

Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI

James Thomas Patrick Decourcy Hallinan, MBChB (Hons), BSc (Hons), FRCR* • Lei Zhu, BSc (Hons)* • Kaiyuan Yang, BSc (Hons) • Andrew Makmur, MBBS, BMedSci, MMed, FRCR • Diyaa Abdul Rauf Algazwi, MBBS • Yee Liang Thian, MBBS, FRCR • Samuel Lau, MBBS • Yun Song Choo, MBBS, FRCR • Sterling Ellis Eide, MBBS, FRCR • Qai Ven Yap, BSc (Hons) • Yiong Huak Chan, PhD • Jiong Hao Tan, MBBS, MRCS, MMed (Orth), FRCS (Orth) • Naresh Kumar, MBBS, MS Orth, DNB Orth, FRCS Ed, FRCS, FRCS (Orth & Trauma), DM (Orth Spinal Surgery) • Beng Chin Ooi, PhD • Hiroshi Yoshioka, MD, PhD • Swee Tian Quek, MBBS, FRCR




From the Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore 119074 (J.T.P.D.H., A.M., Y.L.T., S.L., Y.S.C., S.E.E., S.T.Q.); Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore (J.T.P.D.H., A.M., Y.L.T., S.L., Y.S.C., S.E.E., S.T.Q.); NUS Graduate School, Integrative Sciences and Engineering Programme, National University of Singapore, Singapore (L.Z.); Department of Computer Science, School of Computing, National University of Singapore, Singapore (K.Y., B.C.O.); Department of Radiology, Dammam Medical Complex, Dammam, Saudi Arabia (D.A.R.A.); Biostatistics Unit, Yong Loo Lin School of Medicine, Singapore (Q.V.Y., Y.H.C.); University Spine Centre, Department of Orthopaedic Surgery, National University Health System, Singapore (J.H.T., N.K.); and Department of Radiological Sciences, University of California, Irvine, Orange, Calif (H.Y.). Received November 17, 2020; revision requested January 11, 2021; revision received March 2; accepted March 16. Address correspondence to J.T.P.D.H. (e-mail: james_hallinan@nuhs.edu.sg).

J.T.P.D.H., L.Z., and A.M. supported by an NUHS AI Seed Fund competitive grant (NUHSRO/2018/087/RO5+5/AISeed/02).

* J.T.P.D.H. and L.Z. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the editorial by Hayashi in this issue.

Radiology 2021; 300:130–138 • <https://doi.org/10.1148/radiol.2021204289> • Content codes:   

Background: Assessment of lumbar spinal stenosis at MRI is repetitive and time consuming. Deep learning (DL) could improve productivity and the consistency of reporting.

Purpose: To develop a DL model for automated detection and classification of lumbar central canal, lateral recess, and neural foraminal stenosis.

Materials and Methods: In this retrospective study, lumbar spine MRI scans obtained from September 2015 to September 2018 were included. Studies of patients with spinal instrumentation or studies with suboptimal image quality, as well as postgadolinium studies and studies of patients with scoliosis, were excluded. Axial T2-weighted and sagittal T1-weighted images were used. Studies were split into an internal training set (80%), validation set (9%), and test set (11%). Training data were labeled by four radiologists using predefined gradings (normal, mild, moderate, and severe). A two-component DL model was developed. First, a convolutional neural network (CNN) was trained to detect the region of interest (ROI), with a second CNN for classification. An internal test set was labeled by a musculoskeletal radiologist with 31 years of experience (reference standard) and two subspecialist radiologists (radiologist 1: A.M., 5 years of experience; radiologist 2: J.T.P.D.H., 9 years of experience). DL model performance on an external test set was evaluated. Detection recall (in percentage), interrater agreement (Gwet κ), sensitivity, and specificity were calculated.

Results: Overall, 446 MRI lumbar spine studies were analyzed (446 patients; mean age \pm standard deviation, 52 years \pm 19; 240 women), with 396 patients in the training (80%) and validation (9%) sets and 50 (11%) in the internal test set. For internal testing, DL model and radiologist central canal recall were greater than 99%, with reduced neural foramina recall for the DL model (84.5%) and radiologist 1 (83.9%) compared with radiologist 2 (97.1%) ($P < .001$). For internal testing, dichotomous classification (normal or mild vs moderate or severe) showed almost-perfect agreement for both radiologists and the DL model, with respective κ values of 0.98, 0.98, and 0.96 for the central canal; 0.92, 0.95, and 0.92 for lateral recesses; and 0.94, 0.95, and 0.89 for neural foramina ($P < .001$). External testing with 100 MRI scans of lumbar spines showed almost perfect agreement for the DL model for dichotomous classification of all ROIs (κ , 0.95–0.96; $P < .001$).

Conclusion: A deep learning model showed comparable agreement with subspecialist radiologists for detection and classification of central canal and lateral recess stenosis, with slightly lower agreement for neural foraminal stenosis at lumbar spine MRI.

© RSNA, 2021

Online supplemental material is available for this article.

Lumbar spinal stenosis (LSS) is a potentially debilitating condition affecting more than 200 000 adults in the United States, with considerable impact on livelihood. Most patients with LSS present with lower back pain, which is also the main reason for seeking care (1). A large proportion of patients eventually

undergo lumbar spine MRI for diagnosis and treatment planning.

Lumbar spine MRI is an essential tool in the assessment of LSS for accurate evaluation of the central canal, lateral recesses, and neural foramina. The degree of stenosis at each region plays a role in determining the

Abbreviations

CNN = convolutional neural network, DL = deep learning, LSS = lumbar spinal stenosis, ROI = region of interest

Summary

A deep learning model detected and classified lumbar spinal stenosis at MRI comparable with radiologists for central canal and lateral recess stenosis, with slightly lower agreement for neural foraminal stenosis.

Key Results

- Compared with the reference radiologist, a deep learning (DL) model for detecting and classifying lumbar spinal stenosis at MRI showed high agreement ($\kappa = 0.92$ and 0.96 , respectively; $P < .001$) for dichotomous classification (normal or mild vs moderate or severe) of lateral recess and central canal stenosis in 50 test cases, similar to that of subspecialist radiologists ($\kappa = 0.92$ – 0.98 , $P < .001$).
- The DL model showed high agreement for dichotomous classification of neural foraminal stenosis ($\kappa = 0.89$, $P < .001$) in 50 test cases, slightly lower compared with subspecialist radiologists ($\kappa = 0.94$ and 0.95 , $P < .001$).

appropriate treatment, but detailing such information in a report can be repetitive and time consuming. In addition, there are multiple grading systems for LSS, with a lack of standardization (2–5).

