Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Preliminary Development of a Multimodal Deep Learning Model Integrating Clinical Data and Dual-Plane MRI to Predict 1-Year Patient-Reported Lower-Limb Numbness Improvement After Tubular Microdiscectomy

**Yixuan Liu**
Fujian Medical University
**Wenhui Zhao**
Fujian Medical University Union Hospital
**Xiaoqiang Wei**
Fujian Medical University Union Hospital
**Kaifeng Wang**
Fujian Medical University
**Gang Gao**
Ordos Central Hospital
**Yongjiang Wang**
Ordos Central Hospital
**Rui Wang**
fjxhwr@sina.com

Fujian Medical University Union Hospital

**Research Article**

**Keywords:** Deep learning, Magnetic resonance imaging, Lumbar disc herniation, Microdiscectomy, Numbness

**Posted Date:** November 3rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11305141/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** No competing interests reported.

## Abstract

# Objective

To develop and preliminarily evaluate a multimodal model integrating clinical data and dual-plane magnetic resonance imaging (MRI) features to predict 1-year patient-reported lower-limb numbness improvement after tubular microdiscectomy (TMD).

# Materials and Methods

This retrospective two-center study included 214 adults: 111 in the training cohort, 48 in the internal validation cohort, and 55 in the external validation cohort. Mean ages were 53.2 ± 15.0, 54.1 ± 14.1, and 51.4 ± 14.7 years, respectively; male participants comprised 58 (52.3%), 28 (58.3%), and 28 (50.9%), and female participants 53 (47.7%), 20 (41.7%), and 27 (49.1%), respectively. Deep features were extracted from sagittal and transverse T2-weighted MRI using residual networks and combined with clinical variables in a support vector machine. Outcome-informed feature selection and tuning were confined to the training cohort using 10-fold cross-validation. Candidate models and the decision threshold were selected using Hospital 1 internal validation data. ComBat parameters were estimated from Hospital 1 training data and applied to Hospital 2 without re-estimation; Hospital 2 did not inform model development.

# Results

The final combined model incorporated 25 dual-plane MRI-derived features and four clinical variables. It yielded an area under the receiver operating characteristic curve of 0.868 in the Hospital 1 internal validation cohort used for model selection and 0.833 (95% CI, 0.720–0.947) in the Hospital 2 external validation cohort. Comparisons with the clinical-only and dual-plane imaging-only models were descriptive and should be interpreted as exploratory.

# Conclusion

This study provides preliminary model-development evidence that clinical and dual-plane MRI features may aid prediction of 1-year patient-reported numbness improvement after TMD. Larger prospective, multicenter, multivendor studies using validated sensory outcomes are needed before individualized clinical application.

## 1 Introduction

Lumbar disc herniation (LDH) is a common cause of radicular pain and lower-limb sensory disturbance [1]. Surgery is considered when neurological symptoms substantially impair daily activities or persist despite appropriate conservative treatment [2]. Tubular microdiscectomy (TMD) provides minimally invasive access for neural decompression [3]; however, residual leg numbness has been reported in up to one-third of patients after discectomy [4, 5]. As Huang et al. reported, unlike the rapid alleviation of postoperative pain, the resolution of numbness is often protracted [6]. This persistent condition extends beyond mere sensory discomfort; it significantly diminishes postoperative patient satisfaction [7].

Existing prediction models for lumbar spine surgery have focused primarily on disability, pain, or overall surgical success [8, 9]. However, clinical variables alone may not represent the anatomical extent and spatial pattern of disc herniation and neural compression. Transverse MRI depicts the cross-sectional relationship between the herniated disc, spinal canal, and nerve roots, whereas sagittal MRI characterizes disc degeneration, migration, and the longitudinal extent of compression. Deep-learning features derived from both planes could therefore provide information complementary to preoperative clinical variables [10].

This study aimed to develop and preliminarily evaluate a prognostic model integrating clinical characteristics with deep learning features extracted from sagittal and transverse MRI. The model was designed to estimate patient-reported improvement in lower-limb numbness 1 year after tubular microdiscectomy.

## 2 Materials and Methods

# 2.1 Study Design and Participants

This retrospective investigation adhered to the ethical principles of the Declaration of Helsinki and received approval from the Ethics Committee (approval no. 2022KY026). Data were obtained from two clinical centers and included patients who underwent unilateral TMD between January 2016 and January 2021. The study included 214 adults: 111 in the training cohort, 48 in the internal validation cohort, and 55 in the external validation cohort. Mean ages (mean ± SD) were 53.2 ± 15.0, 54.1 ± 14.1, and 51.4 ± 14.7 years, respectively; male participants numbered 58 (52.3%), 28 (58.3%), and 28 (50.9%), whereas female participants numbered 53 (47.7%), 20 (41.7%), and 27 (49.1%), respectively. The internal validation cohort was used for candidate-model comparison and selection of the final model, whereas the external validation cohort was reserved for a single evaluation after the model specification and decision threshold had been fixed. Hospital 2 images, extracted features, feature distributions, outcome labels, and performance estimates were not used to estimate or update any data-dependent preprocessing or harmonization parameters, select features, fit or tune models, choose architectures or classifiers, select the decision threshold, or make any other model-development decision. This retrospective two-center prediction-model study was reported in accordance with the STROBE and TRIPOD + AI recommendations [11, 12].

Eligible patients were adults with preoperative lower-limb numbness, radicular symptoms, and imaging-confirmed L3–S1 LDH who completed standardized unilateral TMD and postoperative rehabilitation. Patients were excluded for inadequate MRI or follow-up data, lumbar instability, severe systemic or psychiatric disease, or disorders likely to confound neurological assessment. Complete eligibility criteria are provided in Supplementary Material 1, Section S1. The study workflow is shown in Fig. 1.

## 2.2 Surgery and MRI Acquisition

All patients underwent standardized unilateral TMD performed by experienced spine surgeons. Preoperative sagittal and transverse T2-weighted images were acquired using 3.0-T GE Signa systems with 8-channel spine coils. Nine sagittal and 13 transverse slices were available for each patient. Surgical details are provided in Supplementary Material 1, Section S2, and center-specific MRI acquisition parameters in Supplementary Material 1, Section S3.

## 2.3 Outcome Definition

The primary outcome was a binary patient-reported outcome (PRO), defined as improvement in lower-limb numbness 12 months after TMD relative to the preoperative status among patients with baseline numbness. Complete or partial symptomatic relief was classified as improved (label 1), whereas unchanged or worsened symptoms were classified as non-improved (label 0). Detailed labeling criteria are provided in Supplementary Material 1, Section S4. The outcome was predefined before model development and used solely as the supervised learning target. Postoperative measures, including patient satisfaction, were used descriptively and not for outcome labeling or prediction.

## 2.4 Candidate Predictors and MRI Regions

Baseline characteristics and medical history were obtained from electronic medical records, and conventional imaging variables were evaluated in the Picture Archiving and Communication System. Only variables available before surgery were considered candidate predictors; postoperative and intraoperative variables were excluded. A total of 43 preoperative clinical and conventional imaging variables remained.

Three clinicians with 20, 10, and 22 years of experience performed the measurements and initial image review, with disagreements resolved by consensus. Definitions and coding of the evaluated variables are provided in Supplementary Material 1, Section S5.

