thereby affecting the indirect decompression effect of the surgery [3–6]. Patients may require revision surgery or additional interventions when CS progresses to a symptomatic stage [7, 8]. This not only significantly increases the healthcare system burden but also greatly affects patients' quality of life and work capacity. Therefore, accurately determining patients at high risk of CS preoperatively is crucial for reducing surgical risks, optimizing treatment strategies, and enhancing prognosis.

Computed tomography (CT) and magnetic resonance imaging (MRI) are both noninvasive and high-resolution imaging techniques. MRI is popular for its superior tissue contrast and multiplanar imaging capabilities, whereas CT excels in imaging bone structures. Thus, they are both considered important examinations for assessing lumbar degenerative disease. However, existing predictive methods for CS based on medical imaging are often limited by their reliance on subjective clinical judgment, leading to significant variability in the results. While traditional approaches such as visual assessment by clinicians and simple statistical models have been used, they lack consistency and fail to adequately capture the complex relationships between imaging features and the risk of CS. This variability highlights the need for more reliable, objective, and data-driven approaches. Our study aims to address these gaps by integrating radiomics and clinical data, providing a more accurate and robust model for predicting high-risk CS patients.

Radiomics, which extracts quantitative features from medical images, is widely used to assess microstructural changes in the lumbar spine and surrounding tissues. By analyzing image patterns and textures, it offers valuable insights into the disease process [9, 10]. Meanwhile, deep learning has been predominantly used in spinal imaging analysis because of the advantages of three-dimensional convolutional neural networks (3D CNNs) in processing 3D data. Several studies have combined radiomics and machine learning techniques to improve medical efficiency in spinal diseases [11–14]. In this study, we hypothesize that a model integrating radiomic features, deep learning radiomics features, and clinical data will provide more accurate predictions for high-risk CS after lumbar fusion surgery compared to the traditional methods. Our objective is to develop predictive models using clinical data, lumbar CT, and MRI from local medical centers, assess their performance in predicting CS. By constructing a combined model, we aim to identify patients at high risk of CS and improve clinical decision-making.

Results

Clinical baseline characteristics

This study included 305 patients following the inclusion and exclusion criteria. The patients were aged 36–90 years, consisting of 162 females and 143 males. According to

bone mineral density examination results, 143 patients were diagnosed with osteoporosis. Among the included patients, 201, 90, and 14 underwent single-, two-, and three-segment fusion surgeries, respectively. During the follow-up period, 75 patients were diagnosed with CS based on the imaging examination results. Table 1 summarizes the patients' demographic and clinical characteristics. The p values show no significant age difference between the test and training groups, but significant differences exist between the test and validation groups and between the validation and training groups. Binary logistic regression analysis revealed that osteoporosis was significantly associated with CS occurrence (Additional file Table S1).

Radiomics feature selection and construction of the prediction model (radiology-based)

We used the LASSO regression model for feature selection and dimensionality reduction after feature fusion, and the penalty coefficient was $\lambda = 0.018$ (Fig. 1a, b). Finally, 11 radiomics features were retained after feature fusion (Fig. 1c).

We constructed a model based on the fused features and their corresponding regression coefficients using preoperative clinical imaging hand-crafted radiomics features to predict CS after lumbar fusion surgery. Among all the machine learning algorithms tested, AdaBoost proved to be the most effective for radiomics models. The AUC results for the training cohort, validation cohort, and test cohort were 0.872 (95%CI, 0.826–0.923), 0.788 (95%CI, 0.673–0.927), and 0.851 (95%CI, 0.701–0.972), respectively (Fig. 1d–f).

Table 1 Clinical baseline characteristic of patients in the training cohort, validation cohort, and test cohort

Characteristic	Training cohort, ($n=214$)	Validation cohort, ($n=61$)	Test cohort, ($n=30$)	*p*
Sex, no (%)				0.993
Female	114 (53.27)	29 (47.54)	14 (46.67)	
Male	100 (46.73)	32 (52.46)	16 (53.33)	
Age (years)				0.004*
Mean (range)	64.39 ± 9.11	60.00 ± 9.98	64.90 ± 8.13	
Follow-up				0.816
Mean (range)	9.82 ± 3.19	9.87 ± 3.00	9.17 ± 2.20	
BMI				0.687
Mean (range)	26.65 ± 3.86	26.46 ± 4.00	27.21 ± 4.27	
Diagnosis, No (%)				< 0.001*
Lumbar spinal stenosis	113 (52.80)	15 (24.59)	9 (30.00)	
Spondylolisthesis	58 (27.10)	16 (26.23)	11 (36.67)	
Lumbar disc herniation	43 (20.09)	30 (49.18)	10 (33.33)	
Segment				0.035*
1	131 (61.21)	46 (75.41)	24 (80.00)	
2	69 (32.24)	15 (24.59)	6 (20.00)	
3	14 (6.54)			
Osteoporosis, no (%)				0.363
Yes	108 (50.47)	36 (59.02)	12 (40.00)	
No	106 (49.53)	25 (40.98)	18 (60.00)	

*Indicates statistical significance at $p < 0.05$

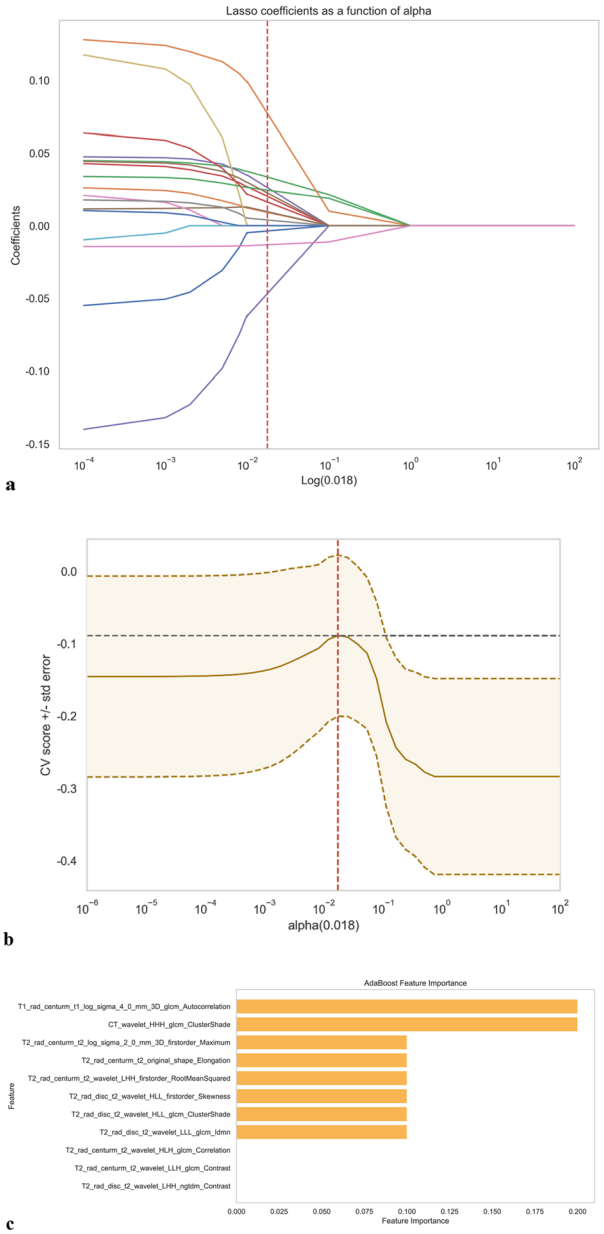


Fig. 1 LASSO regression-based selection of radiomics features. The optimal λ value of 0.018 was selected. Performance of the machine learning model based on the AdaBoost algorithm. **a** Feature coefficients corresponding to the value of parameter λ. Each line represents the change trajectory of each independent variable. **b** The most valuable features were screened out by tuning λ using LASSO regression. The dotted vertical line represents the optimal log(λ) value determined through cross-validation. **c** Feature importance ranking based on the LASSO-selected radiomic features using AdaBoost. The y-axis indicates the selected radiomic features, and the x-axis represents their relative importance. **d** ROC curve. **e** Calibration curve. **f** Decision curve analysis

Feature selection and construction of the prediction model (based on DL)

We used the LASSO regression model for feature selection and dimensionality reduction after combining the features, and the penalty coefficient was λ = 0.012 (Fig. 2a, b). Five DLR features were retained after feature selection (Fig. 2c).