Several recent deep learning (DL) systems have been created to assist in the interpretation of advanced imaging modalities, including knee MRI and CT angiography for cerebral aneurysms (6,7). Prior DL in lumbar spine MRI has shown promise, especially with the use of convolutional neural networks (CNNs) that can automatically learn representative features from images to perform classification tasks.

Automated grading of LSS at MRI using CNNs may improve the consistency, accuracy, and objectivity of assessment. DL in lumbar spine MRI has been mainly limited to counting of vertebra or grading disk degeneration (8,9). Most recently, in 2017, Jamaludin et al (10) developed SpineNet, a multitask architecture for automated classification of lumbar central canal stenosis and other spinal conditions. For central canal stenosis, there was a binary classification (present or absent), and only sagittal MRI was used. In 2018, another group, Lu et al (11), developed a DL algorithm (Deep Spine) to grade LSS at the central canal and neural foramina. This relied on weakly supervised natural language processing labels from existing radiology reports with no predefined grading system.

Currently, to our knowledge, no DL model has been developed to assess stenosis at all three regions of interest (ROIs) along the lumbar spine. Such a model could provide the reporting radiologist with an accurate and reliable diagnostic tool for LSS (12). In this study, a DL model was developed to automatically detect and classify central canal, lateral recess, and neural foraminal stenosis in the lumbar spine using axial and sagittal MRI sequences. Once the DL model was trained, its performance was compared with that of subspecialty radiologists on an internal test set. The DL model's performance and generalizability was also assessed on an external test set.

Materials and Methods

This study was approved by our institutional review board and compliant with the Health Insurance Portability and Accountability Act. A waiver of consent was granted due to the retrospective nature of the study and minimal risk involved.

Data Set Preparation

Retrospective, manual extraction and anonymization of lumbar spine MRI scans was done over a 3-year period from September 2015 to September 2018 at the National University Hospital, Singapore. Adult patients (>18 years) were included, with random selection of MRI lumbar spine studies across different MRI scanners (GE and Siemens 1.5- and 3.0-T platforms). Patients with instrumentation, severe scoliosis, suboptimal image quality, or postgadolinium studies were excluded. Axial T2-weighted and sagittal T1-weighted Digital Imaging and Communications in Medicine images were used. Table E1 (online) provides details on the MRI scanners and sequences.

Scans from the internal data set were randomly split 80%, 9%, and 11% for the training, validation, and test sets, respectively. This is an acceptable split for DL data sets (13).

A data set of MRI lumbar spine studies was also obtained for external testing from Dammam Medical Complex, Saudi Arabia (Philips and GE 1.5- and 3.0-T MRI platforms). The inclusion and exclusion criteria were identical to those of the internal data set. A random selection of 100 MRI lumbar spine studies was available for external testing over a 1-year period from September 2017 to September 2018, encompassing axial T2-weighted and sagittal T1-weighted Digital Imaging and Communications in Medicine images. No further training was performed on this data set.

Data Set Labeling

Training data were manually labeled by four board-certified radiologists with an interest in musculoskeletal radiology (J.T.P.D.H., 9 years of experience; S.E.E., 7 years of experience) and neuroradiology (A.M., 5 years of experience; Y.S.C., 2 years of experience). Each radiologist labeled at least 100 MRI lumbar spine studies independently and was blinded to patient demographics and clinical history. With use of an open-source annotation software, bounding boxes were drawn to segment the ROIs (central canal, lateral recesses, and neural foramina) at and between each lumbar disk level (14). When drawing each bounding box, the annotating radiologist classified the stenosis as one of four categories: normal, mild, moderate, or severe. This was done using well-established criteria for LSS at the central canal (15–17), lateral recesses (18), and neural foramina (15). A visual scale was provided to all annotating radiologists (Fig E1 [online]).

The internal test set was labeled using the same visual scale by an expert musculoskeletal radiologist with 31 years of experience (H.Y.) and served as the reference standard. The expert relabeled the same test set after a 4-week interval to assess intraobserver variability. For comparison with the DL model and to assess interobserver variability, the internal test set was also labeled by a subspecialist neuroradiologist (radiologist 1, A.M.) and a musculoskeletal radiologist (radiologist 2, J.T.P.D.H.).