Images were resampled to 1×1×1 $mm^3$, normalized to an intensity range of 0–1, and centered. Two clinicians, blinded to each other's assessments, independently selected the sagittal and transverse slices showing the most prominent herniation and placed a 224×224-pixel crop encompassing the disc, adjacent spinal canal, dural sac, and nerve-root region [10]. A senior clinician reviewed all selections. One sagittal and one transverse crop were retained per patient. This procedure represented manual cropping rather than pixel-wise segmentation.

## 2.5 Deep Learning Architecture and Feature Extraction

Sagittal and transverse ROIs were utilized to train ResNet18, ResNet50, and ResNet101 architectures (training details are provided in Supplementary Material 1, Section S6). The training process incorporated data augmentation—including random horizontal/vertical translation, vertical flipping, and 90-degree rotations—and employed Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance model interpretability. Features were extracted via average pooling layers. For the ResNet50 and ResNet101 models, 2,048-dimensional features were extracted per plane, yielding a comprehensive 4,096-dimensional dual-plane feature set. Conversely, ResNet18 generated 512 dimensions per plane, resulting in a 1,024-dimensional combined set.

## 2.6 ComBat Normalization Protocol

To mitigate center-related variation in extracted MRI features, ComBat harmonization was performed using the batch-effect removal tool on the Sangerbox platform [13]. The ComBat harmonization parameters were estimated using the Hospital 1 training cohort only (n = 111). After the parameters had been frozen, the same transformation was applied without re-estimation to the Hospital 1 internal validation cohort and the Hospital 2 external validation cohort. No Hospital 2 image, extracted feature, feature distribution, outcome label, or performance estimate contributed to the estimation of the ComBat parameters. Hospital 2 images were processed only after all data-dependent preprocessing and harmonization parameters had been frozen. Hospital 2 data were not used to re-estimate or update these parameters, and Hospital 2 outcome labels and performance estimates were not used for model-development decisions.

## 2.7 Sample Size and Missing Data

The sample size was determined by all eligible patients available during the study period. Patients with missing preoperative MRI, incomplete follow-up, unavailable outcome data, or suboptimal image quality were excluded. Complete-case analysis was used for candidate predictors, and no imputation was performed. The modest development and external validation samples were considered when interpreting model precision and stability.

## 2.8 Feature Set Construction for Modeling

After excluding all postoperative variables, 43 preoperative clinical and conventional imaging variables remained as candidate predictors. Continuous predictors were standardized using means and standard deviations estimated from the training cohort only, and these fixed values were subsequently applied unchanged to the internal validation cohort and the external validation cohort. Five feature sets were specified before feature selection: clinical variables, sagittal MRI features, transverse MRI features, dual-plane MRI features, and a combined set integrating clinical variables with dual-plane MRI features.

## 2.9 Feature Selection and Dimensionality Reduction

To reduce feature dimensionality and limit the risk of overfitting, all outcome-informed feature-selection procedures were restricted to the training cohort. Features were first screened using the Mann–Whitney U test ($P < 0.05$). Redundant features were then removed using Spearman correlation and greedy recursive deletion at an absolute correlation threshold of 0.90. For high-dimensional feature sets, minimum redundancy maximum relevance (mRMR) was used to retain the 40 highest-ranked features; this step was not applied to the clinical set or the single-plane ResNet18 sets because of their lower dimensionality. Least absolute shrinkage and selection operator (LASSO) regression was subsequently used to obtain the final feature subsets.

## 2.10 Model Construction and Validation

Models were developed in Python 3.9 using scikit-learn and open-source components of the OnekeyAI computational framework. Logistic regression (LR), multilayer perceptron (MLP), and support vector machine (SVM) classifiers were evaluated using clinical, single-plane deep learning (transverse and sagittal), dual-plane deep learning, and combined features integrating clinical variables with dual-plane deep learning features. Hyperparameters were optimized by grid search with 10-fold cross-validation within the training cohort. Candidate models were compared using the Hospital 1 internal validation cohort to select the final architecture and classifier. The decision threshold was selected using Hospital 1 data and fixed before external evaluation. After the model specification and decision threshold had been fixed, the final model was evaluated once in the Hospital 2 external validation cohort. Hospital 2 was not used for model selection, feature selection, hyperparameter tuning, threshold selection, or any other model-development decision.

## 2.11 Statistical Analysis

Categorical variables were compared using Fisher's exact test or the chi-square test, and continuous variables were compared using the Mann–Whitney U test or Student's t test, as appropriate. Statistical analyses were performed using Python 3.9. All tests were two-sided, and $P < 0.05$ was considered statistically significant. Decision-curve analysis and calibration curves were generated using R version 4.3.3.

# 3 Results

## 3.1 Participants

A total of 214 patients were included: 111 in the training cohort, 48 in the internal validation cohort, and 55 in the external testing cohort. Numbness improvement occurred in 65 (58.6%), 31 (64.6%), and 34 (61.8%) patients, respectively. Cohort characteristics are summarized in Table 1 (Excerpt); complete characteristics and group comparisons are provided in Supplementary Material 1, Section S7, Table S1.

Table 1
(Excerpt) Clinical indicators of different cohorts.