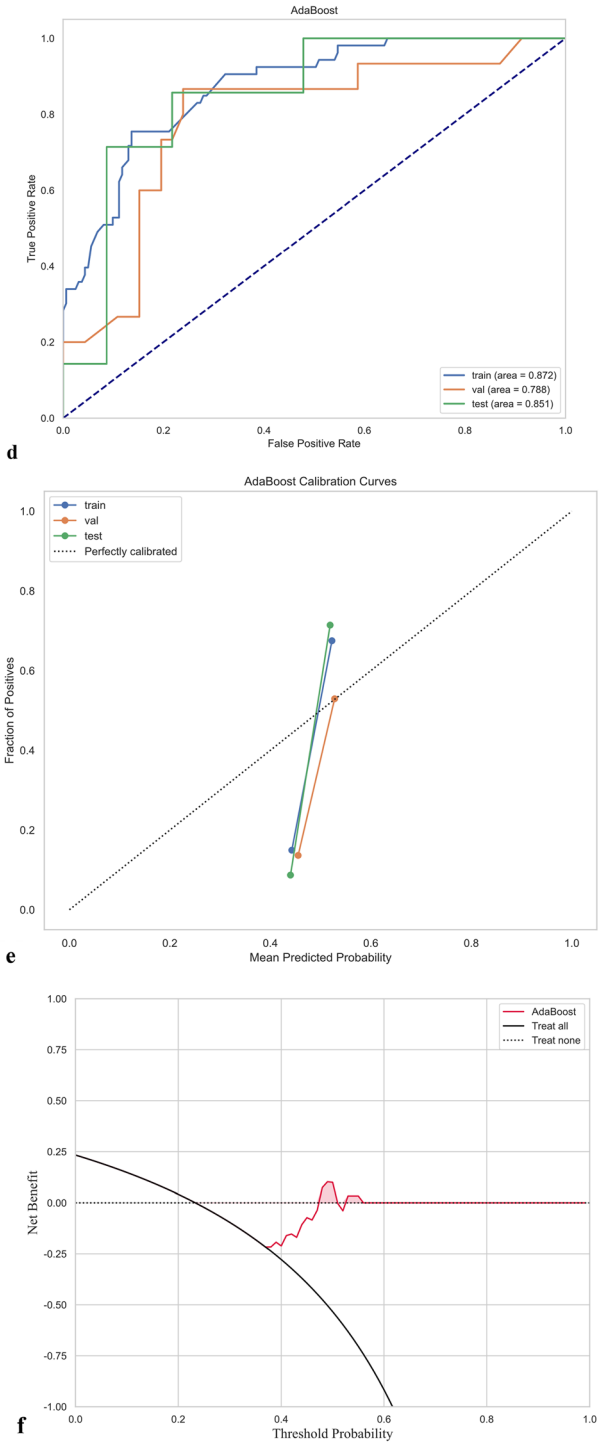


Fig. 1 continued

We constructed a model based on the fused features and their corresponding regression coefficients using preoperative clinical imaging deep learning radiomics features to predict CS after lumbar fusion surgery. AdaBoost continues to demonstrate its strong capability in optimizing model performance (Fig. 2d–f). The AUCs

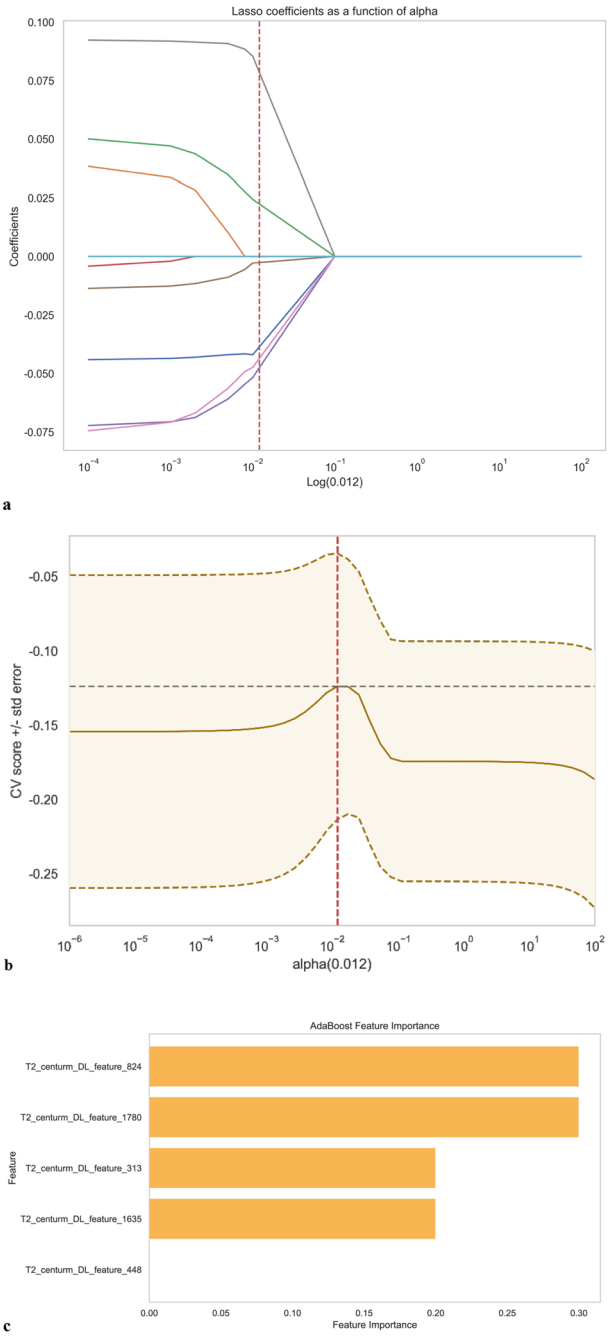


Fig. 2 LASSO regression-based selection of deep learning radiomics features. The optimal λ value of 0.012 was selected. And performance of the machine learning model based on the AdaBoost algorithm. **a** Feature coefficients corresponding to the value of parameter λ. Each line represents the change trajectory of each independent variable. **b** The most valuable features were screened out by tuning λ using LASSO regression. The dotted vertical line represents the optimal log(λ) value determined through cross-validation. **c** Feature importance ranking based on the LASSO-selected radiomic features using AdaBoost. The y-axis indicates the selected deep learning radiomics features, and the x-axis represents their relative importance. **d** ROC curve. **e** Calibration curve. **f** Decision curve analysis