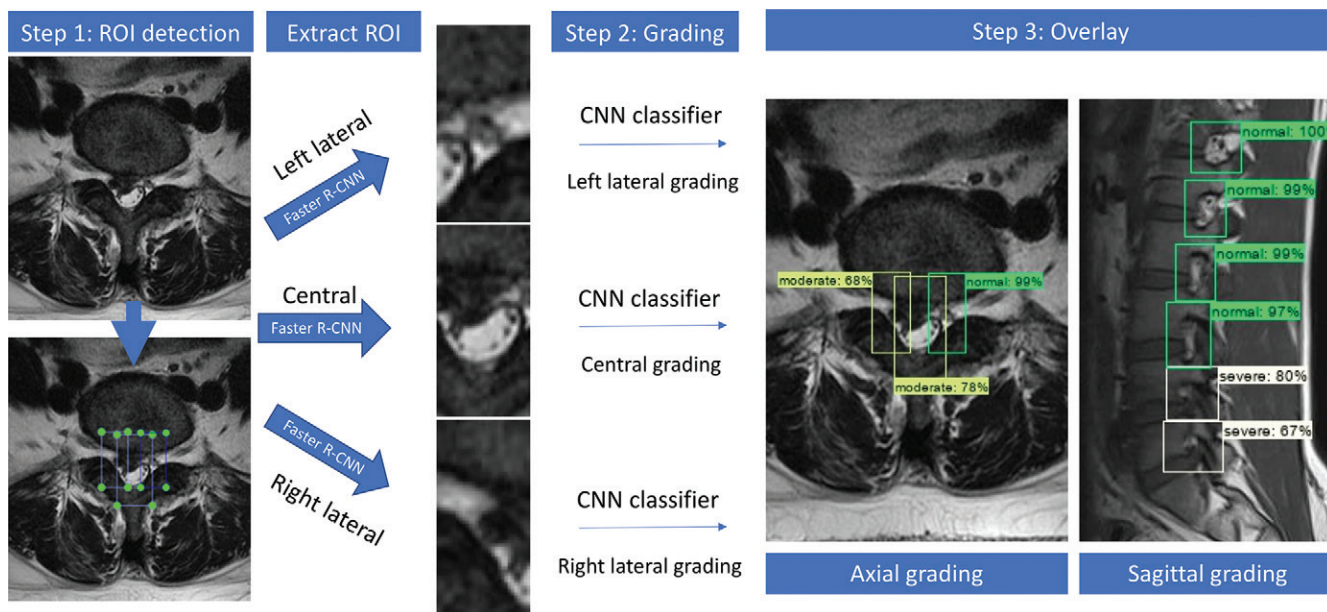


Figure 1: Diagram shows two-component deep learning (DL) model development. First, the model performed region of interest (ROI) detection (central canal and lateral recesses shown) on axial T2-weighted (repetition time msec/echo time msec, 3500/80) images, followed by classification of each ROI into normal, mild, moderate, or severe stenosis. A two-component DL model was also used for neural foraminal stenosis detection and classification on sagittal T1-weighted (500/15) images. The DL model predictions are completely transparent to the user and appear as color-coded bounding boxes at the ROIs along the lumbar spine. Percentage scores next to the grading represent the probability of the predicted class. CNN = convolutional neural network.

In a final step, the external test set was labeled by the subspecialist musculoskeletal radiologist (radiologist 2) to serve as the reference standard.

DL Model Development

A two-component procedure was used to train the models (Fig 1). First, Faster R-CNN (19) with Resnet101 (20) architectures were trained to detect the ROIs (central canal, lateral recesses, and neural foramina). Second, a CNN was trained for ROI classification. The CNN for classification (normal, mild, moderate, or severe) used six convolutional layers, each with a rectified linear unit, or ReLU, activation layer; a batch-normalization layer; and a max-pooling layer. This CNN ended with two fully connected layers, with the last layer using softmax activation. To train the classifier, a weighted categorical loss function was used for multiple classes (Fig E2 [online]). The DL model (SpineAI@NUHS-NUS) code is available at <https://github.com/NUHS-NUS-SpineAI/SpineAI-Detect-Classify-LumbarMRI-Stenosis>.

A recent study used whole axial MRI studies to train the DL CNN models (11). Due to the high capacity of these models, use of the entire image may result in the model learning irrelevant features from the background with poor performance in a real setting. To tackle this issue, we developed an ROI detection model to ensure only features from the relevant region were learned. Once the DL model was trained, its predictions on the test sets were visualized by the radiologist as bounding boxes at the ROIs (Fig 1). To visualize model interpretations, the Integrated Gradients technique was applied to the CNNs to highlight and/or color-code regions of the image (commonly referred to as heatmaps) that predict the classification result (21).

Statistical Analysis

The levels of agreement between the DL model and reference standard for ROI detection and classification were the primary outcome measures. ROI detection was assessed using intersection over union for the bounding boxes with a threshold greater than 0.2–0.4. Recall (in percentage) was used for measuring detection performance, as a high recall ensures the least number of missed ROI bounding boxes.

All analyses were performed using Stata version 16 (StataCorp), with statistically significant difference set at two-sided $P < .05$. Intrarater and interrater agreements using four-category and dichotomous (normal or mild vs moderate or severe) classification were assessed using Gwet κ to account for the paradox effect of a high percentage of normal classification (22,23). Sensitivity and specificity were presented for dichotomous classification.

Levels of agreement were defined for Gwet κ , as follows: less than 0, poor; 0–0.2, slight; 0.21–0.4, fair; 0.41–0.6, moderate; 0.61–0.8, substantial; and 0.81–1, almost-perfect agreement (24). In addition, 95% CIs were calculated.

Results

Patient Characteristics in Data Sets

Of 500 patients, 54 with instrumentation ($n = 15$), severe scoliosis ($n = 10$), suboptimal image quality ($n = 10$), or gadolinium-enhanced images ($n = 19$) were excluded. A total of 446 patients encompassing 12403 axial T2-weighted and 6161 sagittal T1-weighted Digital Imaging and Communications in Medicine images were evaluated. The patient group comprised 206 men (mean age \pm standard deviation, 47 years \pm 20; age range, 18–

Table 1: Patient Demographic and Clinical Characteristics

Characteristic	All Internal Data Sets (<i>n</i> = 446)	Internal Test Set (<i>n</i> = 50)	External Test Set (<i>n</i> = 100)
Age (y)*	52 ± 19 (18–93)	50 ± 20 (18–84)	53 ± 12 (26–75)
Women	240 (54)	27 (54)	46 (46)
Ethnicity			
Chinese	313 (70)	37 (74)	0
Malay	80 (18)	6 (12)	0
Indian	31 (7)	5 (10)	2 (2)
Other	22 (5)	2 (4)	98 (98)
Indication for MRI of the lumbar spine			
Back pain only	161 (36)	14 (28)	46 (46)
Back pain and unilateral radiculopathy	169 (38)	21 (42)	22 (22)
Back pain and bilateral radiculopathy	27 (6)	5 (10)	10 (10)
Other symptoms	89 (20)	10 (20)	22 (22)
Comorbidities			
Hypertension	68 (15)	10 (20)	15 (15)
Hyperlipidemia	27 (6)	2 (4)	8 (8)
Diabetes	18 (4)	5 (10)	14 (14)
Renal disease	40 (9)	3 (6)	5 (5)

Note.—Unless otherwise stated, data are numbers of patients, with percentages in parentheses.

* Data are means ± standard deviations, with ranges in parentheses.