| features | Train (n = 111) | | | | Internal Validation Set (n = 48) | | | | External Validation Set (n = 55) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ALL | Non-Improved | improved | *p* | ALL | Non-Improved | improved | *p* | ALL | Non-Improved | improved |
| | | (n = 46) | (n = 65) | | | (n = 17) | (n = 31) | | | (n = 21) | (n = 34) |
| Pre JOA | 12.28 ± 3.01 | 12.22 ± 3.00 | 12.32 ± 3.04 | 0.501 | 12.38 ± 3.78 | 10.59 ± 3.20 | 13.35 ± 3.76 | 0.027 | 12.18 ± 3.63 | 11.48 ± 2.73 | 12.62 ± 4.06 |
| Surgery Time (h) | 2.95 ± 0.94 | 3.13 ± 1.15 | 2.83 ± 0.74 | 0.454 | 3.06 ± 0.86 | 3.41 ± 0.87 | 2.87 ± 0.81 | 0.035 | 2.75 ± 1.04 | 3.00 ± 1.18 | 2.59 ± 0.92 |
| Pre CTD (month) | 18.54 ± 35.38 | 22.67 ± 32.60 | 15.62 ± 37.19 | 0.060 | 23.62 ± 40.33 | 29.53 ± 45.37 | 20.39 ± 37.67 | 0.400 | 25.53 ± 39.44 | 37.24 ± 46.39 | 18.29 ± 33.16 |
| Hypertension | | | | 0.110 | | | | 0.660 | | | |
| No | 78(70.3) | 28(60.9) | 50(76.9) | | 37(77.1) | 12(70.6) | 25(80.7) | | 44(80.0) | 15(71.4) | 29(85.3) |
| Yes | 33(29.7) | 18(39.1) | 15(23.1) | | 11(22.9) | 5(29.4) | 6(19.4) | | 11(20.0) | 6(28.6) | 5(14.7) |
| Diabetes | | | | 0.590 | | | | 0.240 | | | |
| No | 93(83.8) | 37(80.4) | 56(86.2) | | 44(91.7) | 14(82.4) | 30(96.8) | | 50(90.9) | 17(81.0) | 33(97.1) |
| Yes | 18(16.2) | 9(19.6) | 9(13.9) | | 4(8.3) | 3(17.7) | 1(3.2) | | 5(9.1) | 4(19.1) | 1(2.9) |
| Osteoporosis | | | | 0.060 | | | | 0.150 | | | |
| No | 88(79.3) | 32(69.6) | 56(86.2) | | 38(79.2) | 11(64.7) | 27(87.1) | | 42(76.4) | 15(71.4) | 27(79.4) |
| Yes | 23(20.7) | 14(30.4) | 9(13.9) | | 10(20.8) | 6(35.3) | 4(12.9) | | 13(23.6) | 6(28.6) | 7(20.6) |
| No. of Prot. segments | | | | 0.017 | | | | 0.974 | | | |
| 1 | 0 | 0 | 0 | | 0 | 0 | 0 | | 0 | 0 | 0 |
| 2 | 1(0.9) | 1(2.2) | 0 | | 0 | 0 | 0 | | 1(1.8) | 1(4.8) | 0 |
| 3 | 2(1.8) | 2(4.4) | 0 | | 3(6.3) | 1(5.9) | 2(6.5) | | 4(7.3) | 3(14.3) | 1(2.9) |
| 4 | 76(68.5) | 25(54.4) | 51(78.5) | | 28(58.3) | 10(58.8) | 18(58.1) | | 25(45.5) | 7(33.3) | 18(52.9) |
| 5 | 29(26.1) | 15(32.6) | 14(21.5) | | 15(31.3) | 5(29.4) | 10(32.3) | | 23(41.8) | 9(42.9) | 14(41.2) |
| 6 | 3(2.7) | 3(6.5) | 0 | | 2(4.2) | 1(5.9) | 1(3.2) | | 2(3.6) | 1(4.8) | 1(2.9) |
| Sag. protrusion level | | | | 0.428 | | | | 0.037 | | | |
| -3 | 2(1.8) | 1(2.2) | 1(1.5) | | 0 | 0 | 0 | | 1(1.8) | 0 | 1(2.9) |
| -2 | 4(3.6) | 2(4.4) | 2(3.1) | | 3(6.3) | 3(17.7) | 0 | | 3(5.5) | 2(9.5) | 1(2.9) |
| -1 | 61(55.0) | 21(45.7) | 40(61.5) | | 28(58.3) | 8(47.1) | 20(64.5) | | 18(32.7) | 6(28.6) | 12(35.3) |
| 0 | 42(37.8) | 20(43.5) | 22(33.9) | | 15(31.3) | 4(23.5) | 11(35.5) | | 32(58.2) | 12(57.1) | 20(58.8) |
| 1 | 1(0.9) | 1(2.2) | 0 | | 1(2.1) | 1(5.9) | 0 | | 1(1.8) | 1(4.8) | 0 |
| 2 | 1(0.9) | 1(2.2) | 0 | | 1(2.1) | 1(5.9) | 0 | | 0 | 0 | 0 |
| Pfirrmann Grade | | | | 0.002 | | | | 0.008 | | | |
| 1 | 1(0.9) | 1(2.2) | 0 | | 0 | 0 | 0 | | 0 | 0 | 0 |
| 2 | 9(8.1) | 4(8.7) | 5(7.7) | | 3(6.3) | 0 | 3(9.7) | | 5(9.1) | 1(4.8) | 4(11.8) |
| 3 | 27(24.3) | 7(15.2) | 20(30.8) | | 11(22.9) | 5(29.4) | 6(19.4) | | 16(29.1) | 4(19.1) | 12(35.3) |
| 4 | 50(45.1) | 16(34.8) | 34(52.3) | | 25(52.1) | 5(29.4) | 20(64.5) | | 22(40.0) | 9(42.9) | 13(38.2) |
| 5 | 24(21.6) | 18(39.1) | 6(9.2) | | 9(18.8) | 7(41.2) | 2(6.5) | | 12(21.8) | 7(33.3) | 5(14.7) |

**Note: Continuous variables are expressed as mean ± SD, and categorical variables as n (%). Abbreviations: Pre = preoperative; JOA = Japanese Orthopaed Association score; NRS = Numerical Rating Scale; CTD = conservative treatment duration; No. = number; prot. = protruding; Sag. = sagittal;**

## 3.2 Clinical Model Construction

Six preoperative features—preoperative Japanese Orthopaedic Association (JOA) scores, hypertension, preoperative conservative treatment duration (months), diabetes, osteoporosis, and the number of protruding segments—were selected for clinical model construction. The detailed feature selection process and coefficients are presented in Supplementary Material 1, Section S8, **Fig. S1**. Among the three machine learning classifiers, the SVM demonstrated the highest predictive efficacy, achieving an AUC of 0.719 (95% CI: 0.567–0.870) in the internal validation cohort and 0.721 (95% CI: 0.568–0.874) in the external validation cohort (Fig. 2). Comprehensive performance metrics for all clinical classifiers are summarized in Supplementary Material 1, Section S9, Table S2.

## 3.3 Construction of Single-Plane Deep Learning Models

For the ResNet101 support vector machine, transverse features yielded AUCs of 0.793 (95% CI 0.650–0.936) in internal validation and 0.731 (0.585–0.877) in external testing. The corresponding sagittal-feature AUCs were 0.746 (0.603–0.889) and 0.723 (0.584–0.861) (Fig. 3). Feature-selection results are provided in Supplementary Material 1, Section S8, Figs. S2 and S3, and complete performance metrics in Supplementary Material 1, Section S9, Tables S3 and S4.

## 3.4 Dual-plane Deep Learning Model Construction

The ResNet101 support vector machine based on dual-plane MRI features yielded AUCs of 0.853 (95% CI 0.748–0.958) in internal validation and 0.789 (0.666–0.911) in external testing (Fig. 4). Feature-selection results for all three residual-network architectures are provided in Supplementary Material 1, Section S8, Fig. S4, and complete performance measures in Supplementary Material 1, Section S9, Table S5.

## 3.5 Construction and Evaluation of the Combined Model

The ResNet101 architecture was selected for constructing the combined model. Following feature selection (Fig. 5a-c), a total of 29 features were utilized for model development. The model achieved an AUC of 0.868 (95% CI, 0.769–0.966), specificity of 0.947, and sensitivity of 0.586 in the internal validation cohort. It achieved an AUC of 0.833 (95% CI, 0.720–0.947), specificity of 0.905, and sensitivity of 0.677 in the external validation cohort (Fig. 5 and Table 2). Furthermore, decision curve analysis (DCA) and calibration curves corroborate the reliability of these findings (Fig. 6). To further elucidate the model's decision-making process, Grad-CAM was employed to generate activation heatmaps (Supplementary Material 1, Section S10, **Fig. S5**). Visualization revealed that the model's focus primarily corresponded to key pathological regions, such as protruded discs and compressed nerve roots.