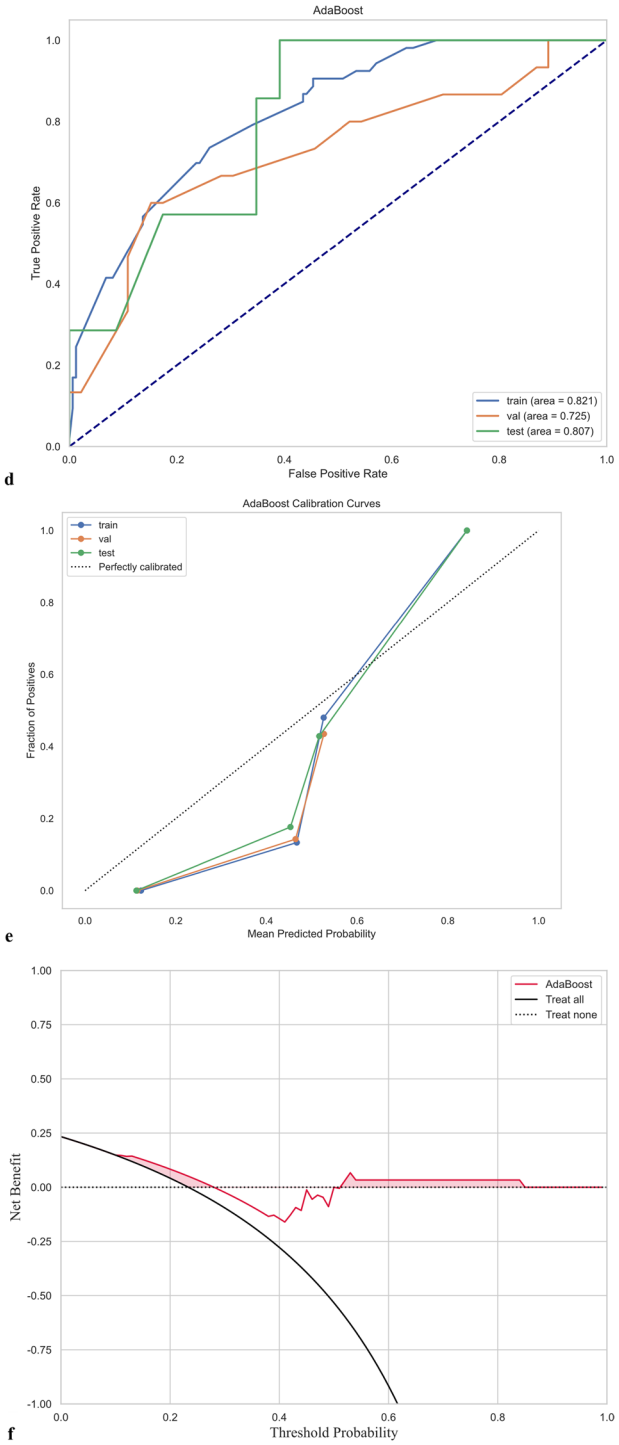


Fig. 2 continued

for the training cohort, validation cohort, and test cohort were 0.821 (95%CI, 0.774–0.874), 0.725 (95%CI, 0.540–0.871), and 0.807 (95%CI, 0.644–0.937), respectively.

Combined model based on baseline clinical data, Rad-sign, and DL-sign

In the combined model, which integrates both Rad-sign and DL-sign alongside clinical data, AdaBoost's ability to enhance model accuracy becomes even more pronounced. The AUCs for the training cohort, validation cohort, and test cohort were 0.941 (95%CI, 0.911–0.969), 0.832 (95%CI, 0.702–0.941), and 0.935 (95%CI, 0.810–0.99), respectively (Fig. 3a).

The feature importance bar chart indicates that the DL-sign, Rad-sign, and osteoporosis contribute the most to the prediction results when combining radiomics and deep learning prediction models with the clinical baseline features to construct a combined model (Fig. 3b). Compared with the previous two models based solely on radiomics features or DLR features, the combined model demonstrated a significant improved predictive performance of the validation and test cohorts. The DCA indicated that the combined model provided higher net benefits to patients compared with the single-feature models (Fig. 3c, d). The DeLong test results reveal that the combined model demonstrates a statistically significant advantage over the single-feature model ($p < 0.05$). The combined model outperformed the predictions made by two experienced spinal surgeons in terms of predictive accuracy. The comparison between the machine and clinicians revealed statistically significant differences (Table 2). This finding highlights the potential of machine learning-based models in providing more accurate and objective predictions, especially when considering the complexity and variability of spinal conditions. Clinicians may benefit from this model as a supportive tool for identifying high-risk patients for CS, potentially reducing the need for revision surgeries and improving patient outcomes. Therefore, this combined model is more intuitive and effective in identifying high-risk patients for CS.

Discussion

Imaging examinations during postoperative follow-up are predominantly used as an important method for diagnosing CS in clinical practice [15]. However, identifying high-risk patients during preoperative evaluation is crucial for improving patient prognosis, considering that CS occurrence increases the likelihood of patients returning for additional treatment. In recent years, several studies have emphasized various risk factors for predicting the possibility of CS after spinal fusion surgery [16, 17]. Factors, such as age, higher BMI, and poor vertebral bone quality, have been determined as important CS predictors [7, 18]. Despite these advances, the current literature has largely overlooked the use of preoperative imaging algorithms for predicting CS, with reported models achieving an AUC of only 0.6–0.7, which is considerably lower than the performance of the model presented in this study. In this study, we used datasets from three hospitals to develop a combined predictive model that integrates deep learning radiomics and traditional radiomics based on CT and multi-sequence MRI, along with baseline clinical data, to determine patients at high risk of developing CS. This combined approach enabled the identification of patients at high risk for developing CS, offering significant potential for improving preoperative risk stratification and clinical decision-making. The superior performance of our model highlights its

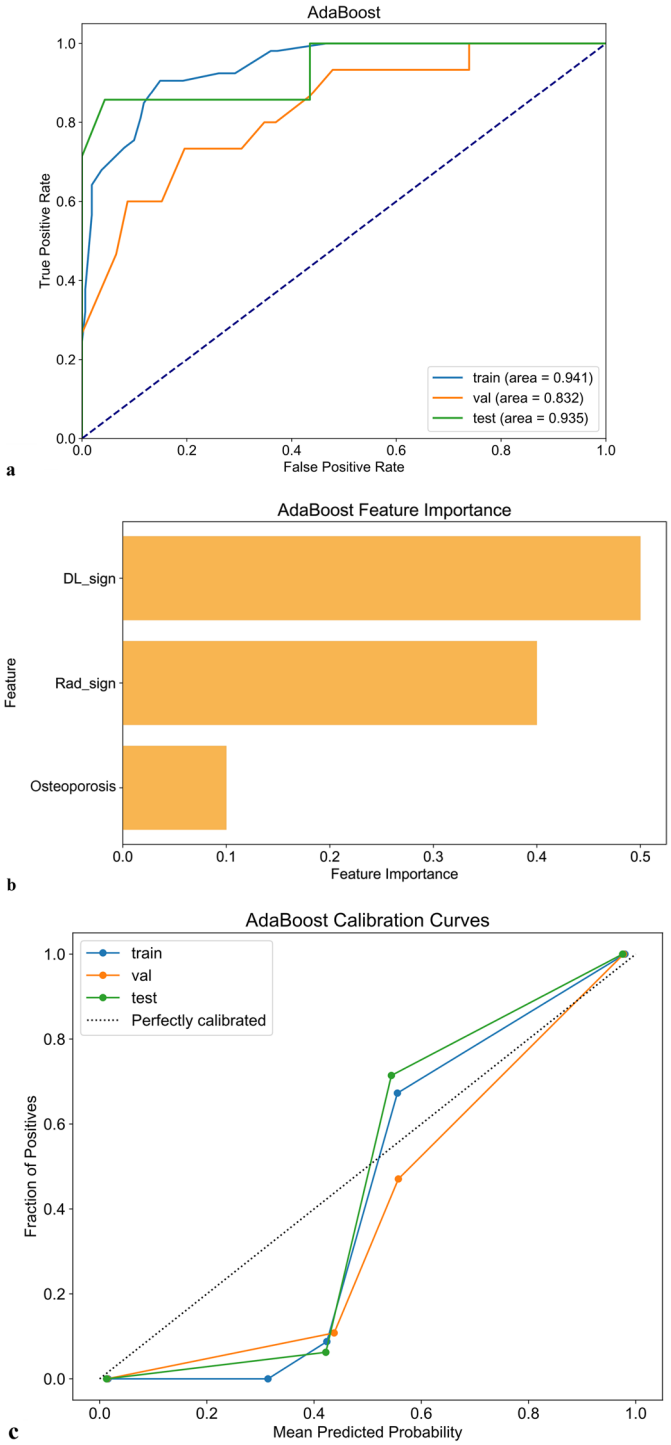
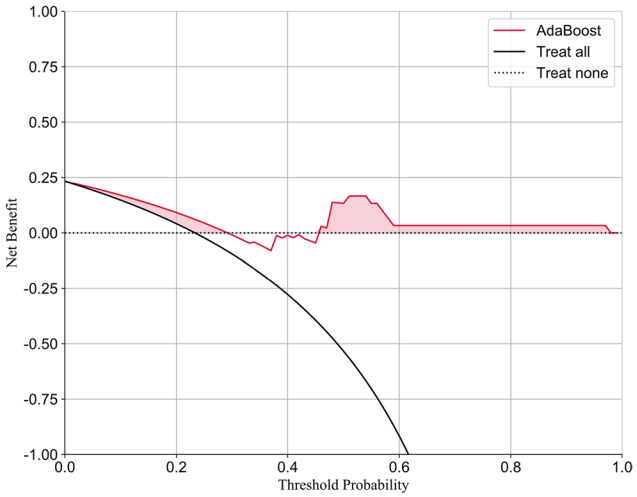


