scientific reports

OPEN

Multicenter study on predicting postoperative upper limb muscle strength improvement in cervical spinal cord injury patients using radiomics and deep learning

Fabin Lin[1,2,6], Kaifeng Wang[2,5,6], Minyun Lai[2,6], Yang Wu[3], Chunmei Chen[1], Yongjiang Wang[4]✉ & Rui Wang[1]✉

Cervical spinal cord injury is often catastrophic, frequently leading to irreversible impairment. MRI has become the gold standard for evaluating spinal cord injuries (SCI). Our study aimed to assess the accuracy of a radiomics approach, based on machine learning and utilizing conventional MRI, in predicting the prognosis of patients with SCI. In a retrospective analysis of 82 SCI patients from three hospitals, we categorized them into good (n = 49) and poor (n = 33) prognosis groups. Preoperative T2-weighted MRI images were segmented using 3D-Region of Interest (ROI) techniques, and both radiomic and deep transfer learning features were extracted. These features were normalized using Z-score and harmonized via ComBat. Feature selection was performed using a greedy algorithm and Least absolute shrinkage and selection operator (LASSO), and others, followed by the calculation of radiomics scores through linear regression. Machine learning was then used to identify the most predictive radiomic features. Model performance was evaluated by analyzing the area under the curve (AUC) and other indicators.Univariate analysis indicated that the demographic characteristics of cervical spinal cord injury were not statistically significant. In the test dataset, the random forest (RF) combined with radiomics and ResNet34 demonstrated better performance, with an accuracy of 0.800 and an AUC of 0.893.Using MRI, deep learning-based radiomics signals show promise in evaluating and predicting the postoperative prognosis of these patients.

Keywords Spinal cord injuries, MRI, Radiomics, Machine learning

Abbreviations

MRI	Magnetic resonance imaging
SCI	Spinal cord injuries
ROI	Region of interest
ResNet	Residual network
AUC	Area under the curve
LR	Logistic regression
SVM	Support vector machine
RF	Random forest
NB	Naive Bayes
MLP	Multilayer perceptron
LightGBM	Light gradient boosting machine
BASIC	Brain and spinal injury center
LASSO	Least absolute shrinkage and selection operator

[1]Fujian Medical University Union Hospital, Fuzhou 350001, Fujian, China. [2]Fujian Medical University, Fuzhou 350001, Fujian, China. [3]The First People's Hospital of ChangDe City, ChangDe 410200, Hunan, China. [4]Ordos Central Hosptial, Ordos 017000, Inner Mongolia, China. [5]Fujian Medical University 2nd Clinical Medical College, Quanzhou, China. [6]These authors contributed equally: Fabin Lin, Kaifeng Wang, Minyun Lai. ✉email: drwangyj@163.com; 3177569230@qq.com

Cervical spinal cord injury (SCI) is considered a catastrophic event, often resulting in permanent impairment[1]. In the United States, over 50% of spinal cord injuries involve the cervical spine, leading to significant upper extremity dysfunction, which profoundly impacts daily living, social engagement, and work activities[2]. Additionally, it poses a long-term risk of complications and demands extensive healthcare resources[3].

Surgery is a crucial treatment modality for Cervical SCI. However, postoperative outcomes for patients have shown considerable variability. The predictors of postoperative outcomes in SCI patients have not been clearly identified[4,5].

Recent studies have focused on correlating intramedullary lesion length, edema length, edema extent, spinal cord expansion, and maximal spinal cord compression with the prognosis of SCI in sagittal/transverse planes, often overlooking the lesion's intact structure[6–8]. Radiomics provides a means to objectively and quantitatively describe lesions in their entirety, enabling precise feature imaging that traditional methods struggle to achieve[7,9]. Additionally, non-invasive medical imaging allows for high-throughput feature extraction by computers. Deep transfer learning (DTL) leverages pre-trained deep learning networks, fine-tuning them to tackle new tasks, thus enabling model training with small datasets[10]. The ResNet model, introduced by Kaiming He in 2015, is a widely used convolutional neural network. It uses skip connections to efficiently train deep networks, effectively addressing vanishing and exploding gradient issues[11]. Integrating radiomics and ResNet holds substantial promise for enhancing clinical diagnosis and treatment strategies[12].

Machine learning can effectively aid in constructing predictive and prognostic models by automatically building classification or prediction algorithms to capture robust statistical patterns in Radiomics signature data[13,14]. The synergy between machine learning and Radiomics has demonstrated benefits in diagnosing and treating spinal cord-related diseases[8,15]. Our goal was to develop a Radiomics prediction model based on T2-MRI sequences to identify the prognosis of SCI patients.

Materials and methods

Patients

This retrospective study was conducted with authorization from the ethical committees of three hospitals (approval no. 2021KY138), where the requirement for obtaining informed consent was waived. Data from 223 patients across the three hospitals, collected between January 2012 and January 2021, were used for this study.

The inclusion criteria were defined as follows: (1) confirmed diagnosis of cervical spinal cord injury (acute injury to nerve structures in the spinal canal leading to temporary or permanent sensory and motor function changes, with or without vesicorectal dysfunction), (2) complete case data and preoperative MRI image. (3) No prior treatment before surgery. The exclusion criteria included incomplete clinical records or inadequate imaging data. Initially, 223 patients were included in the study: 86 from Fujian Medical University Union Hospital, 118 from Ordos Central Hospital, and 19 from The First People's Hospital of ChangDe City. We excluded 86 patients with thoracic or lumbar spinal cord injuries. Of the remaining 137 patients, 55 were excluded due to missing or poor-quality MRI images. Ultimately, 82 patients were analyzed, divided into good prognosis (49 patients) and bad prognosis (33 patients). The patient recruitment map for this study is shown in Fig. 1A.The study design and pipeline are illustrated in Fig. 1C.

Surgical technique

All participants received neck surgery through either an anterior or posterior approach: Left/right anterior cervical approach discectomy + Microdecompression of the spinal canal + Implant fusion + Internal fixation with steel plate or Posterior median cervical approach internal fixation + Implant fusion + Decompression of the spinal canal. Moreover, based on the study conducted by Yanlin Yin et al., a comparative analysis of postoperative cervical and neurological functional recovery between the groups subjected to anterior and posterior surgical interventions revealed no significant statistical differences[16]. Subsequent follow-ups were conducted to monitor the progress. Detailed descriptions of the surgical techniques employed in this investigation are provided below:

1. Patient position: prone (posterior approach), supine (anterior approach).
2. Anesthesia method: tracheal intubation, sedation combined with general anesthesia
3. Preoperative localization: After positioning, the head frame was fixed after retraction (posterior approach) and immobilized. Under the X-ray front and side fluoroscopy of C-arm machine, the surgical segments were localized with a localization needle, and the skin of the surgical cervical segments was precisely marked.
4. Localization of puncture: After routine disinfection and spreading of towel, according to the intraoperative C-arm X-ray positive and lateral fluoroscopy film to confirm the positioning of the patient's surgical segments and puncture site, to make the appropriate size of the surgical incision, layer by layer incision of the skin, subcutaneous tissue and deep fascia.
5. Surgery: The following describes the posterior and anterior surgical operations respectively[16,17].

    5.1 Anterior approach: The procedure encompasses vertebral body resection and thorough excision of the implicated intervertebral discs. To achieve stabilization in the decompressed area, cervical fusion is facilitated using either allograft fibular struts or autografts as fillers, secured with semi-rigid locking plates. Postoperatively, cervical immobilization is maintained with a collar for a fortnight until wound healing is observed.

    5.2 Posterior approach: Utilizing a Mayfield clamp or horseshoe headrest, the cervical spine is stabilized in a neutral position. The extents of cervical laminoplasty are predetermined through preoperative imaging. Transverse mass screws are deployed and affixed with connecting rods to achieve segmental decompression by excising the ligamenta flava and laminae of the targeted segments.