Table 2
Predictive performance of the combined model.

| Model | ML | Task | Acc | AUC | 95% CI | Sen | Spe | PPV | NPV |
|---|---|---|---|---|---|---|---|---|---|
| **ResNet101** | | | | | | | | | |
| | **LR** | **Training** | **0.847** | **0.927** | **0.881–0.973** | **0.791** | **0.932** | **0.946** | **0.746** |
| | **LR** | **Internal validation** | **0.750** | **0.824** | **0.709–0.940** | **0.724** | **0.790** | **0.840** | **0.652** |
| | **LR** | **External validation** | **0.727** | **0.699** | **0.551–0.847** | **0.765** | **0.667** | **0.788** | **0.636** |
| | **SVM** | **Training** | **0.928** | **0.990** | **0.978–1.000** | **0.896** | **0.977** | **0.984** | **0.860** |
| | **SVM** | **Internal validation** | **0.729** | **0.868** | **0.769–0.966** | **0.586** | **0.947** | **0.944** | **0.600** |
| | **SVM** | **External validation** | **0.764** | **0.833** | **0.720–0.947** | **0.677** | **0.905** | **0.920** | **0.633** |
| | **MLP** | **Training** | **0.874** | **0.950** | **0.913–0.987** | **0.866** | **0.886** | **0.921** | **0.813** |
| | **MLP** | **Internal validation** | **0.688** | **0.784** | **0.655–0.913** | **0.586** | **0.842** | **0.850** | **0.571** |
| | **MLP** | **External validation** | **0.746** | **0.804** | **0.679–0.929** | **0.706** | **0.810** | **0.857** | **0.630** |

**Note: Acc, accuracy; AUC, area under the receiver operating characteristic curve; LR, logistic regression; ML, machine learning; MLP, multilayer perceptron; NPV, negative predictive value; PPV, positive predictive value; Sen, sensitivity; Spe, specificity; SVM, support vector machine; 95% CI, 95% confidence interval.**

## 4 Discussion

We developed and preliminarily evaluated a combined clinical and dual-plane MRI model for predicting 12-month patient-reported lower-limb numbness improvement after TMD. The model achieved an AUC of 0.833 in the external validation cohort, with specificity of 0.905 and sensitivity of 0.677. Because the cohort was small and the estimates were uncertain, its numerically higher discrimination than the clinical-only and imaging-only models should be considered exploratory and confirmed in larger independent cohorts.

Previous lumbar spine prediction studies have mainly addressed overall surgical success, disability, or pain. Benítez-Andrades et al. integrated clinical, psychometric, socioeconomic, and laboratory variables [14]. ], whereas Berg et al. and Halicka et al. evaluated disability or pain outcomes using geographic or external validation [8, 9]. Our study focuses on postoperative numbness, which may recover differently from pain.

The dual-plane model showed numerically higher discrimination than either single-plane model. Transverse MRI depicts the relationship among the disc, spinal canal, and nerve roots, whereas sagittal MRI captures disc-height loss, migration, and longitudinal compression [15]. Consistent with Fan et al [16].

combining both planes may capture complementary spatial information, although their prediction target differed from ours.

The combined model retained hypertension, diabetes, osteoporosis, and the number of protruding segments alongside MRI features. Osteoporosis and more extensive disc involvement have been associated with less favorable outcomes after lumbar surgery [5, 17, 18]. These variables should be interpreted as prognostic associations rather than causal determinants. The modest clinical-only performance suggests that routine variables incompletely capture the anatomical and microstructural consequences of nerve-root compression [19].

Ten-fold cross-validation reduced dependence on a single training split but did not eliminate optimism from comparing multiple feature sets, architectures, and classifiers. Because the Hospital 1 internal validation cohort guided candidate-model comparison and final model selection, its performance reflects model-selection performance rather than independent generalization. The decline from training to internal and external validation further supports cautious interpretation.

Center-related variation is a recognized source of instability in MRI-derived predictors. Johnson et al. introduced ComBat as an empirical-Bayes approach for reducing systematic batch effects, and Fortin et al. demonstrated its application to harmonizing multisite diffusion-tensor imaging data [20, 21]. In the present study, the ComBat parameters were estimated using Hospital 1 data only. The fitted transformation was frozen before Hospital 2 data were processed and was applied to the external validation cohort without re-estimation. Therefore, no Hospital 2 image, extracted feature, or feature distribution contributed to the harmonization-parameter estimation, and Hospital 2 remained reserved for the prespecified external validation evaluation. Nevertheless, both institutions used 3.0-T MRI systems from the same manufacturer. The present results therefore address only a limited range of scanner variation.

External validation across heterogeneous imaging platforms provides a stronger assessment of transportability. For example, Nicolaes et al. externally validated a vertebral-fracture convolutional neural-network model using CT examinations from eight institutions, 16 scanners, and three manufacturers, thereby evaluating performance across substantially more diverse acquisition settings [22]. Future work should similarly evaluate the present model in genuinely independent multicenter cohorts that include different MRI vendors, field strengths, acquisition protocols, and clinical populations.

This study has several limitations. Its retrospective complete-case design may have introduced selection and information bias, because patients with missing MRI, follow-up, or outcome data were excluded and no imputation was performed. The training cohort was small relative to the high-dimensional MRI feature space, and the higher training discrimination relative to the internal and external estimates raises concern for overfitting and model-selection optimism. Although outcome-informed feature selection and hyperparameter optimization were restricted to the training process, the Hospital 1 internal validation cohort contributed to candidate-model comparison and final model selection; its performance should therefore not be interpreted as a fully independent estimate of generalization. The primary outcome was a categorical patient-reported assessment rather than a validated quantitative numbness instrument, and classification of any partial relief as improvement may have increased outcome heterogeneity. Although ROI delineation followed a relatively strict protocol and underwent senior review, MRI slice selection and cropping remained manual and operator dependent, introducing potential variability and limiting reproducibility. More automated and objective end-to-end approaches should therefore be considered. Both centers used MRI systems from the same manufacturer, limiting assessment of broader scanner heterogeneity. Calibration estimates were based on relatively small validation cohorts, and only the 12-month endpoint was examined. Grad-CAM maps provided qualitative visualization of the CNN feature extractor rather than a complete explanation of the combined SVM model. Future studies should establish a more robust model using nested resampling or bootstrap-based optimism correction and evaluate it in genuinely independent multicenter cohorts with heterogeneous MRI platforms, acquisition protocols, and validated quantitative sensory outcomes.

## 5 Conclusion

The model achieved an AUC of 0.833 in the external validation cohort, with sensitivity of 0.677 and specificity of 0.905. Given the small external validation cohort, the use of a categorical patient-reported outcome, and the limited scanner heterogeneity, these findings should be interpreted as preliminary model-development evidence rather than definitive evidence supporting individualized clinical prediction. Larger prospective, multicenter, and multivendor studies using validated sensory outcomes are required before clinical implementation can be considered.

## Declarations

### 6.1 Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

### 6.2 Code Availability

The open-source OnekeyAI components used as the computational framework for preprocessing, feature extraction, feature selection, and model construction are available at https://github.com/OnekeyAI-Platform/onekey/tree/main/onekey_comp. The repository provides a reusable computational framework and should not be interpreted as a validated, fully automated raw-DICOM-to-prediction system for the present clinical task.