Fig. 3 Performance of the machine learning model based on the AdaBoost algorithm. **a** ROC curve. **b** Feature importance ranking based on the LASSO-selected features using the AdaBoost. **c** Calibration curve. **d** Decision curve analysis



d

Fig. 3 continued

Table 2 Comparison of performance metrics

Dataset	Prediction Method	Threshold	ACC	AUC	Sensitivity	Specificity	NPV	PPV	F1
val	Combined Model	0.52	0.79	0.83	0.73	0.80	0.90	0.55	0.63
Test	Combined Model	0.52	0.93	0.93	0.86	0.96	0.96	0.86	0.86
val	Surgeon	–	0.62	0.76	0.27	0.74	0.76	0.25	0.26
test	Surgeon	–	0.53	0.53	0.43	0.57	0.76	0.23	0.30

clinical relevance in providing more accurate predictions, ultimately contributing to better patient outcomes and more targeted interventions.

CT and MRI are important diagnostic tools for diagnosing lumbar degenerative diseases and are capable of detecting microenvironmental changes in the spine and surrounding tissues. Based on a previous study [19], lumbar CT images of the vertebrae, sagittal MRI of the vertebrae and intervertebral discs, and axial MRI of the muscle tissue collectively provide comprehensive information related to the spine and its microenvironment. Therefore, in this study, we segmented and analyzed the vertebral bodies on CT images and the vertebral bodies and intervertebral discs on multi-sequence sagittal MRI (T1 and T2 sequences), as well as the quadratus lumborum, psoas major, and paraspinal muscles on axial images. Radiomics based on machine learning has been predominantly applied in the medical field in recent years. Some researchers have utilized machine learning algorithms, such as image segmentation, feature extraction and selection, and predictive modeling, to allow automated medical imaging analysis and offer intelligent diagnostic support to clinicians, thereby improving diagnostic accuracy and efficiency [20]. Some orthopedic scholars have used deep learning to process medical images, thereby developing an effective screening method for ASD based on deep learning and cervical MRI [21]. However, due to the poor interpretability of the deep learning models, this method has certain limitations in its clinical application. DLR has been proposed and has rapidly developed with the emergence of network architectures like ResNet to address these limitations [22, 23]. RadImageNet is an open-source medical

imaging dataset that is specifically designed for medical applications and theoretically outperforms ImageNet in medical tasks [24]. Successful transfer learning depends on the feature similarity between the source and target tasks [25]. Therefore, we conducted deep learning based on DTL using RadImageNet as the pretrained model. In this study, lumbar CT and MRI from local medical centers were conducted to construct DLR predictive models, and we compared the performance of these models in predicting CS after lumbar fusion surgery.

In this study, the optimal model with feature fusion retained 11 radiomics features, 5 DLR features, and 1 clinical feature after final selection. First, the autocorrelation feature in the gray-level co-occurrence matrix (glcm) after applying log transformation with a sigma of 4.0 mm in the lumbar vertebrae MRI (centrum_t1_log_sigma_4_0_mm_3D_glcm_Autocorrelation) revealed the highest correlation coefficient with the occurrence of CS. The T1_log_sigma_4.0mm_3D filter application aims to highlight texture variations at a specific scale ($\sigma = 4.0$ mm) by enhancing the texture features at that scale [26]. CS occurrence is closely related to vertebral bone strength and density changes; hence, filtering at this scale may effectively capture the detailed features associated with bone quality. The gray-level co-occurrence matrix (GLCM) captures the spatial gray-level associations between pixels, whereas the autocorrelation feature describes the similarity between adjacent pixel values in the image [27]. Autocorrelation reflects the uniformity and consistency of the bone structures for bone imaging. These texture features are crucial in evaluating vertebral health, because a uniform bone structure is typically associated with higher bone density and better mechanical properties, which are key factors in preventing CS. This finding may confirm that osteoporosis increases the risk of CS to a certain extent. Overall, the feature importance analysis reveals that the texture, morphology, and 3D structural features of the fused regions and their adjacent segments in preoperative clinical imaging play crucial roles in the model. These features collectively indicate that the complex mechanical and biological states of the spine before fusion exerted multidimensional effects. Moreover, we revealed that the DLR features with the highest correlation coefficients all originated from the vertebral bodies on the T2-weighted sequence. This is because the T2 sequence improves the ability to capture water content and soft-tissue contrast while providing detailed information about bone composition and endplate integrity. Both factors are crucial for assessing the risk of CS. The interpretability of the current DLR features requires further investigation, but this does not prevent us from obtaining preliminary explanations of the lesion-specific features by integrating them with the traditional radiomics features. Additionally, notably, the majority of radiomics features selected for the model and all the DLR features were derived from MRI. This indicates that MRI is more important than CT imaging in the preoperative imaging-based method proposed in this study for predicting patients at high risk of CS. Finally, our results indicate that osteoporosis significantly increases the risk of CS, which is consistent with the previous studies [28–31].

AdaBoost demonstrated superior performance across three modeling frameworks in our study, attributable to its algorithmic adaptability in addressing distinct challenges of radiomic analysis. In traditional radiomics, where high-dimensional features (e.g., texture metrics) frequently exhibit multicollinearity, AdaBoost iteratively optimizes weak classifiers through feature importance weighting, effectively balancing

bias–variance tradeoffs to preserve predictive accuracy [32]. For deep learning-based models, while CNN-derived features encode complex imaging patterns susceptible to overfitting, AdaBoost's sequential reinforcement of generalizable features counteracts data noise and sample size limitations inherent to medical imaging datasets [32]. Notably, the algorithm's discriminative capacity extended to multimodal integration, adaptively reconciling clinical, radiomic, and deep learning features through domain-specific weight optimization—a critical advantage in CS prediction where heterogeneous data synergy determines clinical relevance [33]. This tripartite efficacy—resolving feature redundancy in conventional radiomics, regularizing deep feature overfitting, and harmonizing multimodal inputs—positions AdaBoost as a computationally robust framework. Its consistent outperformance over single-feature models stems from systematic complexity control and dynamic prioritization of cross-domain biomarkers. The algorithm's decision consistency and stability in handling intricate feature interactions suggest translational potential for developing adaptive predictive systems in spinal disorder management. Future investigations should explore AdaBoost's scalability in multicenter cohorts and its integration with emerging neural architectures.