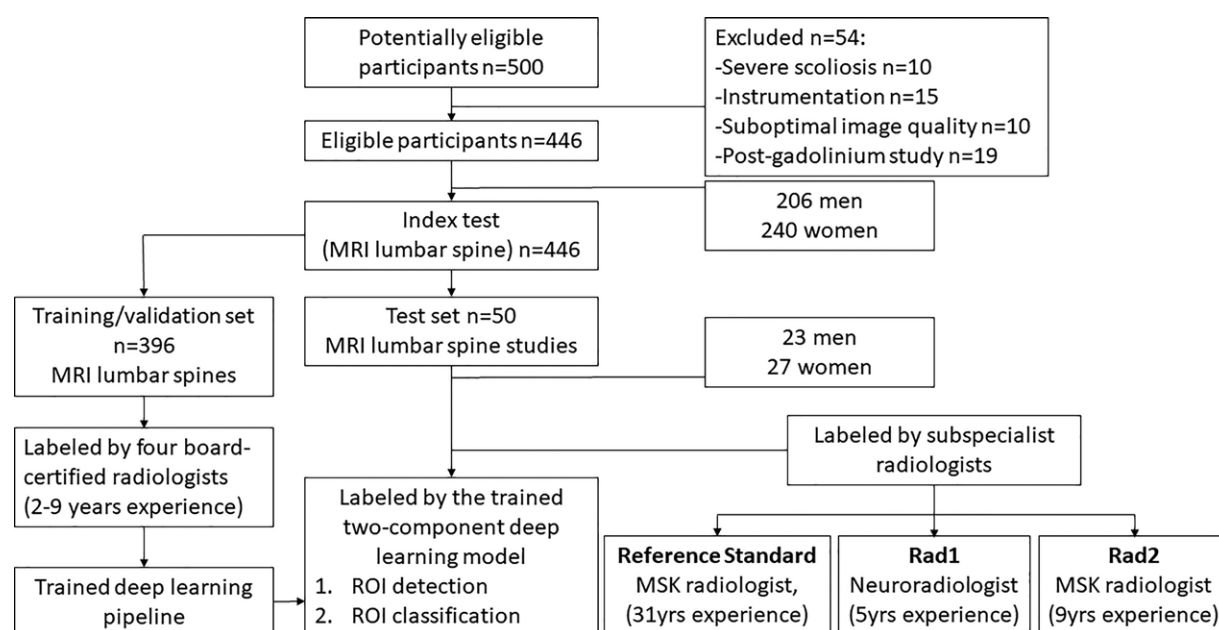


Figure 2: Flowchart of the study design for the internal training, validation, and test sets. The deep learning model's performance was compared with those of an expert musculoskeletal (MSK) radiologist (reference standard) and two subspecialist radiologists. Rad1 = radiologist 1, Rad2 = radiologist 2, ROI = region of interest.

91 years) and 240 women (mean age, 57 years ± 17; age range, 18–93 years). Overall, the mean age of all 446 patients was 52 years ± 19 (age range, 18–93 years) (Table 1).

The internal data set of 446 patients was randomly split into 396 patients for training (80%) and validation (9%) and 50 (11%) patients for testing. The internal test set comprised 23 men (mean age, 46 years ± 21; age range, 18–82 years) and 27 women (mean age, 53 years ± 19; age range, 18–84 years).

A flowchart of the internal data set study design is provided in Figure 2.

A random selection of 100 MRI lumbar spine studies was available for external testing. The patient group comprised 54 men (mean age, 50 years ± 11; age range, 31–75 years) and 46 women (mean age, 55 years ± 12; age range, 26–70 years). The mean age of all 100 patients was 53 years ± 12 (age range, 26–75 years).

Table 2: Reference Standard Bounding Boxes according to Stenosis Severity at Each Region of Interest

Test Set and Stenosis Severity	Central Canal	Lateral Recesses	Neural Foramina
Internal test set			
Normal	1326 (81.6)	514 (61.3)	903 (71.6)
Mild	243 (15.0)	193 (23)	192 (15.2)
Moderate	33 (2.0)	65 (7.7)	87 (6.9)
Severe	23 (1.4)	67 (8.0)	80 (6.3)
Total	1625	839	1262
Radiologist 1: total*	1618 (99.6)	827 (98.6)	1059 (83.9)
Radiologist 2: total*	1624 (99.9)	800 (95.4)	1225 (97.1)
DL model: total*	1624 (99.9)	799 (95.2)	1066 (84.5)
External test set			
Normal	1307 (80.6)	1552 (83.3)	1509 (79.5)
Mild	257 (15.8)	223 (12.0)	278 (14.6)
Moderate	39 (2.4)	59 (3.2)	80 (4.2)
Severe	19 (1.2)	29 (1.6)	31 (1.6)
Total	1622	1863	1898
DL model: total*	1617 (99.7)	1850 (99.3)	1827 (96.2)

Note.—Unless otherwise specified, data are numbers of bounding boxes, with percentage of the total in parentheses. Total bounding box detection numbers and recall are shown for radiologist 1, radiologist 2, and the DL model versus the reference standard. DL = deep learning.

* Data in parentheses are the recall (in percentages).

Reference Standard

The number of reference standard bounding boxes (labels) and classifications in the internal and external test sets are detailed in Table 2. There was a predominance of normal or mild classifications for all ROIs. For the central canal, moderate or severe classification accounted for 3.4% (56 of 1625) and 3.6% (58 of 1622) of labels in the internal and external test sets, respectively. In the internal test set, moderate or severe classification at the lateral recesses and neural foramina accounted for 15.7% (132 of 839) and 13.2% (167 of 1262) of labels, respectively. In comparison, for the external test set, moderate or severe classification at the lateral recesses and neural foramina accounted for 4.7% (88 of 1863) and 5.8% (111 of 1898) of labels, respectively.

Intrarater agreement was high for the internal test set across the ROIs. For the central canal, the κ values were 0.97 (95% CI: 0.96, 0.98; $P < .001$) for four-category gradings and 0.99 (95% CI: 0.99, 1.0; $P < .001$) for dichotomous gradings. For lateral recesses and neural foramina, the κ values were 0.93 (95% CI: 0.91, 0.95; $P < .001$) and 0.96 (95% CI: 0.95, 0.97; $P < .001$), respectively, for four-category gradings and 0.97 (95% CI: 0.96, 0.99; $P < .001$) and 0.98 (95% CI: 0.97, 0.99; $P < .001$) for dichotomous gradings.

ROI Detection

There was high recall for detecting the correct ROI by the radiologists and DL model (Table 2). For the central canal, bounding boxes drawn by the radiologists and DL model showed greater than 99% recall for the internal test set (radiologist 1, 1618 of 1625 labels; radiologist 2, 1624 of 1625 labels; DL model, 1624 of 1625 labels). For the lateral recesses, recall for radiologist 1 was 98.6% (827 of 839 labels), which was higher ($P < .001$) than the recall of 95.4% (800 of 839 labels) for radiologist 2 and

95.2% (799 of 839 labels) for the DL model. For the neural foramina, recall for radiologist 2 was 97.1% (1225 of 1262 labels), which was higher ($P < .001$) compared with the recall of 83.9% (1059 of 1262 labels) for radiologist 1 and 84.5% (1066 of 1262 labels) for the DL model. On the external test set, the recall of the DL model was 99.7% (1617 of 1622 labels), 99.3% (1850 of 1863 labels), and 96.2% (1827 of 1898 labels) for central canal, lateral recesses, and neural foramina, respectively. The high recall for the central canal was expected, as every axial image should have a bounding box. Reduced recall for the DL model on the lateral recesses and neural foramina is likely due to difficulty in determining the exact borders of the ROI (eg, disagreement on a bounding box at the medial or lateral borders of the neural foramen). Examples of DL model ROI detection and classification are provided for axial (Fig 3) and sagittal (Fig 4) images.