### 6.3 Human Ethics and Consent to Participate declarations

This study was approved by the Ethics Committee of a provincial tertiary hospital in China (approval no. 2022KY026), with informed consent waived.

### 6.4 Funding

This work was funded by the Joint Funds for the Innovation of Science and Technology, Fujian Province (Grant No. 2021Y9074 to RW) and the Pingtan Comprehensive Experimental Area 2025 District-level Scientific and Technological Project. The funding sources had no role in the design of the study, collection, analysis, and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

**6.5 Consent for publication**

Not applicable

**6.6 Competing interests**

The authors declare no competing interests.

**6.7 Acknowledgments**

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

**6.8 Author contributions**

Conceptualization: YL, WZ, XW, KW, YW, GG, RW; Data curation: YL, WZ, XW, KW, YW, GG, RW; Formal analysis: YL, WZ, XW, KW; Funding acquisition: YW, GG, RW; Investigation: YL, WZ, XW, KW, YW, GG, RW; Methodology: YL, WZ, XW, KW, YW, GG, RW; Resources: YW, GG, RW; Software: YL, WZ, XW, KW; Supervision: YW, GG, RW; Validation: YW, GG, RW; Writing — original draft: YL, WZ, XW; Writing — review & editing: YL, WZ, XW, KW, YW, GG, RW.

## References

1. Pojskic M, Bisson E, Oertel J, Takami T, Zygourakis C, Costa F. Lumbar disc herniation: Epidemiology, clinical and radiologic diagnosis WFNS spine committee recommendations. World Neurosurg X. 2024;22:100279.
2. Kogl N, Petr O, Loscher W, Liljenqvist U, Thome C. Lumbar Disc Herniation-the Significance of Symptom Duration for the Indication for Surgery. Dtsch Arztebl Int. 2024;121:440–8.
3. Arts MP, Brand R, van den Akker ME, Koes BW, Bartels RH, Peul WC, et al. Tubular diskectomy vs conventional microdiskectomy for sciatica: a randomized controlled trial. JAMA. 2009;302:149–58.
4. Wang Y, Gao F, Zou H. Numbness and Weakness Recovered at a Less Extent in Patients with Lumbar Disc Herniation after Percutaneous Endoscopic Lumbar Discectomy. Pain Res Manag. 2019;2019:4642701.
5. Shi L, Ji X, Tian F, Shi Y, Lou P. Risk factor of residual leg numbness after lumbar microdiscectomy for lumbar disc herniation. Med (Baltim). 2023;102:e35733.
6. Huang P, Sengupta DK. How fast pain, numbness, and paresthesia resolves after lumbar nerve root decompression: a retrospective study of patient's self-reported computerized pain drawing. Spine (Phila Pa 1976). 2014;39:E529–536.
7. Ogura Y, Takahashi Y, Kitagawa T, Yonezawa Y, Yoshida K, Takeda K, et al. Impact of leg numbness on patient satisfaction following decompression surgery for lumbar spinal stenosis. J Clin Neurosci. 2021;93:112–5.
8. Berg B, Gorosito MA, Fjeld O, Haugerud H, Storheim K, Solberg TK, et al. Machine Learning Models for Predicting Disability and Pain Following Lumbar Disc Herniation Surgery. JAMA Netw Open. 2024;7:e2355024.
9. Halicka M, Wilby M, Duarte R, Brown C. Predicting patient-reported outcomes following lumbar spine surgery: development and external validation of multivariable prediction models. BMC Musculoskelet Disord. 2023;24:333.
10. Xu Y, Zheng S, Tian Q, Kou Z, Li W, Xie X et al. Deep Learning Model for Grading and Localization of Lumbar Disc Herniation on Magnetic Resonance Imaging. J Magn Reson Imaging 2024.
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453–7.
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods: a Korean translation. Ewha Med J. 2025;48:e48.
13. Shen W, Song Z, Zhong X, Huang M, Shen D, Gao P, et al. Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. Imeta. 2022;1:e36.
14. Benitez-Andrades JA, Prada-Garcia C, Ordas-Reyes N, Blanco ME, Merayo A, Serrano-Garcia A. Enhanced prediction of spine surgery outcomes using advanced machine learning techniques and oversampling methods. Health Inf Sci Syst. 2025;13:24.
15. Seo J, Lee JW. Magnetic Resonance Imaging Grading Systems for Central Canal and Neural Foraminal Stenoses of the Lumbar and Cervical Spines With a Focus on the Lee Grading System. Korean J Radiol. 2023;24:224–34.
16. Fan Z, Wu T, Wang Y, Jin Z, Wang T, Liu D. Deep-Learning-Based Radiomics to Predict Surgical Risk Factors for Lumbar Disc Herniation in Young Patients: A Multicenter Study. J Multidiscip Healthc. 2024;17:5831–51.
17. Lechtholz-Zey EA, Ayad M, Gettleman BS, Mills ES, Shelby H, Ton AT, et al. Systematic Review and Meta-Analysis of the Effect of Osteoporosis on Reoperation Rates and Complications after Surgical Management of Lumbar Degenerative Disease. J Bone Metab. 2024;31:114–31.
18. Siyaji ZK, Nguyen AQ, Federico VP, Zbeidi S, Zavras AG, Baker JD, et al. Comparison of clinical outcomes in patients undergoing one- and two-level minimally invasive lumbar microdiscectomy. J Neurosurg Spine. 2024;40:169–74.

19. Wu W, Liang J, Ru N, Zhou C, Chen J, Wu Y, et al. Microstructural Changes in Compressed Nerve Roots Are Consistent With Clinical Symptoms and Symptom Duration in Patients With Lumbar Disc Herniation. Spine (Phila Pa 1976). 2016;41:E661–6.
20. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
21. Fortin JP, Parker D, Tunc B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. NeuroImage. 2017;161:149–70.
22. Nicolaes J, Liu Y, Zhao Y, Huang P, Wang L, Yu A, et al. External validation of a convolutional neural network algorithm for opportunistically detecting vertebral fractures in routine CT scans. Osteoporos Int. 2024;35:143–52.