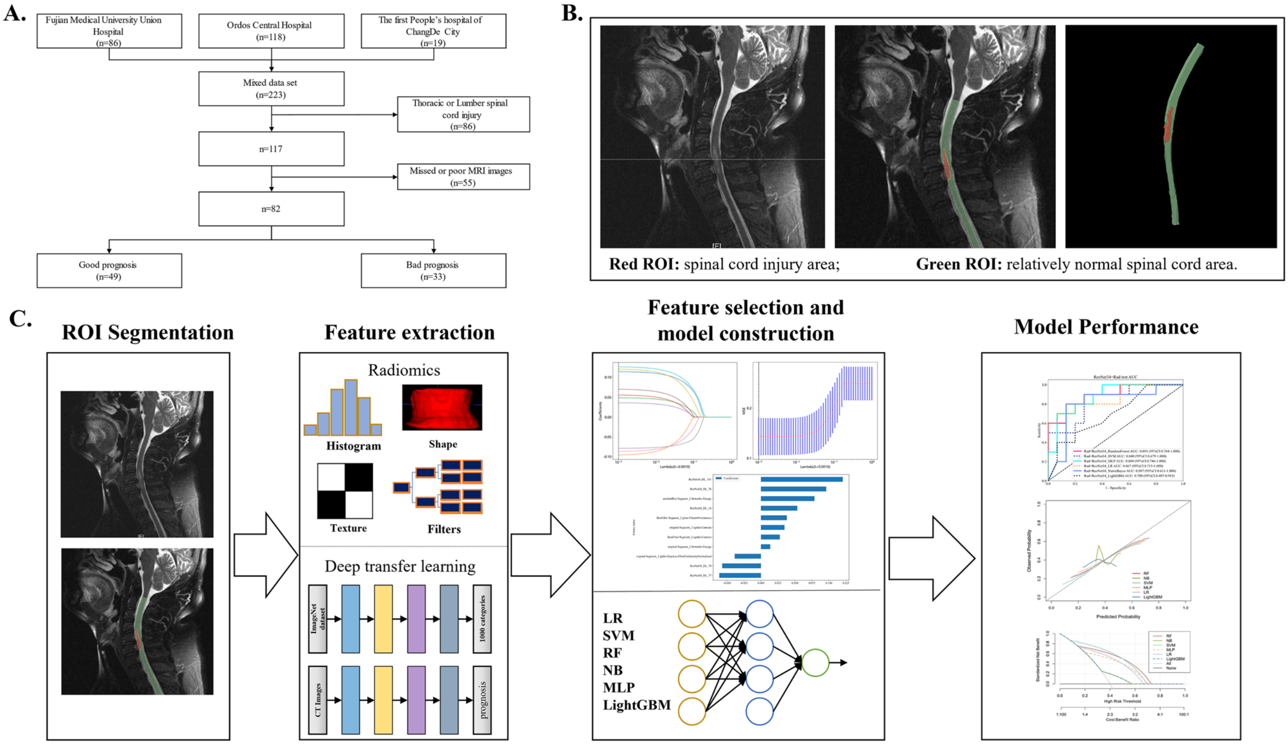
A.
Fujian Medical University Union Hospital (n=86)
Ordos Central Hospital (n=118)
The first People's hospital of ChangDe City (n=19)
Mixed data set (n=223)
Thoracic or Lumber spinal cord injury (n=86)
n=117
Missed or poor MRI images (n=55)
n=82
Good prognosis (n=49)
Bad prognosis (n=33)
B.
Red ROI: spinal cord injury area;
Green ROI: relatively normal spinal cord area.
C.
ROI Segmentation
Feature extraction
Radiomics
Histogram
Shape
Texture
Filters
Deep transfer learning
ImageNet dataset
1000 categories
CT Images
prognosis
Feature selection and model construction
LR
SVM
RF
NB
MLP
LightGBM
Model Performance

Fig. 1. (**A**) Inclusion and exclusion criteria of patients included in this study. (**B**) Examples of ROI segmentation. (**C**) Study flow chart.

6. Ensuring complete hemostasis, paravertebral muscles are repositioned, followed by suturing of the subcutaneous layer and skin. The operative site is then dressed with sterile gauze, inserting a drainage tube.
7. Postoperative Care: Patients are encouraged to start range-of-motion exercises immediately after the drainage tube is removed and are discharged once they are able to walk. Following surgery, both cohorts are managed in alignment with the hospital's clinical guidelines for cervical spine interventions, incorporating a uniform regimen of physical therapy.

Observational indicators

1. Gather perioperative data for two patient cohorts, focusing on metrics such as volume of blood loss, duration of hospitalization, costs associated with the surgery, and the specific sections of surgery performed.
2. Employ the NRS scoring system for pain assessment at the same pre- and post-operative intervals. This method gauges the intensity of pain experienced by the patient.
3. The ASIA spinal nerve function score is applied for assessing motor and sensory functions of the cervical nerves at the same times before and after the procedure. The scoring categorizes function into five grades, detailed as follows: Grade A indicates total spinal cord injury without any motor or sensory function in the sacral segments; Grade B signifies incomplete spinal cord injury with sensory but no motor function; Grade C describes incomplete spinal cord injury with both sensory and motor functions, but muscle strength below level 3; Grade D represents incomplete spinal cord injury with sensory and motor functions, where muscle strength is level 3 or above; Grade E indicates normal sensory and motor functions, with muscle strength intact.

ASIA- scale, clinical data collection

Patient demographic and clinical information, including age, gender, and the cause of spinal cord injury, were collected. The preoperative ASIA scale and postoperative ASIA scale were utilized to calculate the difference. It's worth noting that during postoperative follow-up, the ASIA score criteria remained consistent with preoperative standards. This study classified all patients into good prognosis and poor prognosis groups, computing the average ASIA scale for each group. The classification criteria are as follows:

$$\frac{\sum \text{n}\ \{[\text{ASIA motor score} - \text{Upper extremities}](\text{baseline}) - [\text{ASIA motor score} - \text{Upper extremities}](\text{one year follow})\}}{n}$$
$$= \text{the average of ASIA\_upper.}$$

The average of ASIA_upper is -4.55, Participants were classified into good (mean < -4.55) and poor (mean ≥ -4.55) outcome groups[18].

The acquisition and preprocessing of T2-weighted MRI images

The preoperative MRI scans for all enrolled patients, featuring T2-weighted imaging (T2WI), were conducted using 3.0 Tesla MRI technology. At Fujian Medical University Union Hospital, imaging was performed on a GE Signa machine with TR/TE settings of 2500–3000 ms/100–110 ms. Imaging specifications included a slice thickness of 3 mm, slice spacing of 1 mm, a matrix configuration of 256 × 512, and a field of view (FOV) of 240 mm × 240 mm. Ordos Central Hospital utilized a similar GE Signa machine with TR/TE settings of 2100 ms/120 ms, and imaging parameters mirroring those mentioned above. Likewise, The First People's Hospital of ChangDe City employed a GE Signa machine with TR/TE settings of 2000 ms/80 ms, maintaining imaging specifications consistent with the others. All T2-weighted MRI acquisitions were conducted in the sagittal plane.

After obtaining the T2-weighted MRI images, we conducted preprocessing, which included (1) normalization to the range of 0–1, (2) resampling to a voxel size of 1 × 1 × 1, and (3) Processing image noise using multiple filters: Average filtering, Box filtering, Gaussian filtering, and Median filtering. It is worth noting that each filter outputted as a separate channel rather than being overlaid. These operations were aimed at reducing multicenter effects and enhancing the generalization ability of the model. Detailed descriptions of each filter can be found in the Supplementary File 1.