Internal Test Set ROI Classification

To validate the robustness of the DL model classifier, we randomly split the whole training and validation set into training and validation subsets three times. For each split, we repeated the model training three times, with the validation subset used for model selection. Finally, we produced nine versions of the classifier (DL models version 1–9) and calculated an average DL model.

For central canal classification, radiologists 1 and 2 had similar agreement, with κ values of 0.89 (95% CI: 0.87, 0.91; $P < .001$) and 0.89 (95% CI: 0.87, 0.90; $P < .001$), respectively (Table 3). The average DL model κ value was 0.82 (95% CI: 0.80, 0.85; $P < .001$). For dichotomous classification, both radiologists and the DL model showed almost perfect agreement, with κ values of 0.98 (95% CI: 0.97, 0.99; $P < .001$) for radiologist 1, 0.98 (95% CI: 0.98, 0.99; $P < .001$) for radiologist

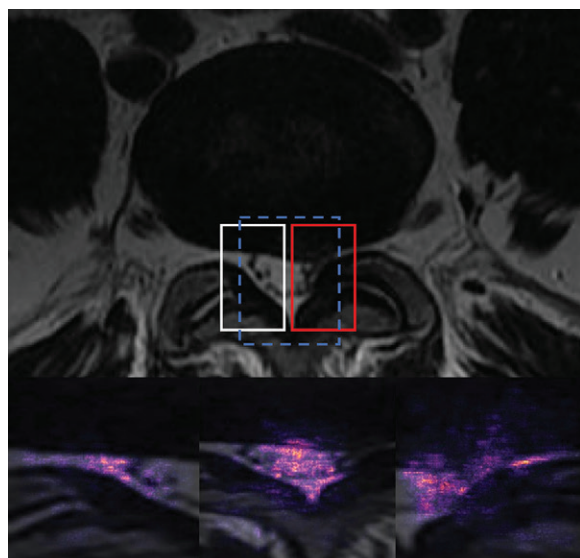


Figure 3: Axial T2-weighted (repetition time msec/echo time msec, 5300/100) MRI scan (top) at the level of L4-L5 intervertebral disk in a 60-year-old man with back pain and left lower limb radiculopathy. Integrated Gradients technique was applied to the deep learning (DL) model and highlights the MRI regions responsible for the prediction as a color-coded overlay, or heatmap (bottom row). In general, the higher the intensity value, the more important the feature for prediction. There is a normal right lateral recess (white box, left heatmap), mild central canal stenosis (blue dashed box, center heatmap), and severe left lateral recess stenosis (red box, right heatmap). Heatmaps show that activation of the DL model (red color) is mainly at the high-signal-intensity regions: cerebrospinal fluid in the central canal and surrounding epidural fat.

2, and 0.96 (95% CI: 0.94, 0.98; $P < .001$) for the average DL model (Table 4). For dichotomous classification, sensitivities for the average DL model (90.9%; 95% CI: 80.3, 96.7) and radiologist 1 (91.1%; 95% CI: 80.4, 97.0) were higher than that for radiologist 2 (76.8%; 95% CI: 63.6, 87.0; $P = .043$ and $P = .039$, respectively) (Table E2 [online]). High specificities were seen for the DL model and both radiologists, with radiologist 1 (98.5%; 95% CI: 97.7, 99.0) and radiologist 2 (99.2%; 95% CI: 98.6, 99.6) having higher specificities compared with the average DL model (97.0%; 95% CI: 96.0, 97.8; $P = .005$ and $P < .001$, respectively).

For lateral recess classification, the κ values for the radiologists and DL model were lower compared with the central canal and neural foramina classifications. Radiologist 2 had the highest κ value, with 0.79 (95% CI: 0.76, 0.82; $P < .001$). The average DL model had a κ value of 0.72 (95% CI: 0.68, 0.76; $P < .001$), similar to that of radiologist 1 ($\kappa = 0.71$; 95% CI: 0.68, 0.75; $P < .001$). For dichotomous classification, radiologist 2 had the highest κ value, with 0.95 (95% CI: 0.93, 0.97; $P < .001$). Similar agreement was seen for radiologist 1 ($\kappa = 0.92$; 95% CI: 0.90, 0.94; $P < .001$) and the average DL model ($\kappa = 0.92$; 95% CI: 0.90, 0.94; $P < .001$). For dichotomous classification, the average DL model had the highest sensitivity, with 80.3% (95% CI: 72.4, 86.7) compared with radiologist 1 (74.6%; 95% CI: 66.2, 81.8) and radiologist 2 (78.7%; 95% CI: 70.4, 85.6), although this was not significant ($P = .274$ and $P = .754$, respectively). High specificities were seen for the DL model and radiologists, and

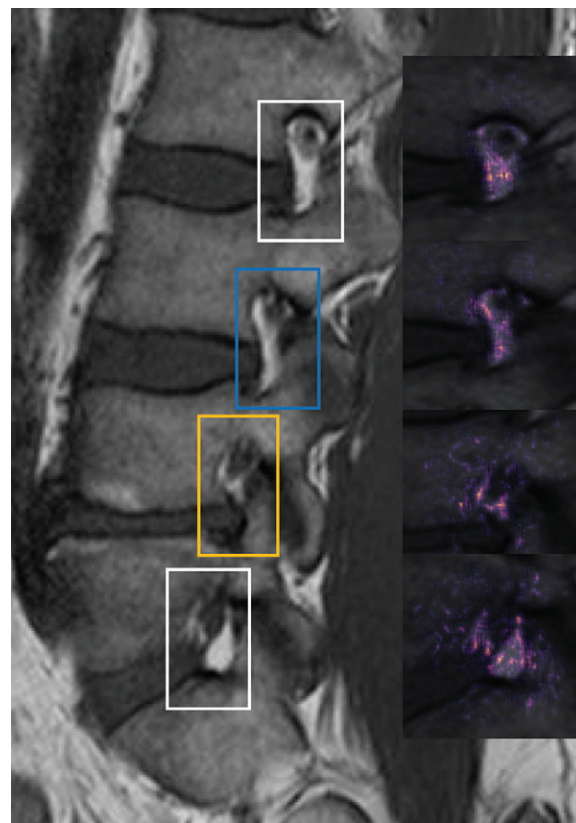


Figure 4: Sagittal T1-weighted (repetition time msec/echo time msec, 500/10) MRI scan (left) is shown off-midline at the level of the right neural foramina in a 66-year-old woman with lower back pain. Integrated Gradients technique was applied to the deep learning (DL) model and highlights the regions of the MRI scan responsible for the prediction as a color-coded overlay, or heatmap (right). In general, the higher the intensity value, the more important the feature for prediction. Examples of normal (white boxes), mild (blue box), and moderate (orange box) neural foraminal stenoses are provided. The heatmaps show that the activation of the DL model (red color) is mainly at the high-signal-intensity epidural fat surrounding the exiting nerve root (keyhole appearance).

the specificity of radiologist 2 (99.0%; 95% CI: 97.9, 99.6) was higher than that of the average DL model (96.7%; 95% CI: 95.0, 97.9; $P = .004$).

For neural foraminal classification, the average DL model agreement ($\kappa = 0.75$; 95% CI: 0.71, 0.78; $P < .001$) was reduced compared with radiologist 1 ($\kappa = 0.80$; 95% CI: 0.77, 0.83; $P < .001$) and radiologist 2 ($\kappa = 0.87$; 95% CI: 0.85, 0.89; $P < .001$). The average DL model showed almost perfect agreement for dichotomous classification of neural foraminal stenosis ($\kappa = 0.89$; 95% CI: 0.86, 0.91; $P < .001$) and was slightly reduced compared with radiologist 1 ($\kappa = 0.94$; 95% CI: 0.92, 0.96; $P < .001$) and radiologist 2 ($\kappa = 0.95$; 95% CI: 0.93, 0.96; $P < .001$). For dichotomous classification, the sensitivity of radiologist 2 (97%; 95% CI: 93.2, 99.0) was higher than that of radiologist 1 (85.8%; 95% CI: 78.7, 91.2; $P < .001$) and the average DL model (91.2%; 95% CI: 85.4, 95.3; $P = .028$). The specificity of the average DL model (91.9%; 95% CI: 90.0, 93.6) was lower than that of radiologist 1 (96.8%; 95% CI: 95.4, 97.8; $P < .001$) and radiologist 2 (95.8%; 95% CI: 94.5, 97.0; $P < .001$).

Table 3: Internal Test Set Classification Using Four Gradings (Normal, Mild, Moderate, and Severe) for Each Region of Interest

Rater	Central Canal		Lateral Recesses		Neural Foramina	
	Gwet κ	<i>P</i> Value	Gwet κ	<i>P</i> Value	Gwet κ	<i>P</i> Value
Radiologist 1	0.89 (0.87, 0.91)	<.001	0.71 (0.68, 0.75)	<.001	0.80 (0.77, 0.83)	<.001
Radiologist 2	0.89 (0.87, 0.90)	<.001	0.79 (0.76, 0.82)	<.001	0.87 (0.85, 0.89)	<.001
Average DL model	0.82 (0.80, 0.85)	<.001	0.72 (0.68, 0.76)	<.001	0.75 (0.71, 0.78)	<.001
DL model version 1	0.81 (0.79, 0.83)	<.001	0.74 (0.70, 0.77)	<.001	0.75 (0.72, 0.78)	<.001
DL model version 2	0.83 (0.81, 0.85)	<.001	0.71 (0.67, 0.75)	<.001	0.79 (0.77, 0.82)	<.001
DL model version 3	0.81 (0.79, 0.84)	<.001	0.70 (0.66, 0.74)	<.001	0.79 (0.76, 0.82)	<.001
DL model version 4	0.83 (0.81, 0.85)	<.001	0.74 (0.70, 0.78)	<.001	0.73 (0.70, 0.76)	<.001
DL model version 5	0.84 (0.82, 0.86)	<.001	0.74 (0.70, 0.77)	<.001	0.69 (0.66, 0.73)	<.001
DL model version 6	0.84 (0.82, 0.86)	<.001	0.71 (0.67, 0.74)	<.001	0.77 (0.74, 0.80)	<.001
DL model version 7	0.84 (0.83, 0.86)	<.001	0.71 (0.68, 0.75)	<.001	0.72 (0.69, 0.75)	<.001
DL model version 8	0.83 (0.81, 0.85)	<.001	0.71 (0.68, 0.75)	<.001	0.70 (0.66, 0.73)	<.001
DL model version 9	0.78 (0.76, 0.80)	<.001	0.72 (0.68, 0.75)	<.001	0.76 (0.73, 0.80)	<.001

Note.—Data in parentheses are 95% CIs. Deep learning (DL) models (versions 1–9) were averaged to obtain the average DL model.

Table 4: Internal Test Set Classification Using Dichotomous Gradings (Normal or Mild vs Moderate or Severe) for Each Region of Interest

Rater	Central Canal		Lateral Recesses		Neural Foramina	
	Gwet κ	<i>P</i> Value	Gwet κ	<i>P</i> Value	Gwet κ	<i>P</i> Value
Radiologist 1	0.98 (0.97, 0.99)	<.001	0.92 (0.90, 0.94)	<.001	0.94 (0.92, 0.96)	<.001
Radiologist 2	0.98 (0.98, 0.99)	<.001	0.95 (0.93, 0.97)	<.001	0.95 (0.93, 0.96)	<.001
Average DL model	0.96 (0.94, 0.98)	<.001	0.92 (0.90, 0.94)	<.001	0.89 (0.86, 0.91)	<.001
DL model version 1	0.96 (0.95, 0.97)	<.001	0.93 (0.91, 0.95)	<.001	0.90 (0.88, 0.92)	<.001
DL model version 2	0.97 (0.96, 0.98)	<.001	0.92 (0.90, 0.94)	<.001	0.92 (0.90, 0.94)	<.001
DL model version 3	0.95 (0.94, 0.97)	<.001	0.92 (0.89, 0.94)	<.001	0.91 (0.89, 0.93)	<.001
DL model version 4	0.97 (0.97, 0.98)	<.001	0.93 (0.90, 0.95)	<.001	0.91 (0.88, 0.93)	<.001
DL model version 5	0.97 (0.96, 0.98)	<.001	0.93 (0.91, 0.95)	<.001	0.88 (0.85, 0.90)	<.001
DL model version 6	0.97 (0.96, 0.98)	<.001	0.90 (0.87, 0.92)	<.001	0.89 (0.87, 0.91)	<.001
DL model version 7	0.98 (0.97, 0.99)	<.001	0.93 (0.90, 0.95)	<.001	0.85 (0.82, 0.88)	<.001
DL model version 8	0.96 (0.95, 0.97)	<.001	0.91 (0.89, 0.94)	<.001	0.85 (0.82, 0.88)	<.001
DL model version 9	0.94 (0.93, 0.96)	<.001	0.92 (0.89, 0.94)	<.001	0.89 (0.87, 0.91)	<.001

Note.—Data in parentheses are 95% CIs. Deep learning (DL) models (versions 1–9) were averaged to obtain the average DL model.

External Test Set ROI Classification

For central canal classification, the average DL model κ value in the external test set was 0.68 (95% CI: 0.66, 0.71; $P < .001$), which was reduced compared with that of the average DL model κ value for the internal test set (Table E3 [online]). For dichotomous classification, the average DL model showed almost perfect agreement, with a κ value of 0.95 (95% CI: 0.94, 0.97; $P < .001$) (Table E4 [online]). The average DL model had high specificity (96.3%; 95% CI: 95.3, 97.2).

For lateral recess classification, the average DL model κ value was 0.77 (95% CI: 0.75, 0.80; $P < .001$). For dichotomous classification, the average DL model κ value showed almost perfect agreement of 0.96 (95% CI: 0.95, 0.97; $P < .001$). The average DL model had high sensitivity (92.2%; 95% CI: 84.5, 96.7) and specificity (96.6%; 95% CI: 95.7, 97.4).

For neural foraminal classification, the average DL model κ value was 0.83 (95% CI: 0.81, 0.85; $P < .001$). For dichotomous classification, the average DL model κ value also showed almost perfect agreement of 0.96 (95% CI: 0.95, 0.97; $P < .001$). The average DL model had high specificity (97.9%; 95% CI: 97.1, 98.5).

Discussion

Lumbar spine MRI scans are an essential tool in the assessment of lumbar spinal stenosis (LSS) for accurate evaluation of central canal, lateral recess, and neural foraminal stenosis (1). Detailing such information in a report can be repetitive and time consuming. In our study, we trained a deep learning (DL) model for the automated detection and classification of LSS at MRI. On an internal test set, the DL model showed almost perfect agreement

for dichotomous classification (normal or mild vs moderate or severe) of lateral recess and central canal stenosis ($\kappa = 0.92$ and 0.96 , respectively; $P < .001$), similar to subspecialist radiologists ($\kappa = 0.92$ – 0.98 , $P < .001$). The DL model also showed almost perfect agreement for dichotomous classification of neural foraminal stenosis ($\kappa = 0.89$, $P < .001$), which was slightly reduced compared with subspecialist radiologists ($\kappa = 0.94$ and 0.95 , $P < .001$). In a further step, external testing of the DL model was performed on a data set from a different geographical region. Substantial levels of agreement were again seen between the DL model and a subspecialist radiologist for all regions of interest ($\kappa = 0.95$ – 0.96 , $P < .001$).

Other groups (9,11) have reported results on DL in lumbar spine MRI. Our DL model showed similar to superior performance to these studies. In addition, to our knowledge, no prior study has assessed the automated detection and classification of lateral recess stenosis, which can be specifically targeted for surgical decompression (18,25).

DL in spine imaging initially focused on automating the numbering of lumbar vertebra and intervertebral disk classification (26–28). More recently, several studies have assessed the automated grading of LSS. SpineNet (2017) is a multitask architecture for automated classification of several spinal conditions, including disk space narrowing and central canal stenosis (9). For central canal stenosis, the model was initially developed to look at binary classification (present or absent) and used sagittal T2-weighted MRI scans. In a further study, SpineNet was trained for grading central canal stenosis on axial T2-weighted images using a similar four-point grading scale to the one in our study and assessment by one expert radiologist (29). SpineNet achieved agreement of 65.7% for ordinal gradings (normal, mild, moderate, or severe) and 94% ($\kappa = 0.75$) for dichotomous grading of normal, mild, or moderate versus severe stenosis. These results compare favorably with the average DL model in our study, which had κ values of 0.82 and 0.96 for ordinal and dichotomous classification of central canal stenosis, respectively ($P < .001$).

Lu et al (11) developed a DL model to grade central canal and neural foraminal stenosis using axial and sagittal T2-weighted images. A large data set of MRI lumbar spine studies was used but relied on weakly supervised natural language processing labels from existing radiology reports. Average class accuracy (normal, mild, moderate, or severe) was 70.6% for central canal and 67.1% for neural foraminal stenosis. Most recently, in 2020, Won et al (30) developed DL algorithms to assess central canal stenosis at L4-L5 alone. Grading was performed by two orthopedic surgeons using a four-point Schizas scale (31). The agreement between the expert labels showed 77.5% accuracy, with the models showing accuracies of 77.9%–83% compared with the expert labels.

Our study has limitations. First, we selected common radiologic grading scales as the reference standard, although there remains controversy concerning LSS classification at MRI (32–34). Second, the reference standard was a single international expert who reviewed the test set independently from the two subspecialist radiologists. No consensus labeling was performed for the three radiologists, as this would have likely been biased toward

the expert. Radiology resident assessment was not analyzed but could also be assessed through further studies that include the use of DL model augmentation. Finally, labeling of images for model development was highly supervised. Manual radiologist labeling of images was believed to be the most accurate method for training the model but was labor-intensive and limited the number of MRI lumbar spine studies that could be used for training.

In conclusion, we demonstrated that our deep learning (DL) model is reliable and may be used to quickly assess lumbar spinal stenosis (LSS) at MRI. In clinical practice, the diagnosis of LSS still relies on the subjective opinion of the reporting radiologist. Our DL model could provide semiautomated reporting under the supervision of a radiologist to provide more consistent and objective reporting. Further development of the DL model could involve a consensus panel of international experts to reduce any labeling errors and biases. The DL model could also be assessed for the longitudinal follow-up of LSS at MRI.

Author contributions: Guarantors of integrity of entire study, J.T.P.D.H., L.Z., A.M., Y.H.C., B.C.O.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.T.P.D.H., L.Z., K.Y., A.M., Y.L.T., S.L., Y.S.C., S.E.E., N.K., B.C.O., S.T.Q.; clinical studies, J.T.P.D.H., A.M., D.A.R.A., B.C.O., H.Y.; experimental studies, J.T.P.D.H., L.Z., K.Y., A.M., Y.S.C., B.C.O., H.Y.; statistical analysis, J.T.P.D.H., L.Z., K.Y., D.A.R.A., Q.V.Y., Y.H.C., B.C.O.; and manuscript editing, J.T.P.D.H., L.Z., K.Y., D.A.R.A., Y.L.T., Y.S.C., S.E.E., Q.V.Y., Y.H.C., J.H.T., N.K., B.C.O., S.T.Q.

Disclosures of Conflicts of Interest: J.T.P.D.H. disclosed no relevant relationships. L.Z. disclosed no relevant relationships. K.Y. disclosed no relevant relationships. A.M. disclosed no relevant relationships. D.A.R.A. disclosed no relevant relationships. Y.L.T. disclosed no relevant relationships. S.L. disclosed no relevant relationships. Y.S.C. disclosed no relevant relationships. S.E.E. disclosed no relevant relationships. Q.V.Y. disclosed no relevant relationships. Y.H.C. disclosed no relevant relationships. J.H.T. disclosed no relevant relationships. N.K. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution received speaker fees from Lilly. Other relationships: disclosed no relevant relationships. B.C.O. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a board member of a ground transport company. Other relationships: disclosed no relevant relationships. H.Y. disclosed no relevant relationships. S.T.Q. disclosed no relevant relationships.

References

- Lurie J, Tomkins-Lane C. Management of lumbar spinal stenosis. *BMJ* 2016; 352:h6234.
- Andreisek G, Imhof M, Wertli M, et al. A systematic review of semi-quantitative and qualitative radiologic criteria for the diagnosis of lumbar spinal stenosis. *AJR Am J Roentgenol* 2013;201(5):W735–W746.
- Andreisek G, Hodler J, Steurer J. Uncertainties in the diagnosis of lumbar spinal stenosis. *Radiology* 2011;261(3):681–684.
- Arana E, Royuela A, Kovacs FM, et al. Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. *Radiology* 2010;254(3):809–817.
- Arana E, Kovacs FM, Royuela A, et al. Influence of nomenclature in the interpretation of lumbar disk contour on MR imaging: a comparison of the agreement using the combined task force and the Nordic nomenclatures. *AJNR Am J Neuroradiol* 2011;32(6):1143–1148.
- Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med* 2018;15(11):e1002699.
- Park A, Chute C, Rajpurkar P, et al. Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open* 2019;2(6):e195600.

8. Zhou Y, Liu Y, Chen Q, Gu G, Sui X. Automatic lumbar MRI detection and identification based on deep learning. *J Digit Imaging* 2019;32(3):513–520.
9. Jamaludin A, Lootus M, Kadir T, et al. ISSLS Prize in Bioengineering Science 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J* 2017;26(5):1374–1383.
10. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. *Med Image Anal* 2017;41:63–73.
11. Lu J, Pedemonte S, Bizzo B, et al. Deep Spine: Automated Lumbar Vertebral Segmentation, Disc-Level Designation, and Spinal Stenosis Grading using Deep Learning. *Proceedings of the 3rd Machine Learning for Healthcare Conference*, in PMLR. 2018; 85:403–419.
12. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. *JOR Spine* 2019;2(1):e1044.
13. England JR, Cheng PM. Artificial Intelligence for medical image analysis: a guide for authors and reviewers. *AJR Am J Roentgenol* 2019;212(3):513–519.
14. LabelImg. <https://github.com/tzutalin/labelImg>. Accessed August 16, 2019.
15. Lurie JD, Tosteson AN, Tosteson TD, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. *Spine* 2008;33(14):1605–1610.
16. Fardon DF, Williams AL, Dohring EJ, Murtagh FR, Gabriel Rothman SL, Sze GK. Lumbar disc nomenclature: version 2.0: recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. *Spine J* 2014;14(11):2525–2545.
17. Andreisek G, Deyo RA, Jarvik JG, et al. Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. *Eur Radiol* 2014;24(12):3224–3232.
18. Bartynski WS, Lin L. Lumbar root compression in the lateral recess: MR imaging, conventional myelography, and CT myelography comparison with surgical confirmation. *AJNR Am J Neuroradiol* 2003;24(3):348–360.
19. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* 2017;39(6):1137–1149.
20. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 770–778.
21. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *ICML* 2017;70:3319–3328. <https://dl.acm.org/doi/10.5555/3305890.3306024>.
22. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 2008;61(Pt 1):29–48.
23. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol* 2013;13(1):61.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159–174.
25. Colak A, Topuz K, Kutlay M, et al. A less invasive surgical approach in the lumbar lateral recess stenosis: direct approach to the medial wall of the pedicle. *Eur Spine J* 2008;17(12):1745–1751.
26. Lootus M, Kadir T, Zisserman A. Vertebrae Detection and Labelling in Lumbar MR Images. In: Yao J, Klinder T, Li S, eds. *Computational Methods and Clinical Applications for Spine Imaging*. Lecture Notes in Computational Vision and Biomechanics, vol 17. Cham, Switzerland: Springer, 2014; 219–230.
27. Castro-Mateos I, Hua R, Pozo JM, Lazary A, Frangi AF. Intervertebral disc classification by its degree of degeneration from T2-weighted magnetic resonance images. *Eur Spine J* 2016;25(9):2721–2727.
28. Lootus M, Kadir T, Zisserman A. Automated Radiological Grading of Spinal MRI. In: Yao J, Glocker B, Klinder T, Li S, eds. *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*. Lecture Notes in Computational Vision and Biomechanics, vol 20. Cham, Switzerland: Springer, 2015; 119–130.
29. Ishimoto Y, Jamaludin A, Cooper C, et al. Could automated machine-learned MRI grading aid epidemiological studies of lumbar spinal stenosis? Validation within the Wakayama spine study. *BMC Musculoskelet Disord* 2020;21(1):158.
30. Won D, Lee HJ, Lee SJ, Park SH. Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks. *Spine* 2020;45(12):804–812.
31. Schizas C, Theumann N, Burn A, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. *Spine* 2010;35(21):1919–1924.
32. Schönström N, Lindahl S, Willén J, Hansson T. Dynamic changes in the dimensions of the lumbar spinal canal: an experimental study in vitro. *J Orthop Res* 1989;7(1):115–121.
33. Mamisch N, Brumann M, Hodler J, et al. Radiologic criteria for the diagnosis of spinal stenosis: results of a Delphi survey. *Radiology* 2012;264(1):174–179.
34. Fu MC, Buerba RA, Long WD 3rd, et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. *Spine J* 2014;14(10):2442–2448.