## Figures

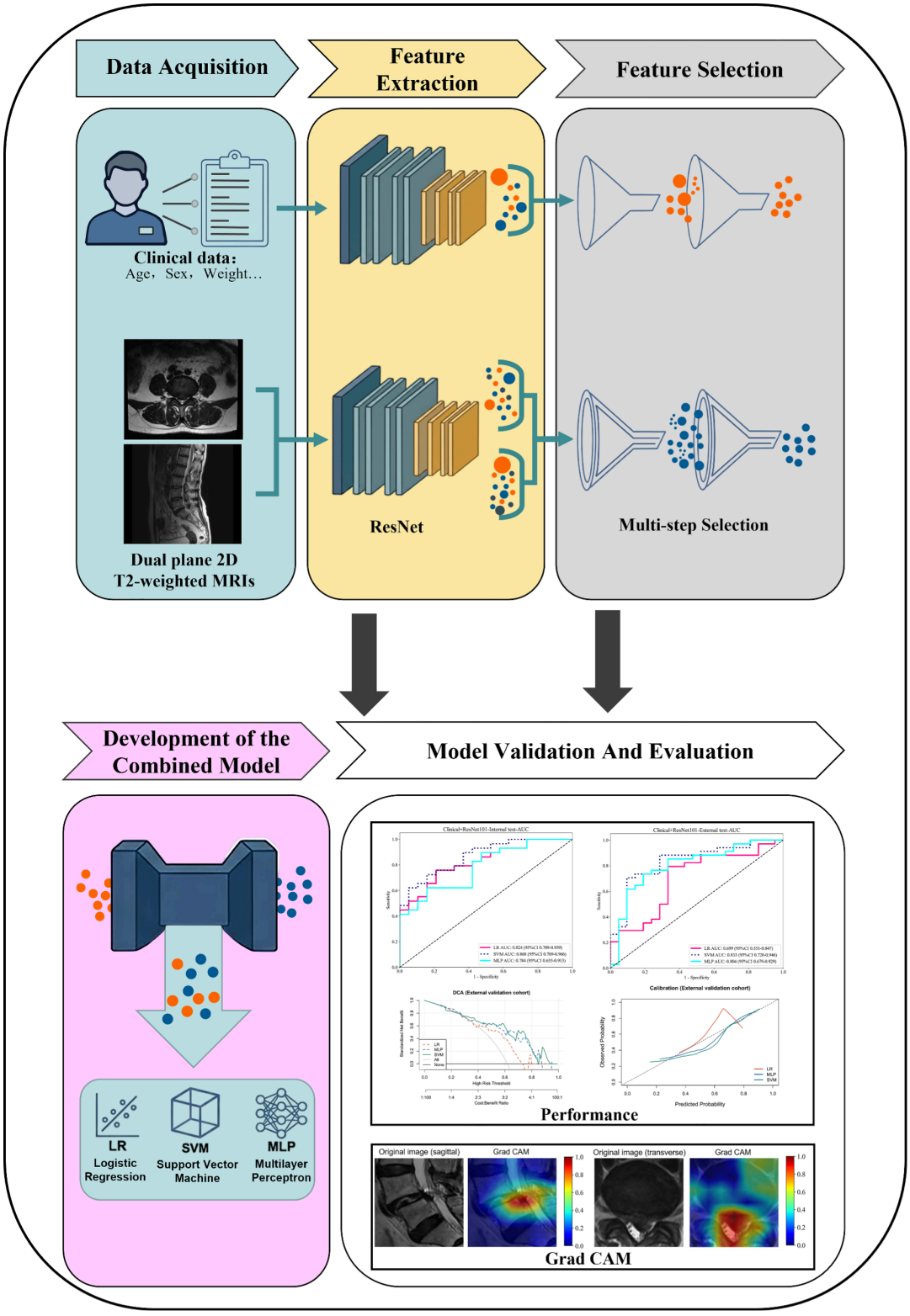


**Figure 1**

Study workflow showing cohort allocation, MRI preprocessing, feature extraction, feature selection, model development, external testing.

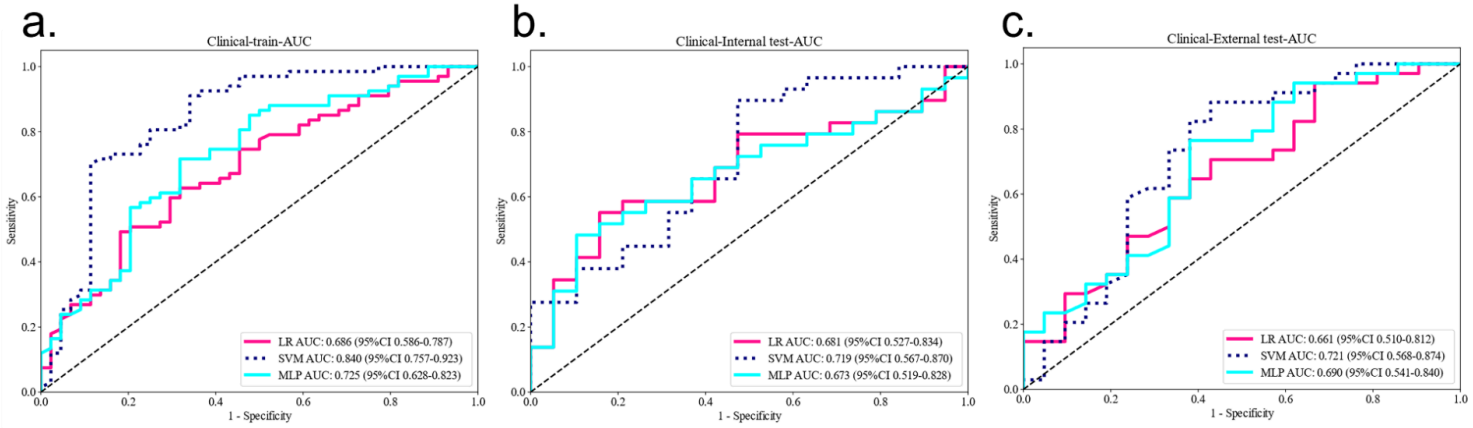



**Figure 2**

Receiver operating characteristic curves of clinical-feature models in the training (a), internal validation (b), and external testing (c) cohorts. LR, logistic regression; MLP, multilayer perceptron; SVM, support vector machine