ROI segmentation and radiomics analysis

ROI segmentation

The 3D Slicer Software Application v5.0.2 (https://www.slicer.org/) was employed for segmenting the Region of Interest (ROI). Each participant's ROIs underwent independent segmentation by two raters, R.W. and Y.W. The ROI comprised two segments: segmentation 1 for the spinal cord injury region and segmentation 2 for the unaffected spinal cord regions. Segmentation 1 is defined as the area of evident spinal cord injury, such as edema and significant spinal cord compression[6–8]. Segmentation 2 is defined as the region excluding the anatomical landmarks of Segmentation 1, including the upper end level aligned with the foramen magnum and the lower end level aligned with the lower edge of the first lumbar vertebra in adults. This defined the extent of the ROI, encompassing the entire spinal cord while carefully excluding vertebral and cerebrospinal fluid regions. This approach ensures that significant biomarkers outside the area of spinal cord injury are not overlooked. Figure 1B illustrates our ROI delineation. Consistency across between original images and four filtering techniques ensured uniformity in ROI delineation for each patient. Intra-rater reliability was assessed by having the same radiologist segment 30 cases two weeks apart. Inter-rater reliability was evaluated by another radiologist segmenting the same 30 cases. Intraclass correlation coefficients (ICCs) were calculated to assess both intra- and inter-rater reliability, with an ICC > 0.75 indicating strong consistency.

Radiomics feature extraction

We employed the Pyradiomics module within 3D Slicer for feature extraction, organizing the radiomics features into three categories: (I) geometry, (II) intensity, and (III) texture. Geometric features described the shape attributes of the ROI, while intensity features depicted the statistical distribution of voxel intensities within the ROI. Texture features characterized patterns and spatial arrangements of intensities through methods such as the gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), and the neighborhood gray-tone difference matrix (NGTDM). Following processing with the four filtering techniques, each patient yielded a total of 1070 radiomics features.

Deep learning analysis

Deep learning model training

Before training the deep learning models, we selected the sagittal plane image with the maximum ROI for each patient and resampled it to 224 × 224 pixels. The other image data preparation steps followed the same preprocessing process as described above. Importantly, we input five different channels into each model: the origin, average filtering, box filtering, Gaussian filtering, and median filtering. Each channel underwent an independent yet identical training process with consistent parameter settings. Notably, the five separately trained channels ultimately converged, working together to contribute to the classification task. We used three deep learning models as our initial models: ResNet18, ResNet34, and ResNet152, all pretrained on the ImageNet dataset[18]. We split the data into training and test sets in a 7:3 ratio. The learning rate was set at 0.01, with a batch size of 64, and training was done for 50 epochs. The classification targets were good prognosis (1) and poor prognosis (0).

We divided the model parameters into two parts: 1. the backbone layer (backbone) and 2. the task-specific layer (task-spec). Task-specific parameters were randomly initialized, while backbone parameters utilized the pre-trained model parameters from ImageNet. For the task-specific parameters, we applied the cosine annealing learning rate decay algorithm[19], with details provided in Supplementary File 2.

Deep learning feature extraction

We selected the penultimate layer of the model (AveragePooling layer) for feature extraction. Our dataset, which includes five categories of images, was independently trained through five identical deep learning channels. Therefore, we extracted deep learning features separately for each channel. For ResNet18 and ResNet34, the feature dimensionality was 512, while for ResNet152, it was 2048. To reduce the risk of overfitting and improve the model's generalization capability, we applied principal component analysis (PCA) to compress the extracted features to 32 dimensions per channel. As a result, each patient had 160 deep learning features(32 × 5).

ComBat compensation procedure

To harmonize post-reconstruction data across different centers or scanners and account for variations in acquisition protocols, adjustments were applied to Radiomic and DL features. This study utilized the ComBat method for harmonization, originally developed for genomic data. ComBat focuses on mitigating the center's influence on numerical feature values extracted from Radiomics and DL. In multi-center studies involving Radiomics and DL, the ComBat compensation procedure addresses differences arising from imaging protocols and various MRI scanners[19,20]. The ComBat compensation method shows considerable promise in improving the reproducibility of research conducted across multiple centers[18].

Feature fusion and selection

We combined 1070 imaging Radiomics features with 160 deep learning features, resulting in a total of 1230 fused features per patient. Before feature selection, we standardized the data using Z-scores to convert all features to a uniform mean of 0 and a standard deviation of 1.

First, we performed the Mann–Whitney *U* test and feature screening, retaining only those with a p-value < 0.05. To address high redundancy among features, we then calculated the Spearman rank correlation coefficient between features, retaining only one feature from any pair with a correlation coefficient greater than 0.9[21]. To maximize the retention of descriptive abilities, we applied a greedy recursive deletion strategy, removing the feature with the highest redundancy in each iteration. Subsequently, we used the MRMR algorithm to select the best features. In the final stage of feature filtering, we employed a tenfold cross-validated LASSO regression to identify and select features with non-zero coefficients for subsequent modeling.

Development and assessment of models

In this study, we employed six machine learning methods, implemented using the scikit-learn package (version: 0.18) in Python 3.90. These methods include Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Multilayer Perceptron (MLP), Light Gradient Boosting Machine (LightGBM).

The dataset was randomly split into two groups with a 70:30 ratio, where 70% of the data comprised the training dataset and the remaining 30% formed the test dataset to evaluate model accuracy. Grid Search was utilized to determine the optimal hyperparameters for each model.

Evaluation metrics for the models included AUC, Decision Curve Analysis (DCA), specificity (spe), sensitivity (sen), accuracy (acc), among others.

Statistical analysis

Categorical variables and normally distributed variables were analyzed using the chi-square test and t-tests, respectively. A p-value < 0.05 from the t-tests indicated statistical significance. The significance level for all statistical analyses was set at 0.05 (two-tailed). Python (version 3.9, http://www.python.org) was utilized for all analyses, and a two-sided p-value < 0.05 was considered statistically significant.

Results

Patient characteristics

A total of 82 patients traumatic cervical spinal cord injury (63males and 19 females) were included in this study. As shown in the methods section. The total dataset was split into good prognosis and bad prognosis. Table 1 summarizes the comparison of demographic characteristics of the two groups. The findings from the univariate analysis indicated that the demographic characteristics of cervical spinal cord injury lack statistical significance (p > 0.05).

Radiomics model construction

Following a series of feature selection steps, a total of six features were incorporated into the radiomics model construction, as detailed in Supplementary File S3. In our comparative analysis of different machine learning algorithms, the Random Forest (RF) and Naive Bayes (NB) algorithms demonstrated notably superior performance. The Radiomics-RF model exhibited an AUC of 0.918 (95% CI 0.8503–0.9848) in the training cohort, with sensitivity and specificity metrics of 0.909 and 0.743, respectively. Transitioning to the test cohort, the model demonstrated an AUC of 0.795 (95% CI 0.6134–0.9763), accompanied by sensitivity and specificity values of 0.750 and 0.692, respectively. (Fig. 2) Further specific information is provided in Table 2

Deep learning model construction

Our three deep learning models, each constructed with five features (Fig. S3). Specifically, the ResNet18-NB model exhibited a training AUC of 0.900 (95% CI 0.8194–0.9811), with sensitivity and specificity metrics of 0.833 and 0.818, respectively. In the test cohort, it displayed an AUC of 0.733 (95% CI 0.5327—0.9340), alongside sensitivity and specificity values of 0.900 and 0.533, respectively (Fig. 2).

The ResNet34-RF model displayed robust performance during training, achieving an AUC of 0.988 (95% CI 0.9707–1.0000), with high sensitivity and specificity values of 0.909 and 0.943, respectively. In the test cohort, its AUC remained strong at 0.853 (95% CI 0.7002–1.0000), with impressive sensitivity and specificity metrics of 0.750 and 0.769, respectively (Fig. 2).

In the training cohort, the ResNet152-RF model yielded predictions with an AUC of 0.982 (95% CI 0.9576–1.0000), along with sensitivity and specificity values of 0.909 and 0.914, respectively. Transitioning to the test cohort, its AUC was 0.788 (95% CI 0.6054–0.9715), accompanied by sensitivity and specificity values of 0.833 and 0.615, respectively (Fig. 2).