This study has some limitations. First, despite our efforts to enhance the accuracy of manual MRI image annotations, manual labeling is inherently subjective and susceptible to certain biases, which may lead to inconsistencies in results when compared to automated segmentation methods. Due to the current limitations in developing a highly accurate automatic soft-tissue segmentation algorithm for lumbar MRI, manual segmentation was used for this study. However, to minimize potential bias, we performed an interobserver consistency evaluation. While this approach helps reduce error, it does not entirely eliminate it. Therefore, future research should aim to optimize automatic MRI image segmentation techniques and combine the strengths of automation with human expertise to achieve more efficient and precise assessments of spinal structures and pathological features. Second, despite being based on multicenter data, the external sample size is limited due to inclusion criteria, and the small test cohort size and high feature-to-sample ratio may increase the risk of overfitting. Additionally, there is a lack of a rigorous assessment of inter-scanner variability among the CT and MRI datasets acquired from three different hospitals, which may have introduced potential biases. While standard preprocessing methods, including resampling and intensity normalization, were deployed to mitigate scanner-related differences, these approaches may not fully compensate for the variability introduced by distinct acquisition protocols and scanning hardware configurations. In the meantime, both the CT and MRI parameters and patient characteristics of the test cohort were comparable to those of the training cohort, which may partly explain the superior performance of the model on the test cohort compared to the validation cohort. Future research needs to include data from more centers and larger sample sizes for prospective multicenter validation. Moreover, radiomics often faces criticism for limited feature interpretability and the time-consuming nature of data annotation—challenges that our study also encounters. Nonetheless, as previously mentioned, similar to our work on automated segmentation of vertebral bodies in CT images, advancements in artificial intelligence and machine learning technologies offer the potential to enhance the application value of radiomics in both scientific research and clinical practice. By refining tools and methodologies, improving

automatic image segmentation techniques, and integrating multiple data sources, we can overcome these obstacles and achieve more efficient and precise assessments.

Conclusion

In conclusion, this study developed a composite model by integrating radiomics, DLR, and clinical features. Compared with models that are based on single-modality features, this combined model significantly improved the identification performance for patients at high risk of CS, offering substantial value in assisting clinical decision-making.

Methods

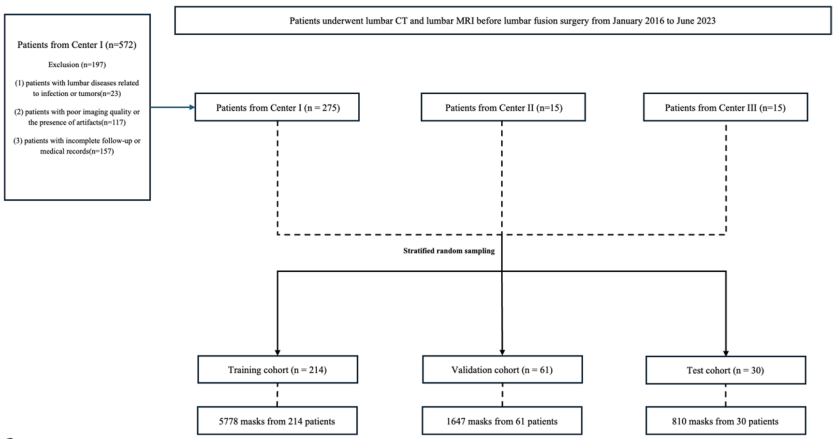
Patients

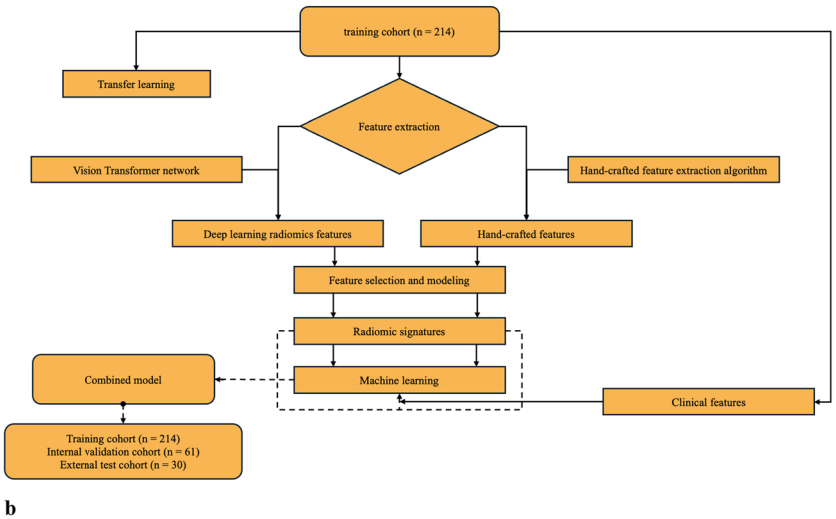
This study was designed and reported in accordance with the TRIPOD reporting guidelines [34], utilizing medical imaging data from three hospitals in China. We collected the clinical baseline and imaging data of patients with lumbar degenerative diseases who underwent posterior lumbar interbody fusion (PLIF) from January 2016 to June 2023 to establish training and validation cohorts as well as a testing cohort. Inclusion criteria were (1) patients who underwent 1–3 segment PLIF at our hospital and two other research centers, (2) those with at least 6 months of complete medical records and postoperative follow-up data, (3) patients with complete clinical data, and (4) patients with no history of spinal surgery. Exclusion criteria were (1) patients with lumbar diseases related to infection or tumors, (2) those with poor imaging quality or the presence of artifacts, and (3) patients with incomplete follow-up or medical records. Figure 4a illustrates a detailed flowchart of the case selection. The training, validation, and test cohorts were allocated from three centers' data using stratified random sampling to ensure an even distribution of positive and negative samples. Stratification was based on the occurrence of CS. The allocation ratio of the training, validation, and test cohorts was 7:2:1. The study design and the deep learning radiomics (DLR) workflow (Fig. 4b, c) demonstrate the processes of case collection and grouping, image preprocessing, feature extraction, feature analysis, and model construction. The Ethics Committee of the three hospitals involved in the study approved this study (Approval No.2024-KE-346) which complies with the Declaration of Helsinki. The Ethics Committee waived the need for informed consent because of the retrospective design study.

Clinical baseline characteristics and medical imaging acquisition

All patients' baseline data, including sex, age, number of surgical segments, body mass index (BMI), osteoporosis status, and diagnosis, were extracted from the clinical medical record system. Explanation Additional file S1 presents details of the CT and MRI equipment and imaging parameters.

Cage subsidence in our study was radiographically defined based on at least one of the following criteria: (1) a reduction in intervertebral height exceeding 10% compared to immediate postoperative imaging; or (2) newly observed, significant endplate damage associated with the fusion cage that was not evident on immediate postoperative imaging.





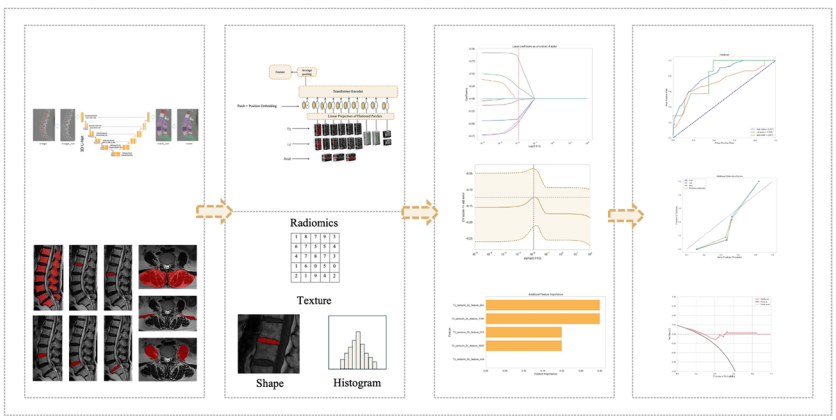


Fig. 4 **a** Flowchart summarizing patient selection and allocation to the training cohort, validation cohort, and test cohort. **b** Flowchart of this study, the process of feature extraction, model development, and validation using radiomics, deep learning, and clinical data. **c** Workflow of deep learning radiomics. The procedure involves four essential steps: (1) Annotation and segmentation: Identifying and outlining areas of interest (ROIs) on medical images. (2) Feature extraction: Deep learning features were extracted using a Vision Transformer-based model. (3) Feature selection and modeling: Applying statistical and machine learning methods approaches to select significant predictive features and establish the model. (4) Model assessment: Evaluating predictive performance and clinical significance with methods, such as calibration curves, ROC analysis, and decision curves

Image analysis and predictions

Two spinal surgeons (A and B) with 6 and 11 years of experience, respectively, independently predicted whether or not patients would develop CS postoperatively based on the same clinical information utilized by the algorithm—the patients' baseline data and clinical images. An expert with over 20 years of experience in spinal surgery (C) served as an adjudicator in cases where the two surgeons' predictions were inconsistent. Cohen's Kappa coefficient was calculated to evaluate the predictive agreement between A and B (Additional file Table S2).

Image segmentation

Accurate tissue segmentation is the premise for subsequent image analysis. The vertebral body segmentation was divided into two parts. A 3D U-Net model was used to segment the vertebral bodies for the CT images, after preprocessing. The CT segmentation

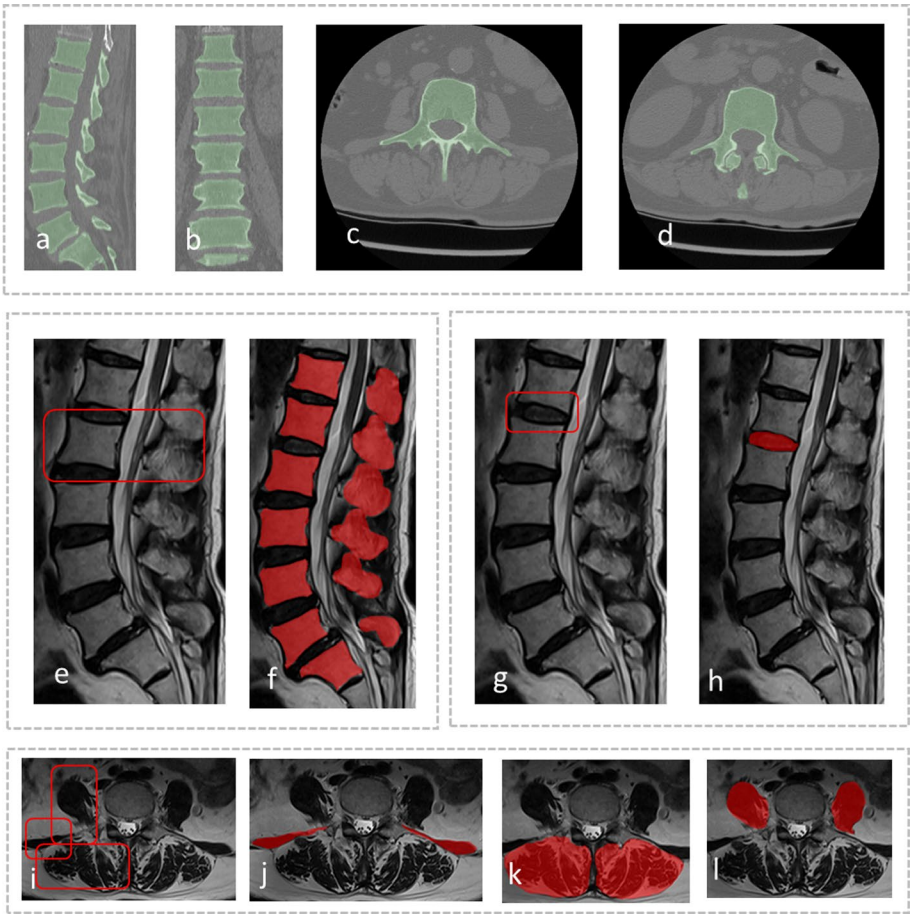


Fig. 5 Segmentation images based on lumbar CT (**a**–**d**) and multi-sequence MRI (**e**–**l**). **e**, **f** Vertebral body segmentation on sagittal MRI images. **g**, **h** Intervertebral disc segmentation on sagittal MRI images. **i**–**l** Muscle ROI on axial MRI images. ROI, region of interest. Separate segmentation of the intervertebral discs and vertebral bodies was carried out on both T1 and T2 MRI sequences to ensure that each structure was accurately delineated. Similarly, the surrounding paraspinal muscles, including the quadratus lumborum and psoas major, were segmented for the respective regions. To maintain the fidelity of each segmented structure, all images were saved as distinct NIfTI (.nii.gz) files

results were then saved as mask files in the NIfTI format (Fig. 5a–d). Two spinal surgeons in this study independently performed manually MRI segmentation. First, the surgeons imported the MRI into the ITK-SNAP software (version 3.8.0, http://www.itksnap.org). The vertebral bodies and intervertebral discs were manually annotated for the sagittal T1 and T2 sequences. The borders of the psoas major, quadratus lumborum, and the entire paraspinal muscles were manually outlined and filled to ensure the inclusion of information on the spine and surrounding soft tissues for the axial sequences. The MRI segmentation results were then saved as mask files in the NIfTI format (Fig. 5e–i). After 1 month, 50 patients were randomly selected from the MRI dataset, and surgeons A and B independently performed the segmentation again. The intraclass correlation coefficient was utilized to evaluate the interobserver consistency of the vertebral segmentation. This step was taken to assess the accuracy and consistency of manual segmentation and reduce potential bias (Additional file Table S3).

Radiomics feature extraction

Feature extraction was performed on the vertebral bodies, intervertebral discs, and paraspinal muscles across all datasets to prevent data leakage. Feature selection, however, was performed only on the training cohort. Z-score normalization was applied to all images before feature extraction. The feature extraction algorithms were optimized based on the guidelines of the Image Biomarker Standardization Initiative, which helps standardize the processing of imaging data across institutions and minimizes inter-institutional variability. Radiomic features were extracted with the open-source package Pyradiomics (http://pypi.org/project/pyradiomics/) based on Python 3.8. These characteristics included first-order statistics, shape, gray-level co-occurrence matrix (GLCM), gray-level size zone matrix, gray-level run-length matrix, neighboring gray-tone difference matrix, and gray-level dependence matrix features. A detailed description of the radiomics features extracted in this study is found in the Pyradiomics documentation (http://pyradiomics.readthedocs.io). Robust normalization was applied to standardize all features by calculating the median and quartiles for each feature after extraction. Robust normalization was achieved by subtracting the median from each feature and then dividing by the interquartile range to reduce data discrepancies between different centers.

Deep transfer learning (DTL) feature extraction

The input images (preprocessed CT and MR images) were resized to $128 \times 128 \times 128$ dimensions using linear interpolation, and pixel intensities were normalized to a mean of 0 ± 1. These preprocessing steps are essential for harmonizing imaging data from different institutions and reducing potential biases caused by inter-institutional variability. We used a DTL method in the deep learning library PyTorch based on Python 3.8, similar to that used in the previous studies [35]. We selected the vision transformer 3D (ViT-3D) [36] as the base model for this study and carefully adjusted the learning rate to improve transfer performance. The final learning rate was set to 2e−05, with the following hyperparameters: image_patch_size = 16, channels = 1, dim = 1024, depth = 6, heads = 8, and mlp_dim = 2048. The fundamental concept behind the ViT is to divide the image into small patches and input them into a neural network for processing. Each

patch is independently processed, and its outputs are combined, enabling the network to learn the global structure and features of the image. In summary, the ViT decomposes clinical images into patches, processes each patch individually, and then pools the findings through concatenation to generate an entire image representation. We used the outputs of the ViT to predict whether CS would occur by applying a binary classifier after processing the raw input data. The model learns the complex features associated with CS through the ViT, enabling accurate prediction in new images. The transfer features are extracted from the penultimate layer of the model (i.e., the average pooling layer that obtains a global image representation by averaging the features of all patches); thus, we divided the model parameters into two parts: the backbone and the task-specific part. The backbone parameters were initialized with the RadImageNet pretrained model [37], whereas the task-specific parameters were randomly initialized.

Feature selection

Early fusion, also known as feature-level fusion, was performed on the radiomics features extracted from the vertebral bodies on CT images and from the vertebral bodies, intervertebral discs, and paraspinal muscles on multi-sequence MRI. Specifically, the radiomics features from these anatomical structures were concatenated into a single radiomics feature vector. We first conducted an independent-sample F test on the histogram of the co-occurrence matrix of regional features to select radiomics features with high repeatability and low redundancy, and features with p values of >0.05 were removed. Second, Pearson correlation coefficients were calculated between features. One of the highly correlated features (correlation coefficient >0.9) was retained. To maximize the representational capability of the features, a greedy recursive deletion strategy [38] was employed, where the feature with the highest correlation to others was removed at each step. We then used the least absolute shrinkage and selection operator (LASSO) algorithm to incorporate stable radiomic features into the analysis by constructing a penalty function λ that shrinks the regression coefficients to zero. Tenfold cross-validation was conducted based on the criterion of maximizing the mean cross-validation score to identify the optimal λ value. Radiomics parameters with non-zero coefficients and their weights were selected according to the model corresponding to the optimal λ. Finally, independent and stable radiomics features were determined.

Feature-level fusion was performed by concatenating them into a single-feature vector for the DLR features from the vertebral bodies, intervertebral discs, and paraspinal muscles. Z-score normalization was applied to the DLR features by subtracting the mean of each feature column and dividing it by its standard deviation, converting them to a standard normal distribution, before principal component analysis (PCA) dimensionality reduction. PCA was then utilized for dimensionality reduction to extract the principal components, which helps improve the model's generalization ability and reduce the risk of overfitting. Finally, LASSO regression was used to perform feature selection on the DLR features, retaining important features with non-zero coefficients.

Model construction and validation

We used the scikit-learn machine learning library after feature selection to construct classification models, including logistic regression (LR); Naive Bayes, linear support

vector machines; polynomial, sigmoid, and radial basis function kernels; decision trees; random forests; extremely randomized trees; eXtreme gradient boosting; AdaBoost; light gradient boosting machine; multilayer perceptions; and gradient boosting machine. We employed grid search algorithms on the training cohort to tune the hyperparameters of all models, enabling the adjustment of commonly used parameters in each model. Tenfold cross-validation was utilized on the training cohort to compare the performance of different classification models and select the optimal hyperparameters.

Receiver-operating characteristic (ROC) curves were plotted, and the area under the curve (AUC), accuracy, sensitivity, F1 score, and specificity were calculated to assess the performance of the predictive models. We separately established hand-crafted radiomics and DLR models. The output probabilities of the optimal classifiers from the radiomics model were defined as the radiomics signature (Rad-sign), and the output probabilities from the deep learning model were defined as the deep learning signature (DL-sign).

Binary LR analysis was conducted to evaluate the baseline clinical variables, such as surgical segments, sex, age, diagnosis, BMI, and osteoporosis. Statistically significant clinical variables were combined with Rad-sign and DL-sign to construct a combined model with the machine learning algorithms described above. This was done to further evaluate the model's effectiveness in determining high-risk patients for CS after lumbar fusion surgery. The modeling was performed with the PixelMed AI platform (https://github.com/410312774).

Statistical analysis

Python (version 3.8.2; https://www.python.org) and SPSS version 21.0 were used for all statistical analyses. Continuous variables with a normal distribution were presented as means ± standard deviation, and between-group comparisons were performed using independent-sample t tests. Data were compared between three groups using analysis of variance. Continuous variables with a non-normal distribution were expressed as the median (interquartile range), and comparisons were performed using the Mann–Whitney U test. Levene's test was utilized to evaluate the homogeneity of variances before conducting t tests. Categorical variables were presented as counts (n) and percentages (%), and comparisons between groups were made using the Chi-square test or Fisher's exact test. Concordance testing was performed using intraclass Correlation Coefficient (ICC). All statistical tests were two-sided, and statistical significance was set at p values of < 0.05. The performance of the classification models was assessed using ROC curves and the AUC. Decision curve analysis (DCA) was conducted by quantifying the net benefit at different threshold probabilities to evaluate the clinical value of the models. The DeLong test was conducted to compare differences in AUCs between the different models.

Clinical trial number

Not applicable.

Abbreviations

MRI	Magnetic resonance imaging
CT	Computed tomography
CS	Cage subsidence

DLR	Deep learning radiomics
ROC	Receiver operating characteristic
3D CNNs	Three-dimensional convolutional neural networks
PLIF	Posterior lumbar interbody fusion
BMI	Body mass index
GLDM	Gray-level dependence matrix
ViT-3D	Vision transformer 3D
PCA	Principal component analysis
LASSO	Least absolute shrinkage and selection operator
LR	Logistic regression
linear SVM	Linear support vector machines
XGBoost	EXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
MLP	Multilayer perceptions
GBM	Gradient boosting machine
AUC	Area under the curve
ACC	Accuracy
SEN	Sensitivity
SPE	Specificity
Rad-sign	Radiomics signature
DL-sign	Deep learning signature
DCA	Decision curve analysis
ROI	Region of interest

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12938-025-01355-y.

Additional files

Acknowledgements
Not applicable.

Author contributions
C.Z., R.C. contributed equally to this work. Conception and design: L.Z.; Acquisition of data: C.Z., R.C., H.S., Q.F.; Analysis and interpretation of data: C.Z. and B.W.; Drafting the article: C.Z. and R.C.; Critically revising the article: L.Z.; All authors have read and agreed to the published version of the manuscript.

Funding
This study was supported by Beijing Natural Science Foundation (7242059), Beijing Hospitals Authority Clinical medicine Development of special funding support (YGLX202305), and Key medical disciplines of Shijingshan district (2023006).

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was conducted in accordance with the ethical principles of the Declaration of Helsinki. The research protocol involving human data was reviewed and approved by the Ethics Committee of Beijing Chaoyang Hospital (Approval No. 2024-KE-346). The requirement for written informed consent was formally waived by the Ethics Committee for the following reasons:1. This is a retrospective analysis of fully anonymized data extracted from institutional databases;2. No personally identifiable information was accessible to researchers during data collection or analysis. Concerning legal provisions: Article 32 of the Measures for the Ethical Review of Life Sciences and Medical Research Involving Human Subjects, released by the People's Republic of China: The ethical assessment for studies utilizing data or biological samples from human subjects may be waived under certain conditions, provided that there is no harm to individuals, no sensitive information is involved, and no commercial interests are present.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 16 December 2024 Accepted: 18 February 2025
Published online: 02 March 2025