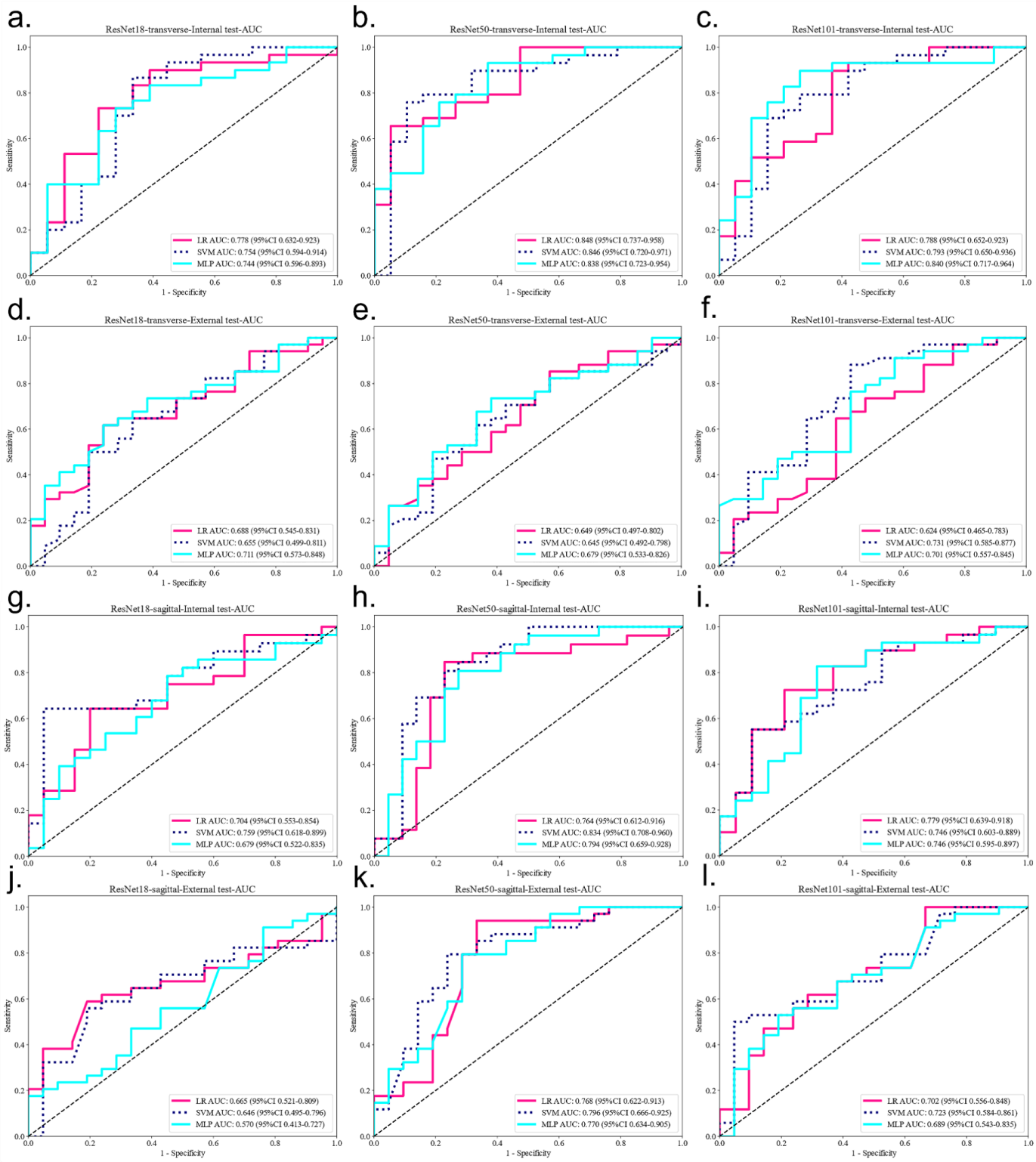



**Figure 3**

Receiver operating characteristic curves of single-plane models based on transverse (a–f) and sagittal (g–l) MRI features. Panels a–c and g–i show internal validation, and d–f and j–l show external testing; columns correspond to ResNet18, ResNet50, and ResNet101

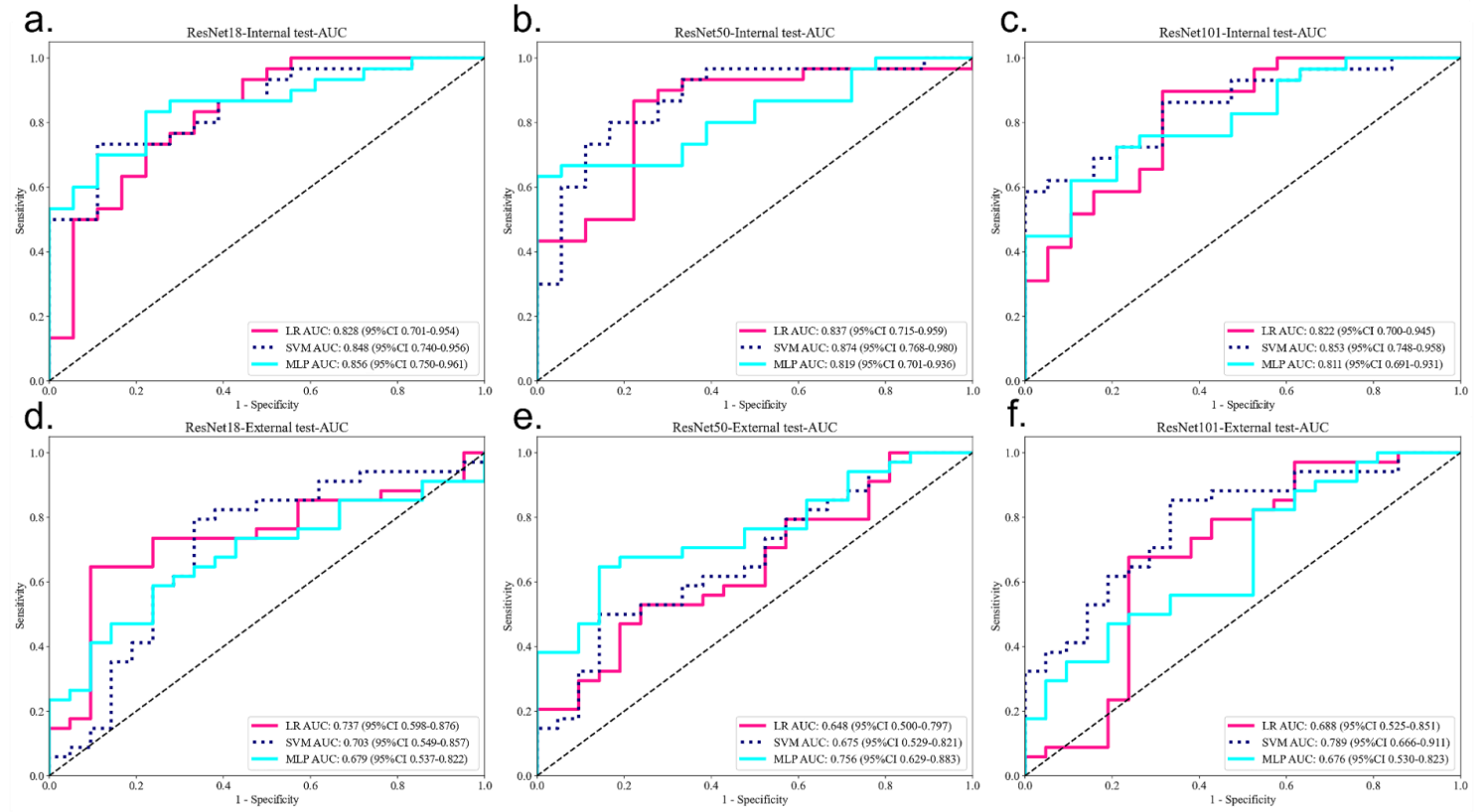



**Figure 4**

Receiver operating characteristic curves of dual-plane ResNet18 (a, d), ResNet50 (b, e), and ResNet101 (c, f) models in the internal validation (a–c) and external testing (d–f) cohorts