Characteristics	Good prognosis n = 49	Bad prognosis n = 33	*P*-value
Age, mean (SD), y	52.07 (12.7)	52.73 (12.53)	0.819
Sex			0.052
Male	34	29	
Female	15	4	
Time from injury to surgery, mean (SD), day	15.3 (17.87)	11.16 (9.73)	0.236
Cause of injury			0.976
Fall	20	15	
Vehicle accident	11	6	
Sports	5	3	
Other	13	9	
ASIA grade at admission			0.479
A	8	5	
B	6	9	
C	16	7	
D	15	10	
E	4	2	
ASIA motor score-upper extremities at admission, mean (SD)	25.57 (16.05)	22.64 (14.41)	0.400
ASIA grade at 1 years			0.564
A	4	3	
B	4	3	
C	5	7	
D	29	18	
E	7	2	
ASIA motor score-upper extremities at one years, mean (SD)	37.49 (11.19)	33.36 (11.4)	0.108

Table 1. The comparison of two groups' characteristics (good prognosis, bad prognosis), including age, sex, time from injury to surgery, cause of injury, ASIA grade and motor score-upper extremities at admission and 1 years. Except for the sex, cause of injury ASIA grade at admission and at 1 years, the data are presented as mean ± SD (standard deviation). The corresponding P-value is attached behind.

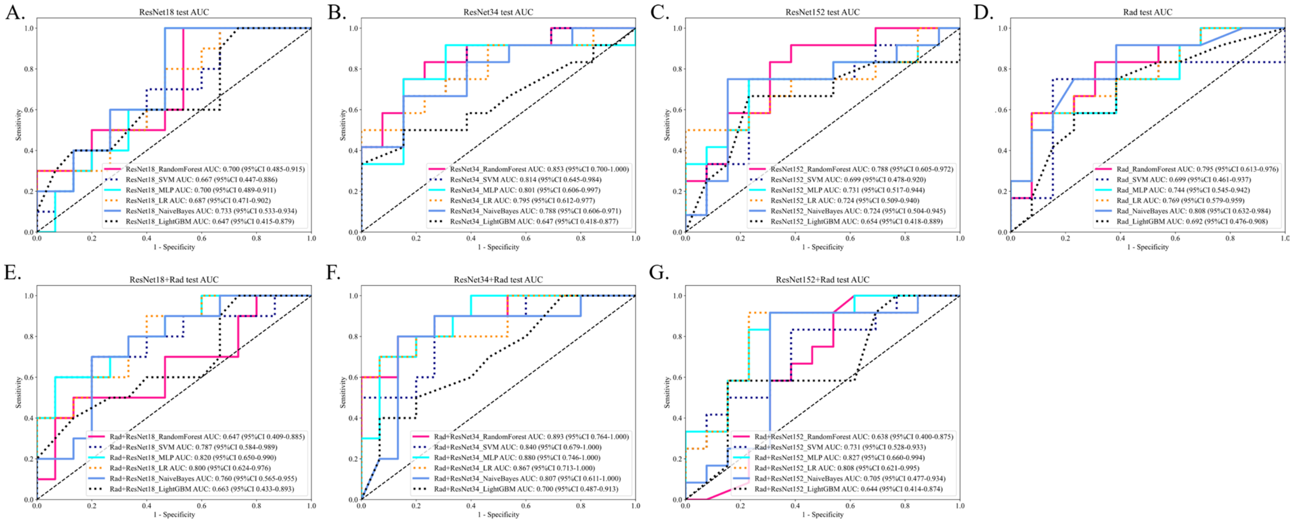

Fig. 2. ROC curves for both the training set and the test set. The receiver operating characteristic (ROC) curve graphically represents the performance of a binary classification model by displaying its sensitivity (true positive rate) versus its false positive rate at different threshold levels. AUC summarizes the ROC curve's performance in a single value, quantifying the model's ability to distinguish between classes. This figure includes the following seven models: (**A**) ResNet18 model (**B**) ResNet34 model (**C**) ResNet152 model (**D**) Radiomics model (**E**) Radiomics + ResNet18 model (**F**) Radiomics + ResNet34 model (**G**) Radiomics + ResNet152 model.

Signature	Cohort	Acc	AUC	95% CI	Sen	Spe	PPV	NPV
ResNet18_RandomForest	Train	0.947	0.984	0.9564–1.0000	0.917	0.970	0.957	0.941
ResNet18_SVM		0.860	0.918	0.8339–1.0000	0.792	0.909	0.864	0.857
ResNet18_MLP		0.825	0.913	0.8426–0.9831	0.833	0.818	0.769	0.871
ResNet18_LR		0.825	0.908	0.8343–0.9813	0.792	0.848	0.792	0.848
ResNet18_NaiveBayes		0.825	0.900	0.8194–0.9811	0.833	0.818	0.769	0.871
ResNet18_LightGBM		0.825	0.877	0.7847–0.9691	0.667	0.939	0.889	0.795
ResNet34_RandomForest		0.930	0.988	0.9707–1.0000	0.909	0.943	0.909	0.943
ResNet34_SVM		0.842	0.882	0.7722–0.9914	0.864	0.829	0.760	0.906
ResNet34_MLP		0.789	0.847	0.7413–0.9522	0.636	0.886	0.778	0.795
ResNet34_LR		0.719	0.803	0.6817–0.9235	0.727	0.714	0.615	0.806
ResNet34_NaiveBayes		0.754	0.827	0.7159–0.9386	0.773	0.743	0.654	0.839
ResNet34_LightGBM		0.789	0.852	0.7490–0.9549	0.591	0.914	0.812	0.780
ResNet152_RandomForest		0.912	0.982	0.9576–1.0000	0.909	0.914	0.870	0.941
ResNet152_SVM		0.877	0.934	0.8725–0.9950	0.773	0.943	0.895	0.868
ResNet152_MLP		0.772	0.868	0.7736–0.9615	0.864	0.714	0.655	0.893
ResNet152_LR		0.737	0.842	0.7389–0.9443	0.909	0.629	0.606	0.917
ResNet152_NaiveBayes		0.789	0.849	0.7492–0.9495	0.773	0.800	0.708	0.848
ResNet152_LightGBM		0.684	0.812	0.7058–0.9189	0.909	0.543	0.556	0.905
Rad_RandomForest		0.807	0.918	0.8503–0.9848	0.909	0.743	0.690	0.929
Rad_SVM		0.754	0.836	0.7189–0.9538	0.773	0.743	0.654	0.839
Rad_MLP		0.719	0.806	0.6894–0.9236	0.909	0.600	0.588	0.913
Rad_LR		0.737	0.813	0.6989–0.9271	0.909	0.629	0.606	0.917
Rad_NaiveBayes		0.789	0.862	0.7606–0.9640	0.773	0.800	0.708	0.848
Rad_LightGBM		0.684	0.727	0.5920–0.8612	0.455	0.829	0.625	0.707
ResNet18_RandomForest	Test	0.640	0.700	0.4849–0.9151	0.900	0.467	0.529	0.875
ResNet18_SVM		0.560	0.667	0.4469–0.8865	0.900	0.333	0.474	0.833
ResNet18_MLP		0.680	0.700	0.4893–0.9107	0.900	0.533	0.562	0.889
ResNet18_LR		0.600	0.687	0.4709–0.9024	0.700	0.533	0.500	0.727
ResNet18_NaiveBayes		0.680	0.733	0.5327–0.9340	0.900	0.533	0.562	0.889
ResNet18_LightGBM		0.680	0.647	0.4147–0.8786	0.300	0.933	0.750	0.667
ResNet34_RandomForest		0.760	0.853	0.7002–1.0000	0.750	0.769	0.750	0.769
ResNet34_SVM		0.720	0.814	0.6447–0.9835	0.833	0.615	0.667	0.800
ResNet34_MLP		0.760	0.801	0.6060–0.9966	0.833	0.692	0.714	0.818
ResNet34_LR		0.720	0.795	0.6123–0.9774	0.417	1.000	1.000	0.650
ResNet34_NaiveBayes		0.720	0.788	0.6062–0.9708	0.583	0.846	0.778	0.687
ResNet34_LightGBM		0.640	0.647	0.4176–0.8773	0.417	0.846	0.714	0.611
ResNet152_RandomForest		0.720	0.788	0.6054–0.9715	0.833	0.615	0.667	0.800
ResNet152_SVM		0.720	0.699	0.4779–0.9195	0.667	0.769	0.727	0.714
ResNet152_MLP		0.720	0.731	0.5172–0.9443	0.667	0.769	0.727	0.714
ResNet152_LR		0.720	0.724	0.5090–0.9397	0.417	1.000	1.000	0.650
ResNet152_NaiveBayes		0.760	0.724	0.5037–0.9451	0.667	0.846	0.800	0.733
ResNet152_LightGBM		0.600	0.654	0.4183–0.8894	0.333	0.846	0.667	0.579
Rad_RandomForest		0.720	0.795	0.6134–0.9763	0.750	0.692	0.692	0.750
Rad_SVM		0.760	0.699	0.4606–0.9368	0.667	0.846	0.800	0.733
Rad_MLP		0.680	0.744	0.5450–0.9421	0.500	0.846	0.750	0.647
Rad_LR		0.720	0.769	0.5792–0.9593	0.500	0.923	0.857	0.667
Rad_NaiveBayes		0.680	0.808	0.6316–0.9838	0.750	0.615	0.643	0.727
Rad_LightGBM		0.600	0.692	0.4764–0.9082	0.583	0.615	0.583	0.615

Table 2. Performance of ResNet18, ResNet34, ResNet152, and Radiomics across 6 machine learning methods. *Acc* accuracy, *AUC* area under the curve, *95% CI* 95% confidence interval, *Sen* sensitivity, *Spe* specificity, *PPV* positive predictive value, *NPV* negative predictive value.

Combined model construction

Among the three combined models we developed (Fig. 2), the ResNet18-Radiomics and ResNet152- Radiomics combined models utilized 10 features each (5 radiomics features and 5 deep learning features). For detailed information, please refer to Table 3 and the supplementary file. The radiomics + ResNet34-RF model demonstrated remarkable predictive capabilities and 11 features (6 radiomics features and 5 deep learning features) were used to build this combined model. During training, it achieved an AUC of 0.968 (95% CI 0.9299–1.000), showcasing sensitivity and specificity values of 0.875 and 0.939, respectively. In the test cohort, the model continued to perform well, with an AUC of 0.893 (95% CI 0.7639–1.000) and sensitivity and specificity metrics of 0.700 and 0.867, respectively. The DCA curves presented in Fig. 3 support our findings. Additionally, we evaluated the feature importance of the optimal model, RF, to enhance interpretability, showing that all features contributed to the model[12] (Fig. 3). For further details, please refer to Table 2. The LASSO and feature weight analysis are shown in Fig. 4.

Grad-CAM

To enhance model interpretability, we employed Grad-CAM to visualize the final convolutional layer. The darker blue regions highlight areas critical for prognosis prediction (Fig. 5). These regions are mainly concentrated on the edematous spinal cord, deformed vertebrae, and even the edematous ligaments.

Signature	Cohort	Acc	AUC	95% CI	Sen	Spe	PPV	NPV
Rad + ResNet18_RandomForest	Train	0.965	0.999	0.9952–1.0000	0.958	0.970	0.958	0.970
Rad + ResNet18_SVM		0.930	0.976	0.9417–1.0000	0.833	1.000	1.000	0.892
Rad + ResNet18_MLP		0.895	0.975	0.9444–1.0000	0.875	0.909	0.875	0.909
Rad + ResNet18_LR		0.860	0.960	0.9182–1.0000	0.958	0.788	0.767	0.963
Rad + ResNet18_NaiveBayes		0.877	0.932	0.8649–0.9988	0.875	0.879	0.840	0.906
Rad + ResNet18_LightGBM		0.807	0.883	0.7976–0.9676	0.833	0.788	0.741	0.867
Rad + ResNet34_RandomForest		0.912	0.968	0.9299–1.0000	0.875	0.939	0.913	0.912
Rad + ResNet34_SVM		0.877	0.920	0.8420–0.9989	0.750	0.970	0.947	0.842
Rad + ResNet34_MLP		0.877	0.934	0.8728–0.9959	0.708	1.000	1.000	0.825
Rad + ResNet34_LR		0.842	0.899	0.8115–0.9865	0.708	0.939	0.895	0.816
Rad + ResNet34_NaiveBayes		0.789	0.823	0.7117–0.9347	0.708	0.848	0.773	0.800
Rad + ResNet34_LightGBM		0.737	0.811	0.7010–0.9215	0.542	0.879	0.765	0.725
Rad + ResNet152_RandomForest		0.930	0.951	0.8896–1.0000	0.864	0.971	0.950	0.919
Rad + ResNet152_SVM		0.912	0.940	0.8700–1.0000	0.818	0.971	0.947	0.895
Rad + ResNet152_MLP		0.895	0.930	0.8525–1.0000	0.773	0.971	0.944	0.872
Rad + ResNet152_LR		0.860	0.905	0.8098–1.0000	0.727	0.943	0.889	0.846
Rad + ResNet152_NaiveBayes		0.772	0.831	0.7229–0.9394	0.636	0.857	0.737	0.789
Rad + ResNet152_LightGBM		0.719	0.790	0.6684–0.9109	0.591	0.800	0.650	0.757
Rad + ResNet18_RandomForest	Test	0.680	0.647	0.4087–0.8847	0.400	0.867	0.667	0.684
Rad + ResNet18_SVM		0.760	0.787	0.5843–0.9891	0.500	0.933	0.833	0.737
Rad + ResNet18_MLP		0.760	0.820	0.6499–0.9901	0.500	0.933	0.833	0.737
Rad + ResNet18_LR		0.680	0.800	0.6240–0.9760	0.800	0.600	0.571	0.818
Rad + ResNet18_NaiveBayes		0.720	0.760	0.5653–0.9547	0.600	0.800	0.667	0.750
Rad + ResNet18_LightGBM		0.680	0.663	0.4335–0.8932	0.300	0.933	0.750	0.667
Rad + ResNet34_RandomForest		0.800	0.893	0.7639–1.0000	0.700	0.867	0.778	0.812
Rad + ResNet34_SVM		0.760	0.840	0.6794–1.0000	0.800	0.733	0.667	0.846
Rad + ResNet34_MLP		0.800	0.880	0.7464–1.0000	0.600	0.933	0.857	0.778
Rad + ResNet34_LR		0.800	0.867	0.7131–1.0000	0.600	0.933	0.857	0.778
Rad + ResNet34_NaiveBayes		0.800	0.807	0.6111–1.0000	0.700	0.867	0.778	0.812
Rad + ResNet34_LightGBM		0.680	0.700	0.4874–0.9126	0.300	0.933	0.750	0.667
Rad + ResNet152_RandomForest		0.680	0.638	0.4004–0.8753	0.917	0.462	0.611	0.857
Rad + ResNet152_SVM		0.680	0.731	0.5281–0.9335	0.750	0.615	0.643	0.727
Rad + ResNet152_MLP		0.760	0.827	0.6597–0.9942	0.833	0.692	0.714	0.818
Rad + ResNet152_LR		0.800	0.808	0.6207–0.9947	0.833	0.769	0.769	0.833
Rad + ResNet152_NaiveBayes		0.760	0.705	0.4766–0.9337	0.833	0.692	0.714	0.818
Rad + ResNet152_LightGBM		0.680	0.644	0.4140–0.8744	0.500	0.846	0.750	0.647

Table 3. Performance of ResNet18 + Rad, ResNet34 + Rad and ResNet152 + Rad across 6 machine learning methods. *Acc* accuracy, *AUC* area under the curve, *95% CI* 95% confidence interval, *Sen* sensitivity, *Spe* specificity, *PPV* positive predictive value, *NPV* negative predictive value.

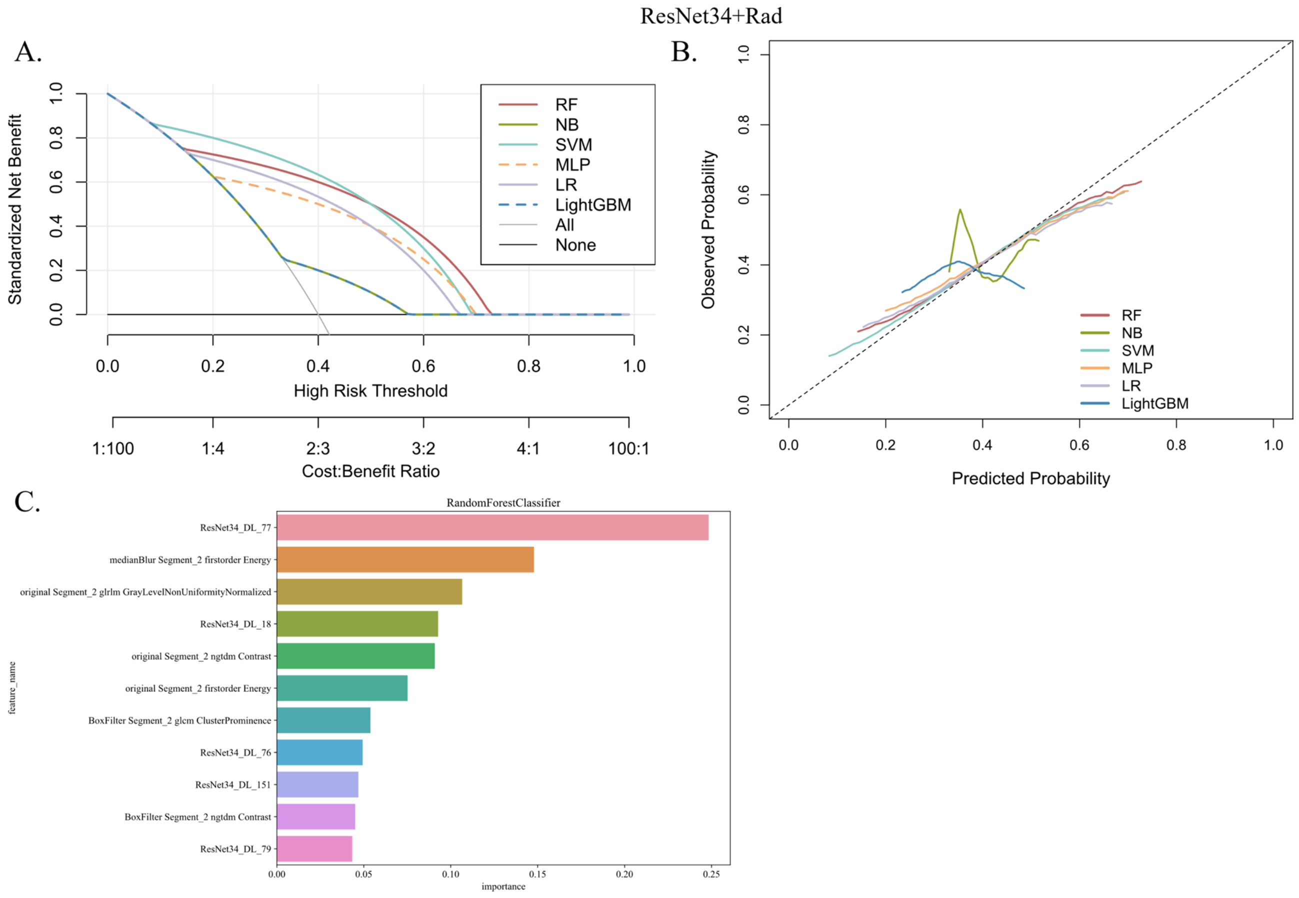

Fig. 3. In the combined ResNet34 + radiomics model: (**A**) decision curve analysis (DCA) (**B**) Calibration curve (**C**) Feature importance of random forest (RF). RF and SVM demonstrated better clinical decision-making capabilities on the DCA (> 20%) and calibration curves.

Discussion

Results of this study

We developed three deep learning models for comparison, with ResNet34 excelling in predicting the one-year prognosis of spinal cord injury. The combined RF model using ResNet34 and Radiomics demonstrated excellent performance (Accuracy = 0.800, AUC = 0.893, Sensitivity = 0.700, Specificity = 0.867). This shows that Radiomics and Deep Learning can create efficient predictive models for forecasting long-term motor function outcomes in SCI patients.

Among the 11 combined model features, DL features were more significant. For Radiomics features, three Radiomics features came from Median Blur and Box Filter . Overall, MedianBlur reduces noise while preserving edges, while BoxFilter smooths but blurs edges. Results (NB-Radiomics: AUC = 0.808) maybe suggest their combination improves generalization. However, Jiang Zhang et al. found high-pass wavelet-filtered texture features are low-repeatable, especially at lower bin counts. Future research should verify each filter's generalization capabilities[22].

Additionally, we observed that the features were not entirely consistent when establishing the three combined models. This inconsistency may be due to certain deep learning features having a high Spearman rank correlation coefficient with radiomics features, leading to the retention of the more meaningful feature among them.

Comparison with previous studies

In the study by Hyun-Joon Yoo et al., machine learning algorithms were utilized to predict gait function at discharge from acute inpatient rehabilitation facilities in SCI patients based solely on clinical variables and achieved promising results[23]. However, this approach ignored the contribution of imaging to prognosis, especially MRI. MRI holds significant prognostic value for assessing lesion sites in the spinal cord, with its metrics being proven reproducible and possessing predictive validity for clinical outcomes[24]. The BASIC (Brain and Spinal Injury Center) score was developed on this basis to predict neurological improvement using pathological high signal intensity on axial T2-weighted MRI[25]. As noted by J. Haefeli et al., the BASIC score is notably effective in predicting neurological improvement[26]. However, their study focused on a limited set of manually observed and measured MRI features (such as injury length, maximum spinal canal compromise, and maximum spinal cord compression), neglecting numerous potential features. In another study utilizing deep learning and machine learning, the model established to predict short-term neurological outcomes in acute cervical spinal cord injury

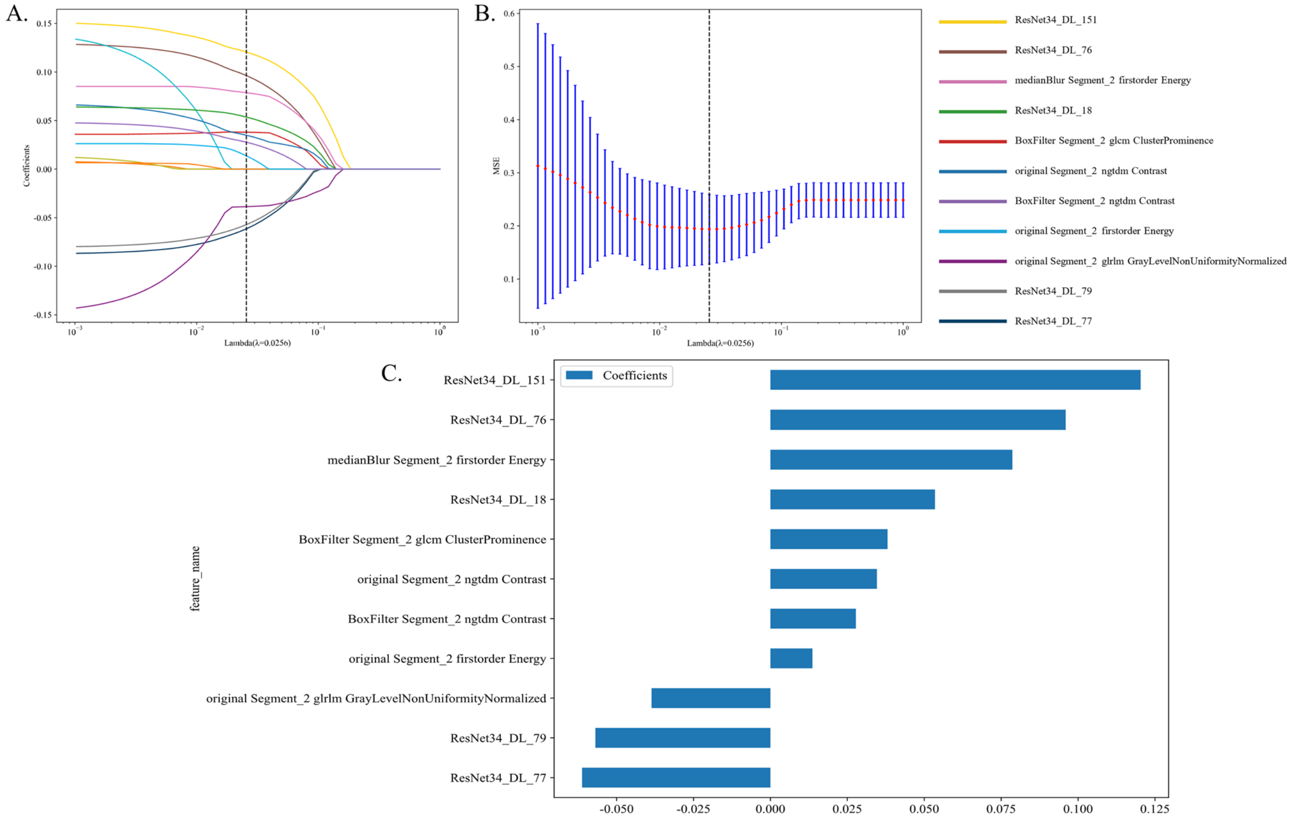

Fig. 4. In the Rad + ResNet34 model: (**A**) Incorporate logistic regression via the least absolute shrinkage and selection operator (LASSO) method, featuring a tenfold validated mean squared error (MSE), and utilizing feature weights and non-zero coefficients for constructing the Rad-score. (**B**) MSE validated across 10 folds. (**C**) Feature weights: Rad-score histogram derived from LASSO-selected features.

achieved an accuracy of 0.714 with a combined model of deep learning and RF. Similarly, in our research, this was enhanced to an ACC of 0.800 and an AUC of 0.893 (Rad + ResNet34_Random Forest).

As mentioned, previous studies have not fully exploited the advantages of MRI. Our study considered not only lesion sites but also normal spinal cord areas as 3D ROIs on T2, while accounting for the degree of spinal cord rotation[7]. The strength of our research lies in the effective use of radiomics and deep learning to analyze and extract data from MRI, thus objectively building predictive models. These methods allow for the extraction of numerous quantitative features from medical images, enabling precise evaluation. Radiomics analysis proves to be more accurate than conventional MRI, especially since many spinal cord injury patients exhibit higher eccentricity and smaller anteroposterior diameter[7,27].

André Wirries et al. demonstrated that deep learning models can be trained on small datasets for clinical decisions in lumbar disc herniation treatment[28]. Our research also shows that deep learning networks significantly contribute to predicting spinal cord injury prognosis. According to Table 2, deep learning models outperformed radiomics models in predictive performance, highlighting their superior capabilities. Notably, ResNet34 exhibited better performance, aligning with Bingbing Xiao's findings where ResNet34 surpassed other ResNet models in classifying microcalcification clusters. Xiao noted that shallow networks struggle to extract significant features, leading to poorer performance compared to deeper CNNs, which excel at identifying complex image features. However, he also highlighted the risk of overfitting with excessively deep networks, emphasizing the need for balanced model depth for optimal performance and generalizability[29]. This insight suggests that while decreasing Top-1 and Top-5 error rates can be achieved with increased depth[11], selecting the appropriate neural network depth based on task complexity and dataset size may yield greater benefits than blindly pursuing deeper networks. Our study reflects this, as smaller datasets may suffice for shallower networks like ResNet34 to extract effective features, whereas deeper networks like ResNet152 might capture too many irrelevant features, reducing model performance. This might indicates that predicting postoperative motor recovery in spinal cord injury may not require very deep neural networks, and shallower networks can provide significant contributions while saving training time and computational resources. Kaiming He et al. found that as the ResNet network deepens, training error increases, a problem also seen in other neural networks[11]. This may also explain why deeper models sometimes underperform.

It's worth noting that a series of measures conducive to model fitting and preventing overfitting are necessary. For instance, we employed strict feature selection processes to maintain feature relevance and avoid the risk of overfitting from excessive features. Additionally, various processing steps during deep learning training, such as random shifts, rotations, and horizontal flips, contribute to enhancing data diversity and preventing overfitting[30].

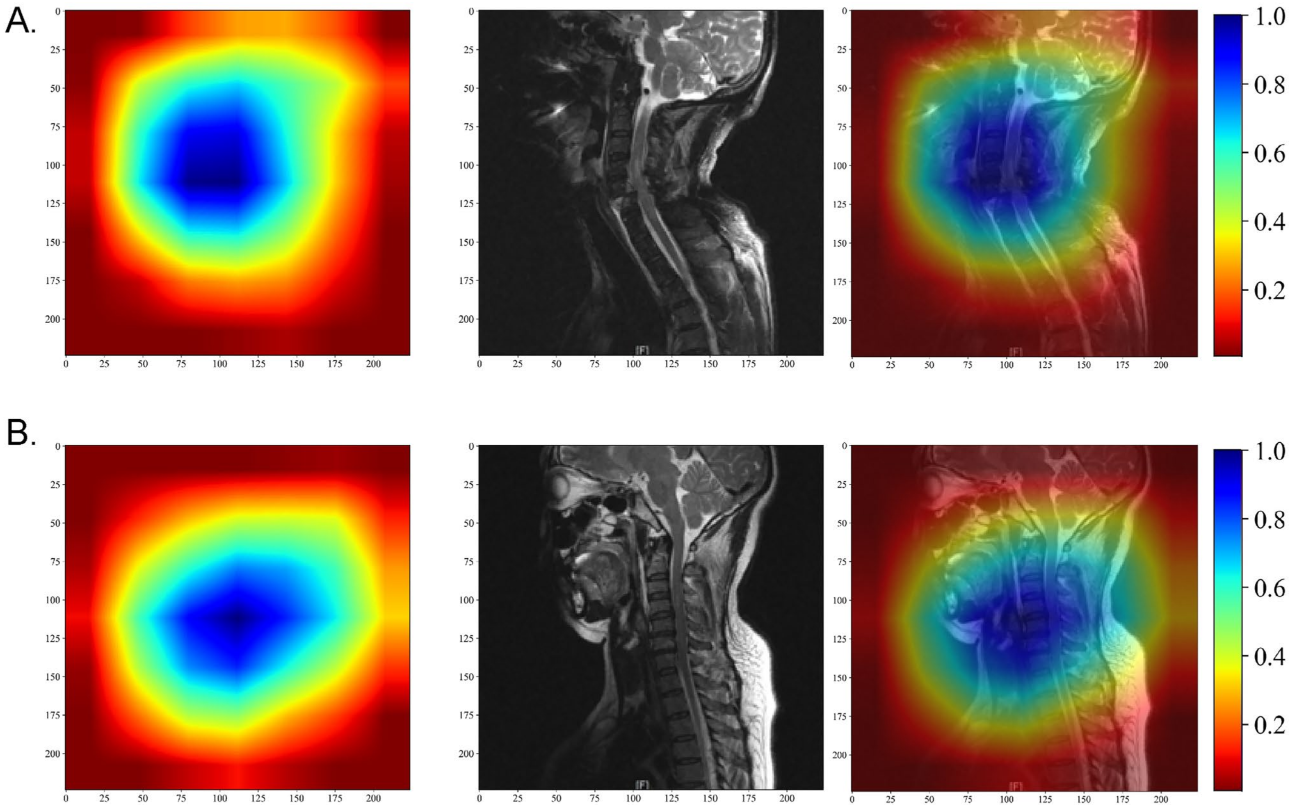

Fig. 5. Two examples of Grad-CAM visualizations. In these examples, the darker blue areas represent regions of attention, indicating that these areas may be significant for prognosis diagnosis.

Model interpretability

While AI has significantly impacted disease diagnosis and prognosis prediction, its interpretability and black-box nature may hinder widespread clinical use[31,32]. Some studies use post-hoc methods or supervised machine learning models to interpret deep learning algorithm outputs[33]. In this regard, we output the feature importance of the optimal RF model[12]. For deep learning, three primary methods improve interpretability[34]: (1). Simplifying models with techniques like LIME and model compression. (2). Visualizing CNN features through gradient ascent, deconvolution, and saliency maps. (3). Applying perturbation tests and gradient-based methods like SmoothGrad, Guided Backpropagation, and Grad-CAM. This study uses Grad-CAM visualization methods, with heatmaps likely indicating regions important for prognosis prediction. This view is endorsed by Kim, Y.[35].

Limitations

However, there are some limitations to this study. First, although this study is a multicenter retrospective study, the sample size is small. Due to the small sample size, it may prevent us from discovering how much clinical features contribute to the predictive model. To confirm the findings, we will need to increase the sample size in future research. Second, we included only a few key clinical features to predict postoperative recovery in SCI patients, and we aspire to enhance and validate the findings of this cohort study by expanding the dataset size and enriching the depth of information content. Additionally, it's important to note that the SCI patients who contributed to developing the model in this study were primarily observed for one year following their enrollment. This limited follow-up duration may constrain the model's predictive accuracy over more extended periods of disease progression.

Conclusion

Recent findings reinforce our theory that utilizing radiomics with standard MRI techniques can accurately predict the postoperative prognosis for spinal cord injury (SCI) patients. Deep transfer learning-based radiomics signatures may be an important method for assessing and monitoring postoperative prognosis in patients with SCI and may help develop more appropriate clinical treatment strategies.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 25 May 2024; Accepted: 9 September 2024
Published online: 17 February 2025

References

1. Yang, F. & Guo, X. Research on rehabilitation effect prediction for patients with SCI based on machine learning. *World Neurosurg.* **158**, e662–e674 (2022).
2. *National Spinal Cord Injury Statistical Center. Facts and Figures at a Glance* (2015).
3. Tator, C. H. *et al.* Contemporary management of spinal cord injury: From impact to rehabilitation. *Spine J.* **1**(5), 384–385 (2001).
4. Aarabi, B. *et al.* Intramedullary lesion length on postoperative magnetic resonance imaging is a strong predictor of ASIA impairment scale grade conversion following decompressive surgery in cervical spinal cord injury. *Neurosurgery* **80**(4), 610–620 (2017).
5. Wichmann, T. O. *et al.* Early clinical predictors of functional recovery following traumatic spinal cord injury: A population-based study of 143 patients. *Acta Neurochir.* **163**(8), 2289–2296 (2021).
6. Tarawneh, A. M. *et al.* Can MRI findings predict the outcome of cervical spinal cord Injury? A systematic review. *Eur. Spine J.* **29**(10), 2457–2464 (2020).
7. Zhang, M. Z. *et al.* Predicting postoperative recovery in cervical spondylotic myelopathy: Construction and interpretation of T2(*)-weighted radiomic-based extra trees models. *Eur. Radiol.* **32**(5), 3565–3575 (2022).
8. Burns, J. E., Yao, J. & Summers, R. M. Vertebral body compression fractures and bone density: Automated detection and classification on CT images. *Radiology* **284**(3), 788–797 (2017).
9. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**(4), 441–446 (2012).
10. Benjamens, S., Dhunnoo, P. & Mesko, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **3**, 118 (2020).
11. He, K.A.Z., Xiangyu and Ren, S. and Sun, J., Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).
12. Huang, Y. *et al.* Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: A multicenter, retrospective study. *EClinicalMedicine* **58**, 101899 (2023).
13. Merali, Z. G. *et al.* Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One* **14**(4), e0215133 (2019).
14. Currie, G. *et al.* Machine learning and deep learning in medical imaging: Intelligent imaging. *J. Med. Imaging Radiat. Sci.* **50**(4), 477–487 (2019).
15. Chang, M. *et al.* The role of machine learning in spine surgery: The future is now. *Front. Surg.* **7**, 54 (2020).
16. Yin, Y. *et al.* The efficacy of anterior cervical corpectomy and fusion and posterior total laminectomy on cervical spinal cord injury and quality of life. *Comput. Math. Methods Med.* **2022**, 8216339 (2022).
17. Seng, C. *et al.* Surgically treated cervical myelopathy: A functional outcome comparison study between multilevel anterior cervical decompression fusion with instrumentation and posterior laminoplasty. *Spine J.* **13**(7), 723–731 (2013).
18. Bi, S. *et al.* Multi-parametric MRI-based radiomics signature for preoperative prediction of Ki-67 proliferation status in sinonasal malignancies: A two-centre study. *Eur. Radiol.* https://doi.org/10.1007/s00330-022-08780-w (2022).
19. Orlhac, F. *et al.* Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* **291**(1), 53–59 (2019).
20. Lucia, F. *et al.* External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* **46**(4), 864–877 (2019).
21. Wang, W. *et al.* Development and validation of a computed tomography-based radiomics signature to predict response to neoadjuvant chemotherapy for locally advanced gastric cancer. *JAMA Netw. Open* **4**(8), e2121143 (2021).
22. Zhang, J. *et al.* Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients. *Radiother. Oncol.* **183**, 109578 (2023).
23. Yoo, H. J. *et al.* Prediction of gait recovery using machine learning algorithms in patients with spinal cord injury. *Medicine (Baltimore)* **103**(23), e38286 (2024).
24. Andreoli, C. *et al.* MRI in the acute phase of spinal cord traumatic lesions: Relationship between MRI findings and neurological outcome. *Radiol. Med.* **110**(5–6), 636–645 (2005).
25. Talbott, J. F. *et al.* The brain and spinal injury center score: A novel, simple, and reproducible method for assessing the severity of acute cervical spinal cord injury with axial T2-weighted MRI findings. *J. Neurosurg. Spine* **23**(4), 495–504 (2015).
26. Haefeli, J. *et al.* Multivariate analysis of MRI biomarkers for predicting neurologic impairment in cervical spinal cord injury. *AJNR Am. J. Neuroradiol.* **38**(3), 648–655 (2017).
27. van de Stadt, S. I. W. *et al.* Spinal cord atrophy as a measure of severity of myelopathy in adrenoleukodystrophy. *J. Inherit. Metab. Dis.* **43**(4), 852–860 (2020).
28. Wirries, A. *et al.* Artificial intelligence facilitates decision-making in the treatment of lumbar disc herniations. *Eur. Spine J.* **30**(8), 2176–2184 (2021).
29. Xiao, B. *et al.* Classification of microcalcification clusters in digital breast tomosynthesis using ensemble convolutional neural network. *Biomed. Eng. Online* **20**(1), 71 (2021).
30. Kuo, B. I. *et al.* Keratoconus screening based on deep learning approach of corneal topography. *Transl. Vis. Sci. Technol.* **9**(2), 53 (2020).
31. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**(141), 20170387 (2018).
32. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
33. Jiang, Y. *et al.* Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun. (Lond.)* **40**(4), 154–166 (2020).
34. Papadimitroulas, P. *et al.* Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys. Med.* **83**, 108–121 (2021).
35. Kim, Y. *et al.* A CT-based deep learning model for predicting subsequent fracture risk in patients with hip fracture. *Radiology* **310**(1), e230614 (2024).

Acknowledgements

We thank the colleagues in our department for their help in our study.

Author contributions

Conceptualization: Y.W., C.C. Formal Analysis: F.L., K.W., M.L., Y.W., R.W. Investigation: F.L., K.W., M.L., Methodology: F.L., K.W. Project Administration: Y.W., C.C. Data curation: F.L., Y.W, R.W Funding acquisition: Y.W., C.C. Writing – Original Draft: F.L., K.W., M.L., Writing – Review & Editing: Y.W., C.C., Y.W., R.W.

Funding

This study was supported by Fujian Provincial Key Clinical Specialty Construction Project (Official document No.884 [2022]）.

Competing interests

The authors declare no competing interests.

Ethical approval

Institutional Review Board approval was obtained. The study was conducted according to the guidelines of the Declaration of Helsinki, and was approved by Institutional Review Board (approval no. 2021KY138).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72539-0.

Correspondence and requests for materials should be addressed to Y.W. or R.W.

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