References

1. Perera RS, Dissanayake PH, Senarath U, Wijayaratne LS, Karunanayake AL, Dissanayake VHW. Associations between disc space narrowing, anterior osteophytes and disability in chronic mechanical low back pain: a cross sectional study. BMC Musculoskelet Disord. 2017;18(1):193.
2. Mobbs RJ, Phan K, Malham G, Seex K, Rao PJ. Lumbar interbody fusion: techniques, indications and comparison of interbody fusion options including Plif, Tlif, Mi-Tlif, Olif/Atp, Llif and Alif. J Spine Surg. 2015;1(1):2–18.
3. Tohmeh AG, Khorsand D, Watson B, Zielinski X. Radiographical and clinical evaluation of extreme lateral interbody fusion: effects of cage size and instrumentation type with a minimum of 1-year follow-up. Spine Phila Pa (1976). 2014;39(26):E1582–91.
4. Malham GM, Parker RM, Blecher CM, Seex KA. Assessment and classification of subsidence after lateral interbody fusion using serial computed tomography. J Neurosurg Spine. 2015;23(5):589–97.
5. Marchi L, Abdala N, Oliveira L, Amaral R, Coutinho E, Pimenta L. Radiographic and clinical evaluation of cage subsidence after stand-alone lateral interbody fusion. J Neurosurg Spine. 2013;19(1):110–8.
6. Rao PJ, Phan K, Giang G, Maharaj MM, Phan S, Mobbs RJ. Subsidence following anterior lumbar interbody fusion (Alif): a prospective study. J Spine Surg. 2017;3(2):168–75.
7. Wu H, Shan Z, Zhao F, Cheung JPY. Poor bone quality, multilevel surgery, and narrow and tall cages are associated with intraoperative endplate injuries and late-onset cage subsidence in lateral lumbar interbody fusion: a systematic review. Clin Orthop Relat Res. 2022;480(1):163–88.
8. Zhao L, Xie T, Wang X, Yang Z, Pu X, Lu Y, Zeng J. Clinical and radiological evaluation of cage subsidence following oblique lumbar interbody fusion combined with anterolateral fixation. BMC Musculoskelet Disord. 2022;23(1):214.
9. Song MX, Yang H, Yang HQ, Li SS, Qin J, Xiao Q. Mr Imaging radiomics analysis based on lumbar soft tissue to evaluate lumbar fascia changes in patients with low back pain. Acad Radiol. 2023;30(11):2450–7.
10. Bach Cuadra M, Favre J, Omoumi P. Quantification in musculoskeletal imaging using computational analysis and machine learning: segmentation and radiomics. Semin Musculoskelet Radiol. 2020;24(1):50–64.
11. Biamonte E, Levi R, Carrone F, Vena W, Brunetti A, Battaglia M, Garoli F, Savini G, Riva M, Ortolina A, Tomei M, Angelotti G, Laino ME, Savevski V, Mollura M, Fornari M, Barbieri R, Lania AG, Grimaldi M, Politi LS, Mazziotti G. Artificial intelligence-based radiomics on computed tomography of lumbar spine in subjects with fragility vertebral fractures. J Endocrinol Invest. 2022;45(10):2007–17.
12. Zhang J, Xia L, Tang J, Xia J, Liu Y, Zhang W, Liu J, Liang Z, Zhang X, Zhang L, Tang G. Constructing a deep learning radiomics model based on X-ray images and clinical data for predicting and distinguishing acute and chronic osteoporotic vertebral fractures: a multicenter study. Acad Radiol. 2024;31(5):2011–26.
13. Cheng L, Cai F, Xu M, Liu P, Liao J, Zong S. A diagnostic approach integrated multimodal radiomics with machine learning models based on lumbar spine CT and X-ray for osteoporosis. J Bone Miner Metab. 2023;41(6):877–89.
14. Liu J, Guo W, Zeng P, Geng Y, Liu Y, Ouyang H, Lang N, Yuan H. Vertebral MRI-based radiomics model to differentiate multiple myeloma from metastases: influence of features number on logistic regression model performance. Eur Radiol. 2022;32(1):572–81.
15. Staartjes VE, Siccoli A, de Wispelaere MP, Schröder ML. Patient-reported outcomes unbiased by length of follow-up after lumbar degenerative spine surgery: do we need 2 years of follow-up? Spine J. 2019;19(4):637–44.
16. Sato J, Ohtori S, Orita S, Yamauchi K, Eguchi Y, Ochiai N, Kuniyoshi K, Aoki Y, Nakamura J, Miyagi M, Suzuki M, Kubota G, Inage K, Sainoh T, Fujimoto K, Shiga Y, Abe K, Kanamoto H, Inoue G, Takahashi K. Radiographic evaluation of indirect decompression of mini-open anterior retroperitoneal lumbar interbody fusion: oblique lateral interbody fusion for degenerated lumbar spondylolisthesis. Eur Spine J. 2017;26(3):671–8.
17. Jin C, Jaiswal MS, Jeun SS, Ryu KS, Hur JW, Kim JS. Outcomes of oblique lateral interbody fusion for degenerative lumbar disease in patients under or over 65 years of age. J Orthop Surg Res. 2018;13(1):38.
18. Ran L, Xie T, Zhao L, Huang S, Zeng J. Low hounsfield units on computed tomography are associated with cage subsidence following oblique lumbar interbody fusion (Olif). Spine J. 2022;22(6):957–64.
19. Li YC, Chen HH, Horng-Shing Lu H, HondarWu HT, Chang MC, Chou PH. Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? Clin Orthop Relat Res. 2021;479(7):1598–612.
20. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186(8):1772–91.
21. Goedmakers CMW, Lak AM, Duey AH, Senko AW, Arnaout O, Groff MW, Smith TR, Vleggeert-Lankamp CLA, Zaidi HA, Rana A, Boaro A. Deep learning for adjacent segment disease at preoperative mri for cervical radiculopathy. Radiology. 2021;301(3):664–71.
22. Pitonakova L, Bullock S. The robustness-fidelity trade-off in grow when required neural networks performing continuous novelty detection. Neural Netw. 2020;122:183–95.
23. Csermely P. The wisdom of networks: a general adaptation and learning mechanism of complex systems: the network core triggers fast responses to known stimuli; innovations require the slow network periphery and are encoded by core-remodeling. Bioessays. 2018;40(1).
24. Parakh A, Lee H, Lee JH, Eisner BH, Sahani DV, Do S. Urinary stone detection on ct images using deep convolutional neural networks: evaluation of model performance and generalization. Radiol Artif Intell. 2019;1(4): e180066.
25. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019;54:280–96.
26. Xu L, Lin S, Wang Y, Xu Z. Shrinkage degree in L_2-rescale boosting for regression. IEEE Trans Neural Netw Learn Syst. 2017;28(8):1851–64.
27. Mahmood U, Apte AP, Deasy JO, Schmidtlein CR, Shukla-Dave A. Investigating the robustness neighborhood gray tone difference matrix and gray level co-occurrence matrix radiomic features on clinical computed tomography systems using anthropomorphic phantoms: evidence from a multivendor study. J Comput Assist Tomogr. 2017;41(6):995–1001.
28. Phan K, Rogers P, Rao PJ, Mobbs RJ. Influence of obesity on complications, clinical outcome, and subsidence after anterior lumbar interbody fusion (Alif): prospective observational study. World Neurosurg. 2017;107:334–41.

29. Buerba RA, Sharma A, Ziino C, Arzeno A, Ajiboye RM. Bisphosphonate and teriparatide use in thoracolumbar spinal fusion: a systematic review and meta-analysis of comparative studies. Spine (Phila Pa 1976). 2018;43(17):1014–23.
30. Hou Y, Yuan W. Influences of disc degeneration and bone mineral density on the structural properties of lumbar end plates. Spine J. 2012;12(3):249–56.
31. Hou Y, Luo Z. A study on the structural properties of the lumbar endplate: histological structure, the effect of bone density, and spinal level. Spine (Phila Pa 1976). 2009;34(12):427–33.
32. Hu W, Gao J, Wang Y, Wu O, Maybank S. Online adaboost-based parameterized methods for dynamic distributed network intrusion detection. IEEE Trans Cybern. 2014;44(1):66–82.
33. Wang X, Li J, Huang T. Cnvabnn: an adaboost algorithm and neural networks-based detection of copy number variations from Ngs data. Comput Biol Chem. 2022;99:107720.
34. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. BMJ. 2015;350: g7594.
35. Sharma AK, Nandal A, Dhaka A, Koundal D, Bogatinoska DC, Alyami H. Enhanced watershed segmentation algorithm-based modified Resnet50 model for brain tumor detection. Biomed Res Int. 2022;2022:7348344.
36. Pan D, Shen J, Al-Huda Z, Al-Qaness MAA. Vcanet: vision transformer with fusion channel and spatial attention module for 3d brain tumor segmentation. Comput Biol Med. 2025;186: 109662.
37. Mei X, Liu Z, Robson PM, Marinelli B, Huang M, Doshi A, Jacobi A, Cao C, Link KE, Yang T, Wang Y, Greenspan H, Deyer T, Fayad ZA, Yang Y. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell. 2022;4(5): e210315.
38. Lai T, Chen R, Yang C, Li Q, Fujita H, Sadri A, Wang H. Efficient robust model fitting for multistructure data using global greedy search. IEEE Trans Cybern. 2020;50(7):3294–306.

Publisher's Note