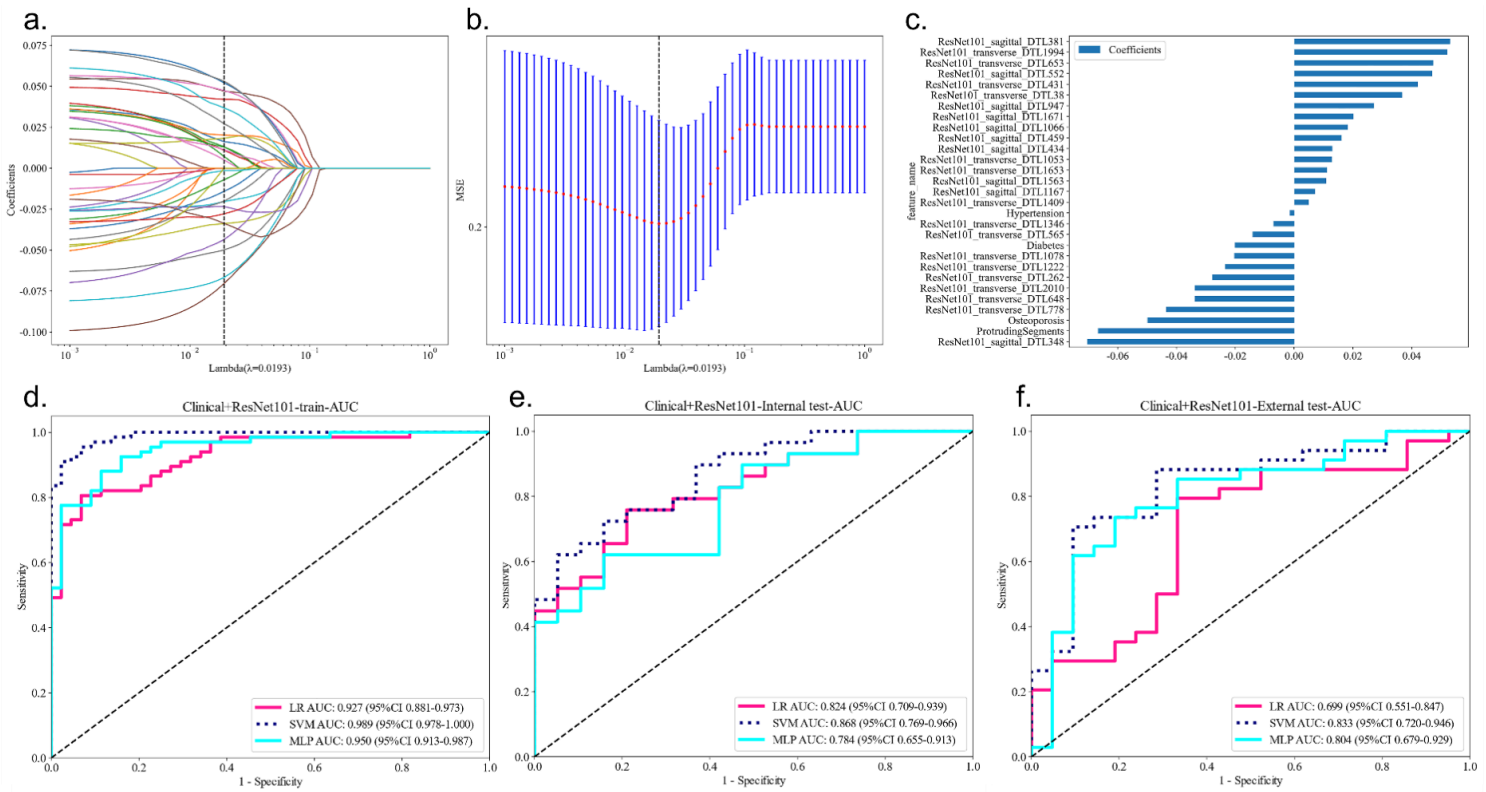



**Figure 5**

Feature selection and performance of the combined model. LASSO coefficient paths (a), cross-validated selection of λ (b), selected-feature weights (c), and receiver operating characteristic curves in the training (d), internal validation (e), and external testing (f) cohorts

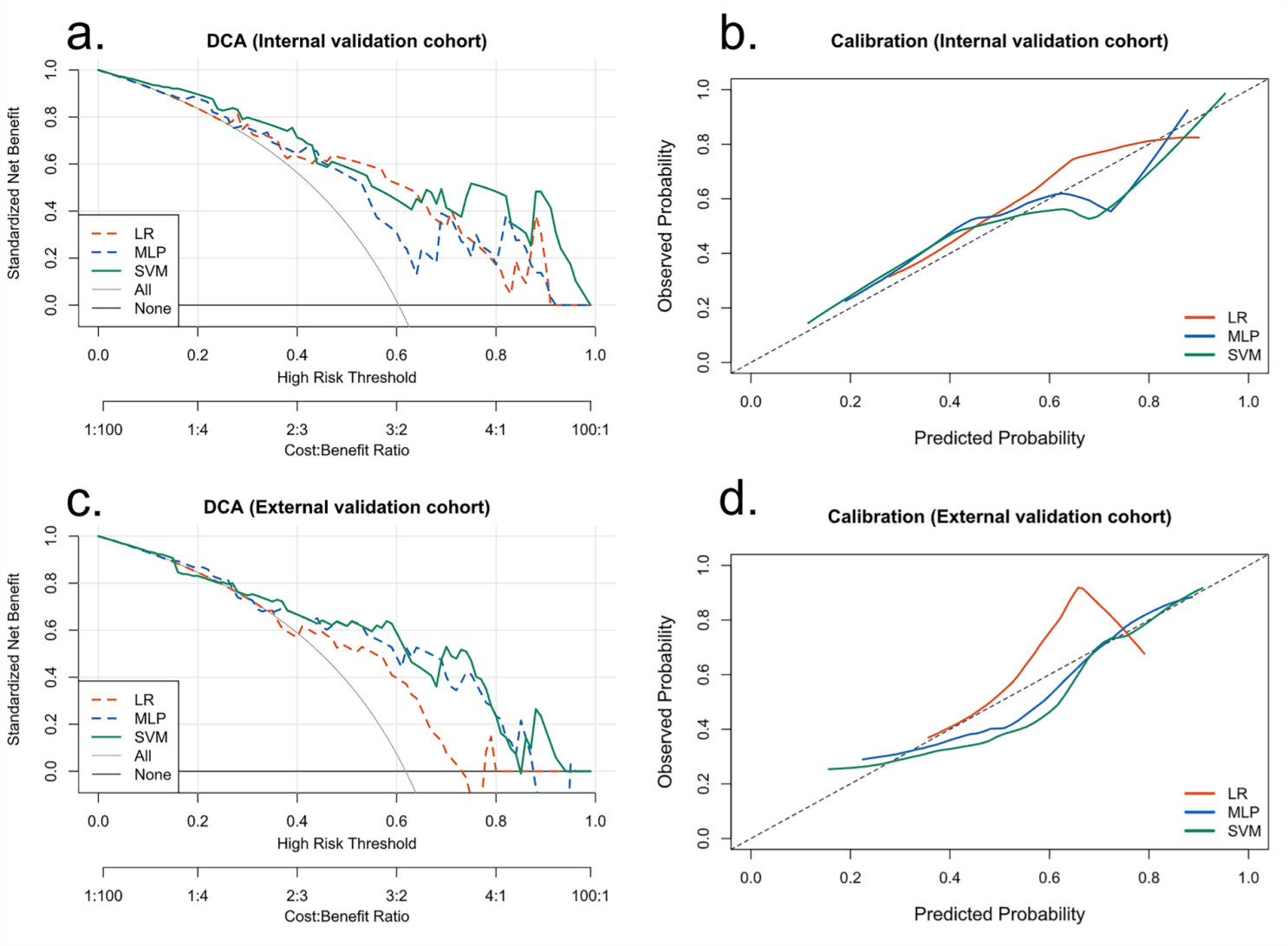



**Figure 6**

Decision-curve analyses (a, c), calibration curves (b, d), Panels a–b show internal validation and c–d show external testing

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx