Online Resource 1 - PubMed search string; adapted for other databases

((("Algorithms"[Mesh] OR "Algorithms"[tw] OR "Algorithm"[tw] OR "Algorithmic"[tw] OR "Algorithm*"[tw] OR "Artificial Intelligence"[mesh] OR "Artificial Intelligence"[tw] OR "Machine Learning"[mesh] OR "Machine Learning"[tw] OR "Deep Learning"[tw] OR "Neural Networks"[tw] OR "Neural Network"[tw] OR "Artificial Intelligence"[title/abstract:~4] OR "Machine Learning"[title/abstract:~4] OR "Deep Learning"[title/abstract:~4]) AND ("Spinal Stenosis"[Mesh] OR "Spinal Stenosis"[tw] OR "Spine Stenosis"[tw] OR "Spinal Stenosis"[title/abstract:~4] OR "Spine Stenosis"[title/abstract:~4] OR "Spines Stenosis"[title/abstract:~4] OR "Spinal Degeneration"[title/abstract:~4] OR "Spine Degeneration"[title/abstract:~4] OR "Spines Degeneration"[title/abstract:~4] OR "Spinal Degenerative"[title/abstract:~4] OR "Spine Degenerative"[title/abstract:~4] OR "Spines Degenerative"[title/abstract:~4] OR "Intervertebral Disc Degeneration"[Mesh] OR "Intervertebral Disc Degeneration"[tw] OR "Intervertebral Disk Degeneration"[tw] OR "Sciatica"[Mesh] OR "Sciatica"[tw] OR "Intervertebral Disc Displacement"[Mesh] OR "Intervertebral Disc Displacement"[tw] OR "Intervertebral Disk Displacement"[tw] OR "Disc Degeneration"[tw] OR "Disk Degeneration"[tw] OR "Disc Displacement"[tw] OR "Disk Displacement"[tw] OR "Disc Herniation"[tw] OR "Disk Herniation"[tw] OR "Disc Herniations"[tw] OR "Disk Herniations"[tw] OR "Disc Hernia"[tw] OR "Disk Hernia"[tw] OR "Disc Hernias"[tw] OR "Disk Hernias"[tw] OR "Disc Degeneration"[title/abstract:~4] OR "Disk Degeneration"[title/abstract:~4] OR "Disc Degradation"[title/abstract:~4] OR "Disk Degradation"[title/abstract:~4] OR "Disc Displacement"[title/abstract:~4] OR "Disk Displacement"[title/abstract:~4] OR "Disc Herniation"[title/abstract:~4] OR "Disk Herniation"[title/abstract:~4] OR "Disc Herniations"[title/abstract:~4] OR "Disk Herniations"[title/abstract:~4] OR "Disc Hernia"[title/abstract:~4] OR "Disk Hernia"[title/abstract:~4] OR "Disc Hernias"[title/abstract:~4] OR "Disk Hernias"[title/abstract:~4]) AND ("Lumbar Vertebrae"[Mesh] OR "Lumbosacral Region"[Mesh] OR "Lumbar"[tw] OR "Lumbosacral"[tw])) OR (("Algorithms"[Mesh] OR "Algorithms"[tw] OR "Algorithm"[tw] OR "Algorithmic"[tw] OR "Algorithm*"[tw] OR "Artificial Intelligence"[mesh] OR "Artificial Intelligence"[tw] OR "Machine Learning"[mesh] OR "Machine Learning"[tw] OR "Deep Learning"[tw] OR "Neural Networks"[tw] OR "Neural Network"[tw] OR "Artificial Intelligence"[title/abstract:~4] OR "Machine Learning"[title/abstract:~4] OR "Deep Learning"[title/abstract:~4]) AND ("Spinal Stenosis"[majr] OR "Spinal Stenosis"[ti] OR "Spine Stenosis"[ti] OR "Spinal Stenosis"[title:~4] OR "Spine Stenosis"[title:~4] OR "Spines Stenosis"[title:~4] OR "Spinal Degeneration"[title:~4] OR "Spine Degeneration"[title:~4] OR "Spines Degeneration"[title:~4] OR "Spinal Degenerative"[title:~4] OR "Spine

Degenerative"[title:~4] OR "Spines Degenerative"[title:~4] OR "Intervertebral Disc Degeneration"[majr] OR "Intervertebral Disc Degeneration"[ti] OR "Intervertebral Disk Degeneration"[ti] OR "Sciatica"[majr] OR "Sciatica"[ti] OR "Intervertebral Disc Displacement"[majr] OR "Intervertebral Disc Displacement"[ti] OR "Intervertebral Disk Displacement"[ti] OR "Disc Degeneration"[ti] OR "Disk Degeneration"[ti] OR "Disc Displacement"[ti] OR "Disk Displacement"[ti] OR "Disc Herniation"[ti] OR "Disk Herniation"[ti] OR "Disc Herniations"[ti] OR "Disk Herniations"[ti] OR "Disc Hernia"[ti] OR "Disk Hernia"[ti] OR "Disc Hernias"[ti] OR "Disk Hernias"[ti] OR "Disc Degeneration"[title:~4] OR "Disk Degeneration"[title:~4] OR "Disc Degredation"[title:~4] OR "Disk Degredation"[title:~4] OR "Disc Displacement"[title:~4] OR "Disk Displacement"[title:~4] OR "Disc Herniation"[title:~4] OR "Disk Herniation"[title:~4] OR "Disc Herniations"[title:~4] OR "Disk Herniations"[title:~4] OR "Disc Hernia"[title:~4] OR "Disk Hernia"[title:~4] OR "Disc Hernias"[title:~4] OR "Disk Hernias"[title:~4])))

Online Resource 2 – Cowley's Adjusted Risk of Bias Tool

A maximum of ten points could be awarded to each assessed study. For proper description of the studied population, model architecture, comparison and ground truth (1) a maximum of four points could be awarded; for data variability control and/or full automatization of the model (2) one point could be awarded; for assessment of ground truth, comprehensive outcome reporting and data/code sharing (3) a total of five points could be awarded. Points were assigned if an item matched a criterion.

1	1.	Description of cohort	*max = 1*
	2.	Model selection, development, implementation	*max = 1*
	3.	Comparison	*max = 1*
	4.	Ground truth description	*max = 1*
2	1.	Data variability control and model automatization	*max = 1*
3	1.	Ground truth assessment	*max = 1*
	2.	Outcome reporting	*max = 2*
	3.	Sharing of data and code	*max = 2*
			Total = 10

1.1 Description of cohort

a) Detailed description (e.g., origin, age, male:female ratio) = 1
b) No details provided on patient population = 0

1.2 Model selection, development, implementation

a) Adequate description = 1
b) Somewhat adequate/no description = 0

1.3 Comparison

a) Comparison with ground truth (e.g., clinical expert, existing model) = 1
b) No comparison to model performance = 0

1.4 Ground truth description

a) Clear description of ground truth = 1
b) No clear description = 0

2.1 Data variability control and model automatization

a) Model controls for variation in data (e.g., demographic variation between patients in test and training set), other factors (any correction) and/or is fully automatic (i.e., no pre-processing steps like cropping of images) = 1
b) No controlling for variation/factors and not fully automatic = 0

3.1 Ground truth assessment

a) Ground truth performance objectified = 1

b) Ground truth performance not objectified = 0

3.2 Outcome reporting

a) Two or more outcome measures with uncertainty metrics reported = 2
b) More than two outcome measures but less than two with uncertainty measures = 1
c) Two or less outcome measures; less than two without uncertainty measures = 0

3.3 Sharing of data and code

a) Data and code sharing = 2
b) Data or code sharing = 1
c) No sharing/training data already shared by third party/publicly available = 0

Online Resource 3 – Overview of study characteristics

Study	Study aim	Modality	Scanner characteristics	Input	Conventional ML or DL	Model name	Model architecture	Original or external validation	Transfer Learning	Loss function	Optimizer	Comparison	Ground Truth	Degree of automation
Al Kafri et al., 2019 [20]	Segmentation	MRI	<u>Axial:</u> TR: T1 385-953 ms, T2 1900-5000 ms TE: T1 11.0 ms, T2 84.0-96.0 ms Slice thickness: T1 4.0 mm, T2 3.0-5.0 mm Slice gap: T1 4.4 mm, T2 3.3-6.5 mm FoV: T1 220 mm, T2 220 mm Flip angle: T1 150 degrees, T2 150 degrees <u>Sagittal:</u> TR: T1 330-926 ms, T2 3190-4000 ms TE: T1 9.2-12.0 ms, T2 67.0-96.0 ms Slice thickness: T1 3.0-4.0 mm, T2 3.0-5.0 mm Slice gap: T1 3.3-4.8 mm, T2 3.3-6.5 mm FoV: T1 280 mm, T2 280 mm Flip angle: T1 150 degrees, T2 150 degrees	Axial, T1,T2	DL	SegNet-TL80	SegNet (based on VGG16)	Original	Yes	Cross-entropy of probability distribution	(1+1)-Evolutionary Strategy, stochastic gradient descent with momentum algorithm	Model vs. human	Manual labelling by five radiologists	Fully
Al Kafri et al., 2018 [21]	Segmentation	MRI	No description	Axial, T2	DL	PALMSNet	CNN	Original	No	Cross-entropy of probability distribution	Stochastic gradient descent with momentum algorithm	Model vs. human and model vs. model	Human: physicians labelled binary image marking (GT) Models: fine Gaussian SVM, medium Gaussian SVM, cubic k-nearest	Fully

													neighbour, medium k-nearest neighbour, weighted k-nearest neighbour, coarse k-nearest neighbour, Complex Tree, Medium Tree, Boosted Tree Ensemble, Bagged Tree Ensemble (all ML), SegNet (DL)	
Altun et al., 2023 [22]	Classification	MRI	No description	Axial, T2	Both	Conventional ML: Sobel-RF, Sobel-SVM, Gabor-RF, Gabor-SVM	Gabor-RF: Gabor operator + random forest Gabor-SVM: Gabor operator + SVM Sobel-RF: Sobel operator + random forest Sobel-SVM: Sobel operator + SVM	Original	No	No description	No description	Model vs. model	No description	Semi
						DL: InceptionNet, MobileNet, ResNet, VGG16	InceptionNet: CNN MobileNet: CNN ResNet: CNN VGG16: CNN	Original	Yes, MobileNet	No description	No description	Model vs. model	No description	Fully
Altun et al., 2022 [23]	Segmentation	MRI	Vendor/type: GE Optima 360 advance model Field strength: 1.5T Size: 512x512 pixels Number of slices per patient: 41 ± 4	Axial, T2	DL	-	3D-U-Net	Original	No	Cross-entropy	ADAM	Model vs. human	20 students supervised by expert neurosurgeon (15 yrs experience)	Fully
Gaonkar et al., 2019 [24]	Segmentation	MRI	Resolution per pixel: mean 0.53 ± 0.125 x 0.53 ± 0.125 mm (range 0.27 x 0.27 - 1.5 x 1.5 mm)	Central canal segmentation: axial, T2	Conventional ML	-	Linear SVMs followed by an ensemble of regression trees	Original	No	Negative of the Dice score	ADAM	Model vs. human	Student delineated canals. Attending physician reviewed each section and corrected the	Semi

			Mean TR: 3756 ± 738 ms Mean TE: 107 ± 12 ms										student-generated delineations.	
			Resolution per pixel: range 0.5x0.5 - 2x2 mm	<u>Disc segmentation:</u> sagittal, T2	DL	-	Deep-U-Net	Original	No	Negative of the Dice score	ADAM	Model vs. human	Student delineated lumbar discs. Attending physician reviewed each section and corrected the student-generated delineations.	Semi
Ghosh et al., 2014 [25]	Segmentation	MRI	Vendor/type: Philips MRI scanner Field strength: 3T	Axial, sagittal, T1, T2	Conventional ML	-	HOG feature descriptors and random forests as the classifier	Original	No	No description	No description	Model vs. human	Manual segmentation using own labeling tool	Fully
Grob et al., 2022 [26]	Classification	MRI	Variety of MR imaging techniques and acquisition protocols	Sagittal, T2	DL	SpineNet (v1.1)	CNN	External validation	No	Multi-task loss functions, class-balanced loss	Stochastic Gradient Descent with momentum	Model vs. human	Annotations by an expert radiologist (10 yrs experience)	Fully
Hallinan et al., 2021 [27]	Classification	MRI	Vendor/type: Internal dataset: GE Optima and Discovery, and Siemens Aera and Skyra External dataset: Philips Healthcare Systems Achieva, and GE Optima and Discovery Field strength: 1.5 and 3.0T <u>Axial:</u> TR: 3400-5300 ms TE: 80-120 ms Slice thickness: 4-5 mm Slice gap: 2 mm Field of view: 160x160 - 200x200 mm^2	<u>Central canal, lateral recess classification:</u> axial, T2	DL	SpineA@NUHS-NUS	CNN	Original	No	Weighted categorical loss function	No description	Model vs. human	Annotation by a musculoskeletal radiologist (31 yrs experience)	Fully
			<u>Sagittal:</u> TR: 450-900 ms TE: 10-21 ms Slice thickness:	<u>Neural foramina classification:</u> sagittal, T1										

			3.5-4 mm Slice gap: 1-2 mm Field of view: 300x300 mm^2											
Han et al., 2018 [28]	Segmentation and classification	MRI	Field strength: 1.5T TR: 380-4,000 ms, mean: 1,529 ms TE: 8.144-151 ms, mean: 38.35 ms Slice thickness: 0.879-4 mm, mean: 3.14 mm Pixel spacing: 0.38×0.38 - 0.86x0.86 mm, mean: 0.56x0.56 mm	Sagittal, T1, T2	DL	Spine-GAN	ACAE, local LSTM based RNN and CNN	Original	No	Hybrid loss function: segmentation loss and GAN loss	RMSProp, ADAM	Model vs. human and model vs. model	Human: extraction from clinical report of spinal surgery, double-checked by two experienced spinal radiologists Models: FCN, SegNet, DeepLab v3+, U-Net, ACAE, ACAE+ LSTM, ACAE+CNN, Spine-GAN	Fully
Hou et al., 2022 [29]	Segmentation	MRI	Resolution: 320 × 320 Slice thickness: 4 mm Field of view: 220 mm	Axial, T2	DL	Spine-Seg-2 phase model	FC-DenseNet (variant of U-Net)	Original	No description	Gaussian divergence loss, contour loss, cross-entropy loss	ADAM	Model vs. human and model vs. model	Human: consensus from 5 expert radiologists (GT) Models: Seg-Net-TL-80	Semi
Huber et al., 2019 [30]	Classification	MRI	No description	Axial, T2	Conventional ML	Weka version 3	J48 (c4.5) decision tree	Original	No	No description	No description	Model vs. reference standard	Reference standard: CSA cut-off value	Semi
Ishimoto et al., 2020 [31]	Classification	MRI	Vendor/type: Toshiba Excelart Field strength: 1.5T TR: 4000 ms TE: 120 ms Field of view: 180×180 mm	Axial, T2	DL	SpineNet (v1.0)	CNN	External validation	No	Multi-task loss functions, class-balanced loss	Stochastic Gradient Descent with momentum	Model vs. human	Spinal surgeon annotated gradings	Fully
Jamaludin et al., 2017 (1) [32]	Classification	MRI	Varying acquisition protocols used	Sagittal, T2	DL	-	CNN	Original	No	No description	No description	Model vs. human	Single expert spinal radiologist	Fully
Jamaludin et al., 2017 (2) [33]	Classification	MRI	Varying machines and acquisition protocols used Slice thickness: 2.6-6.0 mm, median 4 mm	Sagittal, T2	DL	SpineNet (v1.0)	CNN	Original	No	Multi-task loss functions, class-balanced loss	Stochastic Gradient Descent with momentum	Model vs. human	Single expert spinal radiologist	Fully
Kim et al., 2022 [34]	Classification	X-ray	Neutral, flexion, and extension	Sagittal	DL	Efficient1, ResNet50, VGG16, VGG19	Efficient1: CNN ResNet50: CNN	Original	Yes, ImageNet-pretrained models	Categorical cross-entropy loss function	Stochastic Gradient Descent	Model vs. human	Human: confirmation of severe central LSS based on a	Semi

							VGG16: CNN VGG19: CNN		were used as an initial parameter for algorithm training				formal MRI reading Model: Efficient1, ResNet50, VGG16, VGG19	
Koompairojn et al., 2010 [35]	Segmentation	MRI	TR: 1000-2500 ms, mostly 1290 TE: 25-30 ms, mostly 26	Axial, T2	Conventional ML	-	Multilayer Perceptron framework	Original	No	No description	Gradient descent	Model vs. human	Radiologists verified the extracted spinal features	Semi
	Classification	MRI	No description	Axial, T2	Conventional ML	Weka	Multilayer Perceptron framework	Original	No description	No description	Gradient descent	Model vs. human	Agreement between at least one radiologist and one orthopedist	Semi
Lehnen et al., 2021 [36]	Classification	MRI	Vendor/type: Philips Healthcare Achieva Field strength: 1.5 and 3.0T TR: 3000-5000 ms TE: 80-120 ms Axial: Slice thickness: 3.5 mm Slice gap: 0.35 mm Sagittal: Slice thickness: 4 mm Slice gap: 0.4 mm	Axial, sagittal, T2	DL	CoLumbo	U-Net-based CNN	External validation	No	Pixelwise cross-entropy loss	Stochastic Gradient Descent with momentum	Model vs. human	Annotation by an experienced radiologist (5 yrs experience)	Fully
Lewandrowski et al., 2020 (1) [37]	Classification	MRI	No description	Axial, sagittal, T1, T2	DL	Multus RadBot	VGG-like classifier	Original	No	Binary cross-entropy log loss function	Stochastic gradient descent	Model vs. human	Radiology report from expert radiologists	Semi
Lewandrowski et al., 2020 (2) [38]	Classification	MRI	No description	No description	DL	Multus RadBot	VGG-like classifier	External validation	No description	No description	No description	Model vs. human	Radiology report from expert radiologists	Semi
Li et al., 2021 [39]	Segmentation	MRI	No description	Axial, T2	DL	MANet	Multi-scale attention U-Net	Original	No	Generalized dice loss and the cross-entropy loss	Stochastic gradient descent	Model vs. human and model vs. model	Human: manual labelling by four spine surgeons and two imaging surgeons, which were double-checked by one spine surgery specialist (>30 yrs experience) and two spine	Semi

													surgeons (>10 yrs experience) (GT) Model: FCN, U-Net, Res2net, Deeplabv3 +, U-Net + multi-scale, U-Net + attention (all DL)	
Lin et al., 2021 [40]	Classification	MRI	Vendor/type: Philips Medical Systems Ingenia TR: 3228.15 ms TE: 120 ms Flip angle: 90°	Axial, T2	DL	OSBP-Net	OSBP-Net: object-specific bi-path CNN	Original	No	Primary loss, weighted IPDC loss, weighted IICR loss	No description	Model vs. human and model vs. model	Manual measurements by two physicians (8 yrs experience) (GT) MF+RF: multiple features with random forest (ML) HOG+AKRF: HOG with adjusted K-clustered random forest (ML) HOG+SSVR: HOG with structured support vector regression (ML) CARN: cascade amplifier network with linear regression model (ML) DE-Net: dense-enhanced CNN (DL)	Semi
Lu et al., 2018 [41]	Classification	MRI	Vendor/type: GE Healthcare and Siemens Medical Imaging Field strength: at least 1.5T Slice thickness: 0.9-4.0 mm	Axial, sagittal, T2	DL	DeepSPINE	Bi-path ResNeXt-50-based CNN	Original	No	Weighted categorical cross-entropy loss function	Adadelta optimizer	Model vs. human	Report from a single radiologist from neuroradiology or musculoskeletal radiology division	Fully
Natalia et al., 2020 [42]	Segmentation	MRI	Resolution: 320×320 pixels Slice thickness: 4 mm Field of view: 220 mm	Axial, T1, T2	DL	SegNet with a contour evolution algorithm	SegNet	Original	Yes, SegNet model with pre-trained VGG16 coefficients on the ImageNet database	Softmax cross-entropy multi-taks loss	Stochastic Gradient Descent with Momentum algorithm, (1+1)-Evolutionary Strategy	Model vs. human	Manual annotation of two region boundaries by expert radiologists	Fully

Pang et al., 2023 [43]	Classification	MRI	Resolution: 310x320 - 320x320 pixels Pixel spacing: 0.6875x0.6875 mm Slice thickness: 4 mm Field of view: 220 mm	Axial, T1, T2	DL	SGRNet	U-Net based CNN with VG-11 encoder, convolutional block attention module-based decoder	Original	No	Multitask loss function: regression loss, segmentation loss, deep supervision loss functions	ADAMW	Model vs. model	Automatically generated spine indices from previous study* (GT) VGG-11: CNN (DL) ResNet18: CNN (DL) ResNet50: CNN (DL) ResNet101: CNN (DL) HourglassNet-2-1: CNN (DL) HourglassNet-4-1: CNN (DL) HourglassNet-8-1: CNN (DL) HourglassNet-4-2: CNN (DL) CARN: cascade amplifier network with linear regression model (ML) DENet: dense-enhanced CNN (DL)	Semi
Su et al., 2022 [44]	Classification	MRI	Vendor/type: Siemens Healthineers Magnetom Verio Field strength: 3.0T TR: 3500-3775 ms TE: 94-120 ms Field of view: 153 × 153 mm^2 Slice thickness: 4.0-4.5 mm	Axial, T2	DL	-	ResNet50 with three fully connected networks	Original	No description	Cross-entropy loss function	ADAM	Model vs. human	Annotation by 2-3 clinicians (8-25 yrs experience)	Fully
Tang et al., 2020 [45]	Segmentation	CT	No description	Axial	DL	DDU-Net	Two U-Net sub-networks	Original	No	Weighted cross-entropy loss function	Stochastic gradient descent (mini-batch SGD)	Model vs. human and model vs. model	Human: consensus from manual segmentation by four radiologists using a custom designed interactive segmentation tool (GT) Models: FCN-8s,	Fully

													U-Net, DeeplabV3, DDU-Net (all DL)	
Won et al., 2020 [46]	Classification	MRI	Vendor/type: Siemens Healthineers Magnetom Verio Field strength: 1.5T Slice thickness: 4 mm	Axial, T2	DL	-	Fully automatic model (FA): Faster R-CNN10 with VGG-16 Semi-automatic model (SA): VGG-16	Original	Yes, the base network in R-CNN is pretrained	Cross-entropy with softmax function	ADAM	Model vs. human	Manual annotation by orthopaedic and spine surgeon (>10 yrs experience)	Fully and semi

ACAE: atrous convolution autoencoder; CNN: convolutional neural network; CSA: cross-sectional area; CT: computed tomography; DDU-Net: dual-decoder-U-net; DL: deep learning; FCN: fully convolutional network; FoV: field of view; GAN: generative adversarial network; GT: ground truth; HOG: histogram of oriented gradients; IICR: inter-index correlation regularization; IPDC: inter-path dissimilarity constraint; LSTM: long short term memory; ML: machine learning; MRI: magnetic resonance imaging; RF: random forest; RMSprop: root mean squared propagation; RNN: recurrent neural network; SGAM: segmentation-guided attention module; SVM: support vector machine; TE: echo time; TR: repetition time.

* = Natalia F, Meidia H, Afriliana N, et al. Automated measurement of anteroposterior diameter and foraminal widths in MRI images for lumbar spinal stenosis diagnosis. PLoS One. 2020;15:e0241 309.

Online Resource 4 – Risk of Bias scores

Study	1.1 Description of cohort	1.2 Model selection, development, implementation	1.3 Comparison	1.4 Ground truth description	2.1 Data variability control and model automatization	3.1 Ground truth assessment	3.2 Outcome Reporting	3.3 Sharing of data and code	Total Score (0-10)
Al Kafri et al., 2019 [20]	0	1	1	1	1	1	0	2	7
Al Kafri et al., 2018 [21]	0	1	1	1	1	0	1	0	5
Altun et al., 2023 [22]	1	1	1	0	1	0	1	0	5
Altun et al., 2022 [23]	0	1	1	1	1	0	0	0	4
Gaonkar et al., 2019 [24]	0	1	1	0	0	1	2	0	5
Ghosh et al., 2014 [25]	0	1	1	0	1	0	1	0	4
Grob et al., 2022 [26]	1	1	1	1	1	1	1	1	8
Hallinan et al., 2021 [27]	1	1	1	1	1	1	2	1	9
Han et al., 2018 [28]	0	1	1	0	1	0	2	0	5
Hou et al., 2022 [29]	0	1	1	0	0	1	0	0	3
Huber et al., 2019 [30]	1	0	1	1	0	1	1	0	5
Ishimoto et al., 2020 [31]	1	1	1	1	1	1	1	2	9
Jamaludin et al., 2017 (1) [32]	0	1	1	0	1	1	1	0	5
Jamaludin et al., 2017 (2) [33]	0	1	1	0	1	1	2	0	6
Kim et al., 2022 [34]	1	1	1	0	0	0	2	0	5
Koompairojn et al., 2010 [35]	0	1	1	0	0	0	0	0	2
Lehnen et al., 2021 [36]	1	1	1	1	1	0	1	0	6
Lewandrowski et al., 2020 (1) [37]	1	1	1	1	0	0	1	0	5
Lewandrowski et al., 2020 (2) [38]	1	0	1	1	0	0	2	0	5
Li et al., 2021 [39]	1	1	1	1	0	0	1	0	5
Lin et al., 2021 [40]	0	1	1	1	0	0	2	0	5
Lu et al., 2018 [41]	0	1	1	1	1	0	2	0	6

Natalia et al., 2020 [42]	1	1	1	0	1	0	0	2	6
Pang et al., 2023 [43]	0	1	1	1	0	1	1	0	5
Su et al., 2022 [44]	1	1	1	1	1	1	1	0	7
Tang et al., 2020 [45]	0	1	1	0	1	0	1	0	4
Won et al., 2020 [46]	0	1	1	1	0	1	1	0	5

Online Resource 5 – Data extraction for ML algorithms on segmentation

Study	ROI segmented	Segmentation labels	Sample size	Pre-processing	Model augmentation	Data augmentation	Data imbalance	Split	Validation step applied	Validation method	Testing method	Outcome measures	Uncertainty metrics	Results	Conclusion
Gaonkar et al., 2019 [24]	IVD (sagittal), spinal canal (axial)	IVD: part of IVD, not part of IVD Spinal canal: part of spinal canal, not part of spinal canal	209 patients, number of slices (per patient) not explicitly stated	Nonparametric bias correction, linear histogram matching to a common template, intensity normalization to the 0–1 range, linear image registration, resampling to 256x256 pixels	No description	No	No description	Training: 48%; test: 52%	No	-	Out-of-sample	Dice score, Hausdorff distance (mm), average surface distance (mm)	95% CI for all outcomes	<u>Spinal canal segmentation:</u> Dice score: Auto vs. Rater 1: 0.84 ± 0.08 Auto vs. Rater 2: 0.83 ± 0.08 Hausdorff distance: Auto vs. Rater 1: 7.89 ± 9.42 Auto vs. Rater 2: 9.41 ± 11.2 Average surface distance: Auto vs. Rater 1: 0.84 ± 0.08 Auto vs. Rater 2: 0.83 ± 0.08	*"Machine-generated segmentations were qualitatively comparable with those generated by human experts"* *"Machine generated segmentation agrees almost as well with each human expert as the human experts agree among themselves"*
Ghosh et al., 2014 [25]	Dural sac, IVD	Background, dural sac, IVD, vertebra	262 cases, 24 or 30 slices per patient	No	No description	No description	No description	Training: 19%; test: 81%	Yes	5-fold cross-validation	Out-of-sample	Dice similarity index, precision, relative overlap (Jaccard index), sensitivity	No	<u>Disc segmentation:</u> Dice score: 0.84 Precision: 0.79 Relative overlap: 0.73 Sensitivity: 0.90 <u>Dural sac segmentation:</u> Dice score: 0.87 Precision: 0.83 Relative overlap: 0.78 Sensitivity: 0.92	*"This method performs substantially better than previously reported approaches, but there is still scope for improvement. It suffers from oversegmentation of the dural sac and intervertebral discs"*
Koompairojn et al., 2010 [35]	IVD, ligamentum flavum	Bottom of IVD, left ligamentum	50 subjects,	Images downsampled	No	No description	For CSS: L2-L3 mild imbalance	Training: 90%;	No	-	10-fold cross-	Segmentation quality	No	<u>Segmentation quality:</u> Spinal canal:	*"The accuracy consistently ranges above*

	and facets, spinal canal, superior articular facets	flavum, right ligamentum flavum, left superior articular facet, right superior articular facet, spinal canal	number of slices (per patient) not explicitly stated				For LSS: L3-L4 mild imbalance, L2-L3 moderate imbalance	test: 10%			validation	(TP/(TP+FP+FN)		92.47% IVD: 91.47% Right superior articular facet: 93.33% Left superior articular facet: 91.25% Right ligamentum flavum & facet: 97.15% Left ligamentum flavum & facet: 98.21%	*91%, which is adequate for our diagnosis purposes"*

Conclusions written in italics were quoted from the article. CI: confidence interval; CSS: central spinal stenosis; FN: false negatives; FP: false positives; IVD: intervertebral disc; LSS: lateral spinal stenosis; ROI: region of interest; TP: true positives.

Online Resource 6 – Data extraction for DL algorithms on segmentation

Study	ROI segmented	Segmentation labels	Sample size	Pre-processing	Model augmentation	Data augmentation	Data imbalance	Split	Validation step applied	Validation method	Testing method	Outcome measures	Uncertainty metrics	Results	Conclusion
Al Kafri et al., 2019 [20]	AAP, IVD, PE, TS	AAP, IVD, PE, TS, other, unregistered	515 patients, 48,345 MRI slices, 3 axial slices per patient, 1 for each IVD from L3-4 to L5-S1	Resizing to 360x480 pixels	No	No	Yes, class population imbalances	Training: 80%; test: 20%	No	-	Out-of-sample	Class accuracy, class intersection-over-union (Jaccard index)	No	Class accuracy: AAP: 0.93 IVD: 0.99 PE: 0.96 TS: 0.96 Other: 0.98 Unregistered: 0.50 Iouc: AAP: 0.53 IVD: 0.92 PE: 0.78 TS: 0.85 Other: 0.98 Unregistered: 0.21	*„We concluded that the model's performance is within the range of manual labelling performance. (…) Through visual inspection of these results, we can confidently claim that our proposed method is sufficiently accurate, and the results are suitable for computer-aided-diagnosis purposes."*
Al Kafri et al., 2018 [21]	AAP region (TS and nerve roots)	AAP, not-AAP	10 patients, 572,400 image patches, number of slices per patient not explicitly stated	Resizing to 25x25 pixels	No	No	No	Training: 70%; test: 30%	No	-	Out-of-sample	Mean accuracy, pixel accuracy, mean intersection-over-union (IoUm) (Jaccard index), frequency weighted intersection-over-union (IoUfw) (frequency weighted Jaccard index)	No	Mean accuracy: Fine Gaussian SVM: 0.77 Medium Gaussian SVM: 0.78 Cubic k-nearest neighbour: 0.78 Medium k-nearest neighbour: 0.78 Weighted k-nearest neighbour: 0.77 Coarse k-nearest neighbour: 0.79 Complex Tree: 0.75 Medium Tree: 0.76	*„The architecture design of PALMSNet, a Deep Neural Network suitable to use for image segmentation of lumbar spine MRI."*

														Boosted Tree Ensemble: 0.76 Bagged Tree Ensemble: 0.75 SegNet: 0.67 PALMSNet: 0.85 Pixel accuracy: Fine Gaussian SVM: 0.78 Medium Gaussian SVM: 0.79 Cubic k-nearest neighbour: 0.79 Medium k-nearest neighbour: 0.79 Weighted k-nearest neighbour: 0.77 Coarse k-nearest neighbour: 0.79 Complex Tree: 0.75 Medium Tree: 0.75 Boosted Tree Ensemble: 0.76 Bagged Tree Ensemble: 0.75 SegNet: 0.75 PALMSNet: 0.86 IoUm: Fine Gaussian SVM: 0.62 Medium Gaussian SVM: 0.64 Cubic k-nearest neighbour: 0.64 Medium k-nearest neighbour: 0.64	

														Weighted k-nearest neighbour: 0.62 Coarse k-nearest neighbour: 0.64 Complex Tree: 0.59 Medium Tree: 0.59 Boosted Tree Ensemble: 0.60 Bagged Tree Ensemble: 0.59 SegNet: 0.53 PALMSNet: 0.75 <u>IoUfw:</u> Fine Gaussian SVM: 0.64 Medium Gaussian SVM: 0.66 Cubic k-nearest neighbour: 0.66 Medium k-nearest neighbour: 0.66 Weighted k-nearest neighbour: 0.64 Coarse k-nearest neighbour: 0.66 Complex Tree: 0.61 Medium Tree: 0.61 Boosted Tree Ensemble: 0.62 Bagged Tree Ensemble: 0.61 SegNet: 0.58 PALMSNet: 0.76	
Altun et al., 2022 [23]	Canal, IVD, PE, TS	Background, canal, IVD, PE, TS, other	300 LSS patients, approximately 777,778 images,	Application of Gaussian filter	No	No	No description	Training: 90%; test: 10%	No	-	Out-of-sample	Dice score, intersection-over-union (IoU)	No	<u>Dice score:</u> Background: 0.81 Canal: 0.61 IVD: 0.88	*"The 3D automatic segmentation success rates obtained are*

			number of slices per patient not explicitly stated									(Jaccard index)		Other: 0.96 PE: 0.80 TS: 0.81 IoU: Background: 0.83 Canal: 0.61 IVD: 0.91 Other: 0.97 PE: 0.82 TS: 0.81	*quite high and show that a Computer Aided Diagnosis system can be created in LSS diagnosis"*
Gaonkar et al., 2019 [24]	IVD (sagittal), spinal canal (axial)	IVD: part of IVD, not part of IVD Spinal canal: part of spinal canal, not part of spinal canal	209 patients, number of slices (per patient) not explicitly stated	Nonparametric bias correction, linear histogram matching to a common template, intensity normalization to the 0–1 range, linear image registration, resampling to 256x256 pixels	No description	No	No description	Training: 48%; validation: 52%	No	-	Out-of-sample	Dice score	No	Disc segmentation: Dice score: 0.88	*"Machine-generated segmentations were qualitatively comparable with those generated by human experts"* *"Machine generated segmentation agrees almost as well with each human expert as the human experts agree among themselves"*
Han et al., 2018 [28]	IVD, neural foramen	Background, IVD, neural foramen	253 patients, 253 images, 1 sagittal slice per patient	No	Yes, advanced modules with hybrid learning strategy, dynamic optimization algorithm	No	Yes, class imbalances	Training: 80%; test: 20%	Yes	Five-fold cross-validation	Five-fold cross-validation	Pixel accuracy, Dice score	Standard deviation	Pixel accuracy: FCN: 0.917 ± 0.004 SegNet: 0.945 ± 0.002 DeepLabv3+: 0.953 ± 0.001 U-Net: 0.920 ± 0.004 ACAE: 0.958 ± 0.002 ACAE+LSTM: 0.959 ± 0.002 ACAE+CNN: 0.960 ± 0.004 Spine-GAN: 0.962 ± 0.003 Dice score: FCN: 0.754 ± 0.033 SegNet: 0.760 ± 0.032 DeepLabv3+: 0.812 ± 0.021 U-Net: 0.797 ± 0.013	*„Spine-GAN has achieved the pixel accuracy of 96.2%, which demonstrates that our system can provide a very visual presentation for clinical application."*

														ACAE: 0.841 ± 0.013 ACAE+LSTM: 0.848 ± 0.009 ACAE+CNN: 0.863 ± 0.006 Spine-GAN: 0.871 ± 0.004	
Hou et al., 2022 [29]	AAP, dural sac, IVD, lamina	AAP, background, dural sac, IVD, lamina	515 patients, 1545 images, 3 axial slices per patient, 1 for each IVD from L3-4 to L5-S1	Images downsampled	No description	Random flipping and clipping	No description	Training: 75%; validation: 6%; test: 19%	Yes	Out-of-sample	Out-of-sample	Pixel accuracy, intersection-over-union (IoU) (Jaccard index)	No	Pixel accuracy: Spine-Seg-2Phase: 0.9935 IoU: Spine-Seg-2Phase: 0.8493 SegNet-TL80: 0.72	*"Our model can achieve a higher intersection ratio overall and achieve better performance in most categories (4 out of 5 categories) compared with previous models"*
Li et al., 2021 [39]	Dural sac, lamina	Background, dural sac, lamina, vertebral body	120 patients, 1080 images, average of 9 slices per patient, 3 for each IVD from L3-4 to L5-S1	Brightness, contrast adjustment and normalization processing; size of all images was unified to 512 x 512.	No description	No description	Yes, vertebral body, lamina, and dural sac classes are extremely small in comparison to the background	Training: 70%; validation: 20%; test: 10%	Yes	Out-of-sample	Out-of-sample	Dice score, average surface distance (mm)	No	Mean Dice score: FCN: 0.9037 U-Net: 0.9008 Res2net: 0.9140 Deeplabv3 +: 0.9111 U-Net + multi-scale: 0.9159 U-Net + attention: 0.9152 MANet: 0.9252 Mean average surface distance: FCN: 5.41 U-Net: 6.40 Res2net: 3.21 Deeplabv3 +: 4.15 U-Net + multi-scale:	*"Compared with several commonly used medical image segmentation methods, the proposed method has achieved better segmentation results"*

														4.02 U-Net + attention: 3.68 MANet: 2.71	
Natalia et al., 2020 [42]	AAP, IVD, PE, TS	AAP, IVD, PE, TS, other, unregistered	515 patients, 1545 composite images, 3 axial slices per patient	No	No description	Parametric bias field estimation is applied before being corrected using the PABIC method, resizing to 320x320, contour evolution technique applied	Yes, to make sure that small-sized classes, such as the TS and AAP, are not underrepresented in the training data the class weighting was set to be inversely proportional to the class population	Training: 80%; test: 20%	No	-	Out-of-sample	BF-score*	No	BF-Score: IVD boundary: 0.80 PE boundary: 0.87	*"Our experiments show that our method can locate all the important points that are on region boundaries accurately."*
Tang et al., 2020 [45]	Dural sac, spinal canal	Background, dural sac, spinal canal, vertebral body	279 patients, 2393 images, average of 9 axial slices per patient	No	No	Yes: rotation, horizontal flipping, random cropping, random standard Gaussian noise applied	Yes, imbalanced labels	Training: 50%; validation: 20%; test: 30%	Yes	Out-of-sample	Out-of-sample	Pixel accuracy, mean pixel accuracy, mean intersection-over-union (IoUm), frequency weighted intersection-over-union (IoUfw)	No	Pixel accuracy: FCN-8s: 0.9869 U-Net: 0.9804 DeeplabV3: 0.9896 DDU-Net: 0.9913 Mean pixel accuracy: FCN-8s: 0.7588 U-Net: 0.7316 DeeplabV3: 0.8718 DDU-Net: 0.9099 IoUm: FCN-8s: 0.7705 U-Net: 0.7746 DeeplabV3: 0.8053 DDU-Net: 0.8331 IoUfw: FCN-8s: 0.9766 U-Net: 0.9646 DeeplabV3: 0.9802 DDU-Net: 0.9835	*"Our method is precise, and by comparing the segmentation results to those of existing state-of-the-art methods on our new dataset, the proposed method proved superior in terms of segmentation accuracy"*

Conclusions written in italics were quoted from the article. AAP: area between anterior and posterior vertebrae elements; ACAE: atrous convolution autoencoder; CNN: convolutional neural network; DDU-net: dual-decoder-U-net; FCN: fully convolutional network; GAN: generative adversarial network; IVD: intervertebral disc; LSS: lumbar spinal stenosis; LSTM: long short term memory; MRI: magnetic resonance imaging; PABIC: parametric bias field correction; PE: posterior element; ROI: region of interest; SVM: support vector machine; TS: thecal sac.

* Boundary F1 score: a semantic contour-based metric. The authors refer to: Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011; 2(1):37–63

Online Resource 7 – Data extraction for ML algorithms on classification

Study	ROI classified	Classification labels	Sample size	Pre-processing	Model augmentation	Data augmentation	Data imbalance	Split	Validation step applied	Validation method	Testing method	Outcome measures	Uncertainty metrics	Results	Conclusion
Altun et al., 2023 [22]	Canal, IVD, PE, TS	LSS or control (no LSS present), one label for all ROIs	1030 patients, 3090 images, 3 axial slices per patient, 1 for each IVD from L3-4 to L5-S1	ROI extraction	No	No description	No	No description	No	No	10-fold cross-validation	Accuracy, AUC, sensitivity, specificity	No	<u>Accuracy:</u> Gabor-RF: 0.762 Gabor-SVM: 0.701 Sobel-RF: 0.672 Sobel-SVM: 0.661 <u>AUC:</u> Gabor-RF: 0.79 Gabor-SVM: 0.71 Sobel-RF: 0.67 Sobel-SVM: 0.62 <u>Sensitivity:</u> Gabor-RF: 0.721 Gabor-SVM: 0.698 Sobel-RF: 0.693 Sobel-SVM: 0.689 <u>Specificity:</u> Gabor-RF: 0.827 Gabor-SVM: 0.72 Sobel-RF: 0.705 Sobel-SVM: 0.703	*"It is seen that both of the classifications made in the dataset created using the Gabor operator make more accurate predictions".* *"It is the most successful conventional classification method for the Gabor-RF dataset".*
Huber et al., 2019 [30]	Small ROI: dural sac Large ROI: dural sac, epidural fat, bilateral recesses	Severe or non-severe LSS	82 patients, 343 images, 1-5 slices in total per patient, one for each IVD	Rescaling, manual slice extraction	No	No description	No description	No description	Yes	No description	10-fold cross-validation	AUC, NPV, precision, sensitivity, specificity	No	<u>AUC:</u> Small ROI: 0.962 Large ROI: 0.940 <u>NPV:</u> Small ROI: 96.06 Large ROI: 98.04 <u>Precision:</u> Small ROI: 95.00 Large ROI:	*"Diagnostic performance descriptors increased significantly when using TA for detection of severe LSS with sensitivity and specificity of 94% and 98%, respectively when compared to*

														94.32 <u>Sensitivity:</u> Small ROI: 94.33 Large ROI: 94.32 <u>Specificity:</u> Small ROI: 96.53 Large ROI: 98.04	*standard CSA measurements using expert qualitative scores as reference.“*
Koompairojn et al., 2010 [35]	Central canal, lateral recess or neural foramina, ligaments or facet	Central canal: stenosis or not Lateral recess/foraminal: stenosis or not Ligaments or facet: hypertrophy or not	50 patients, number of slices (per patient) not explicitly stated	Downsampling	No	No description	Hypertrophy of ligamentum flavum and facet: moderate imbalance Disc space narrowing: moderate imbalance Lateral spinal stenosis: mild imbalance	Training: 90%; test: 10%	Yes	10-fold cross-validation	10-fold cross-validation	Accuracy	No	<u>Accuracy:</u> CSS: 92.66 % Hypertrophy of ligament flavum & facet: 96.82% Lateral spinal stenosis: 96.29%	*"We have developed a fully automatic computer-aided diagnosis system for lumbar spine stenosis. The average accuracy ranges from about 92% to 97%."*

Conclusions written in italics were quoted from the article. AUC: area under the curve; CSA: cross-sectional area; CSS: central spinal stenosis; IVD: intervertebral disc; LSS: lumbar spinal stenosis; NPV: negative predictive value; PE: posterior element; RF: random forest; ROI: region of interest; SVM: support vector machine; TA: texture analysis; TS: thecal sac.

Online Resource 8 – Data extraction for DL algorithms on classification

Study	ROI classified	Classification labels	Sample size	Pre-processing	Model augmentation	Data augmentation	Data imbalance	Split	Validation step applied	Validation method	Testing method	Outcome measures	Uncertainty metrics	Results	Conclusion
Altun et al., 2023 [22]	Undefined	LSS or control (no LSS present)	1030 patients, 3090 images, 3 axial slices per patient, 1 for each IVD from L3-4 to L5-S1	ROI extracted	No	No description	No	No description	No	No	10-fold cross-validation	Accuracy, AUC, sensitivity, specificity	No	Accuracy: InceptionNet: 0.753 MobileNet: 0.835 ResNet: 0.867 VGG16: 0.877 AUC: InceptionNet: 0.75 MobileNet: 0.83 ResNet: 0.87 VGG16: 0.88 Sensitivity: InceptionNet: 0.704 MobileNet: 0.807 ResNet: 0.893 VGG16: 0.921 Specificity: InceptionNet: 0.84 MobileNet: 0.869 ResNet: 0.9 VGG16: 0.84	*"As a result of the study, the most successful method was VGG16. In this direction, LSS-VGG16 model is proposed for an expert decision support system in LSS diagnosis."*
Grob et al., 2022 [26]	Central canal, IVD	Central canal: normal, mild, moderate, severe stenosis IVD: degree of disc degeneration according to Pfirrmann grade (1,2,3,4,5)	882 patients, number of slices (per patient) not explicitly stated	No	No	No	Yes	-	-	-	External test: 100%	Class average accuracy, Lin's concordance correlation coefficient, precision, sensitivity, specificity, weighted Cohen's kappa	No	Class average accuracy: CSS: 43% Pfirrmann grade: 55% Lin's concordance correlation coefficient: CSS: 0.60 Pfirrmann grade: 0.72 Precision: CSS: 0.43 Pfirrmann grade: 0.55	*"Our findings show that SpineNet V1.1 is a useful and accurate tool for providing automated annotation of sagittal lumbar spine MRIs."*

														Sensitivity: CSS: 0.50 Pfirrmann grade: 0.42 Specificity: CSS: 0.88 Pfirrmann grade: 0.81 Weighted Cohen's kappa: CSS: 0.60 Pfirrmann grade: 0.72	
Hallinan et al., 2021 [27]	Central canal, lateral recesses, neural foramina	Normal, mild, moderate, severe stenosis	Internal dataset: 446 patients, 12403 axial T2 images, average of 28 per patient, 6161 sagittal T1 images, average of 14 per patient External dataset: 100 patients, number of slices (per patient) not explicitly stated	ROI extracted	No	No	Yes	Internal dataset: Training 80%; validation 9%; test 11% External dataset: Test: 100%	Yes	3-fold cross validation	Out-of-sample	Gwet's AC1 for four classes, Gwet's AC1 dichotomized, sensitivity, specificity	95% CI, p-value	**Internal dataset:** Gwet's AC1 (4 classes): Central canal: 0.82 (0.80-0.85) (P < 0.001) Lateral recesses: 0.72 (0.68-0.76) (P < 0.001) Neural foramina: 0.75 (0.71-0.78) (P < 0.001) Gwet's AC1 (dichotomized): Central canal: 0.96 (0.94-0.98) (P < 0.001) Lateral recesses: 0.92 (0.90-0.94) (P < 0.001) Neural foramina: 0.89 (0.86-0.91) (P < 0.001) Sensitivity: Central canal: 90.9% (80.3-96.7%) Lateral recesses: 80.3 (72.4-86.7%) Neural foramina:	*"We demonstrated that our deep learning (DL) model is reliable and may be used to quickly assess lumbar spinal stenosis (LSS) at MRI."*

														91.2% (85.4-95.3%) <u>Specificity:</u> Central canal: 97.0% (96.0-97.8%) Lateral recesses: 96.7% (95.0-97.9%) Neural foramina: 91.9% (90.0-93.6%) **External dataset:** <u>Gwet's AC1 (4 classes):</u> Central canal: 0.68 (0.66-0.71) (P < 0.001) Lateral recesses: 0.77 (0.75-0.80) (P < 0.001) Neural foramina: 0.83 (0.81-0.85) (P < 0.001) <u>Gwet's AC1 (dichotomized):</u> Central canal: 0.95 (0.94-0.97) (P < 0.001) Lateral recesses: 0.96 (0.95-0.97) (P < 0.001) Neural foramina: 0.96 (0.95-0.97) (P < 0.001) <u>Sensitivity:</u> Central canal: 84.3% (72.0-92.8%) Lateral recesses: 92.2% (84.5-96.7%) Neural foramina:	

														73.0% (63.7-81.0%) <u>Specificity:</u> Central canal: 96.3% (95.3-97.2%) Lateral recesses: 96.6% (95.7-97.4%) Neural foramina: 98.3% (97.6-98.9%)	
Han et al., 2018 [28]	IVD, neural foramen	IVD: normal, disc degeneration Neural foramen: normal, foraminal stenosis	253 patients, 253 images, 1 sagittal slice per patient	No	Yes	No	Yes	Training: 80%; test: 20%	Yes	Five-fold cross-validation	Five-fold cross-validation	Sensitivity, specificity	Standard deviation	<u>Sensitivity:</u> IDD: FCN: 0.726 ± 0.039 SegNet: 0.755 ± 0.032 DeepLabv3+: 0.852 ± 0.025 U-Net: 0.811 ± 0.049 ACAE: 0.844 ± 0.027 ACAE+LSTM: 0.854 ± 0.025 ACAE+CNN: 0.869 ± 0.029 Spine-GAN: 0.871 ±0.029 FS: FCN: 0.672 ± 0.029 SegNet: 0.662 ± 0.022 DeepLabv3+: 0.865 ± 0.012 U-Net: 0.769 ± 0.060 ACAE: 0.821 ± 0.027 ACAE+LSTM: 0.829 ± 0.026 ACAE+CNN: 0.856 ± 0.039 Spine-GAN: 0.876 ± 0.029 <u>Specificity:</u> IDD:	*„Spine-GAN achieved accurate segmentation, radiological classification, and pathological correlations representation of three types of spinal diseases. (…) Spine-GAN has also achieved 89.1% specificity and 86% sensitivity on both three types of spinal structures simultaneously, which demonstrates that our system is capable of assisting clinicians to improve the diagnosis efficiency.“*

														FCN: 0.638 ± 0.072 SegNet: 0.731 ± 0.012 DeepLabv3+: 0.717 ± 0.027 U-Net: 0.746 ± 0.031 ACAE: 0.810 ± 0.039 ACAE+LSTM: 0.818 ± 0.055 ACAE+CNN: 0.804 ± 0.028 Spine-GAN: 0.844 ± 0.063 FS: FCN: 0.745 ± 0.041 SegNet: 0.746 ± 0.015 DeepLabv3+: 0.786 ± 0.035 U-Net: 0.814 ± 0.057 ACAE: 0.867 ± 0.032 ACAE+LSTM: 0.870 ± 0.032 ACAE+CNN: 0.893 ± 0.020 Spine-GAN: 0.907 ±0.047	
Ishimoto et al., 2020 [31]	Central canal	None, mild, moderate, severe stenosis	971 patients, 4855 images, 5 axial slices per patient, 1 for each IVD from L1-2 to L5-S1	No	No	No description	Yes: underrepresentation of class 0 and 3	Training: 90%; test: 10%	No description	No description	Out-of-sample	Cohen's kappa, Lin's concordance correlation coefficient	No	Cohen's kappa: 0.75 Lin's concordance correlation coefficient: 0.73	*"We have shown that MRI grading of central LSS can be predicted with a high degree of reliability and consistency after a period of learning of the reference standard."*
Jamaludin et al., 2017 (1) [32]	Central canal, IVD	Central canal: normal, narrow/constricted canal	2009 patients, 12,018 images,	ROI extraction	No	No description	Yes	Training: 90%; test: 10%	No description	No description	Out-of-sample	Class average accuracy, Cohen's	No	<u>Class average accuracy:</u> CSS: 94.7%	*"Our system excelled at determining Pfirrmann*

		IVD: Pfirrmann grade (5 grades), disc narrowing (normal, slight, moderate, severe)	number of slices per patient not explicitly stated									kappa, Lin's concordance correlation coefficient		Disc narrowing: 75.4% Pfirrmann grade: 70.1% <u>Cohen's kappa:</u> CSS: 0.52 <u>Lin's concordance correlation coefficient:</u> Disc narrowing: 0.89 Pfirrmann grade: 0.88	*and disc narrowing grades on a relevant data set of clinical images, both in terms of accuracies and reliability scores (Kappa values)."*
Jamaludin et al., 2017 (2) [33]	Central canal, IVD	Central canal: normal, narrow/constricted canal IVD: Pfirrmann grade (5 grades), disc narrowing (normal, slight, moderate, severe)	2009 patients, 12,018 images, number of slices per patient not explicitly stated	No	No	Rotation, translation, rescaling, intensity variation, random slice-wise flip across the mid-sagittal	Yes, class-balanced loss, class-specific weights, reweight the loss	Training: 80%; validation: 10%; test: 10%	Yes	Two-fold cross-validation	Two-fold cross-validation	Class average accuracy	Standard deviation	<u>Class average accuracy:</u> CSS: 94.3 ± 0.0 Disc narrowing: 75.9 ± 0.4 Pfirrmann grade: 71.9 ± 1.5	*"Training with multiple tasks in general improves over the performance of training on in- dividual tasks. Furthermore, we have shown that we can also pro- duce high quality visualizations of pathology hotspots that can aid understanding of the predictions of models trained only on weak class labels."*
Kim et al., 2022 [34]	Undefined	LSS or no stenosis	Internal dataset: 4,644 patients, 12,442 images, 3 sagittal images per patient (neutral, flexion, extension)	Images resized, histogram equalization, images rescaled	No description	Zoom, rotation, width and height shift augmentation	No description	Internal dataset: Training: 60%; validation: 20%; test: 20% External dataset: Test: 100%	Five-fold cross-validation	Five-fold cross-validation	Five-fold cross-validation	Accuracy, AUROC, NPV, precision, sensitivity, specificity		**Internal dataset:** <u>Accuracy:</u> VGG19: 0.827-0.836 VGG16: 0.819-0.824 ResNet50: 0.798-0.809 Efficient1: 0.785-0.802 <u>AUROC:</u> VGG19: 0.898-0.901	"Our proposed models showed acceptable accuracy with high sensitivity and specificity values" "This study suggests the possibility of

			External datatset : 199 patients, 3 sagittal images per patient (neutral, flexion, extension)											VGG16: 0.888-0.896 ResNet50: 0.871-0.883 Efficient1: 0.856-0.876 <u>NPV:</u> VGG19: 0.818-0.831 VGG16: 0.806-0.816 ResNet50: 0.786-0.809 Efficient1: 0.790-0.802 <u>Precision:</u> VGG19: 0.836-0.841 VGG16: 0.822-0.845 ResNet50: 0.809-0.815 Efficient1: 0.769-0.812 <u>Sensitivity:</u> VGG19: 0.810-0.826 VGG16: 0.792-0.812 ResNet50: 0.773-0.806 Efficient1: 0.782-0.810 <u>Specificity:</u> VGG19: 0.842-0.846 VGG16: 0.825-0.856 ResNet50: 0.811-0.824 Efficient1: 0.759-0.820 **External dataset:** <u>Accuracy:</u> VGG19: 0.815-0.819 VGG16: 0.819-0.824 ResNet50: 0.799-0.807 Efficient1: 0.758-0.793 <u>AUROC:</u> VGG19: 0.883-0.895 VGG16:	utilizing lateral radiographs with a CNN as a screening test for diagnosing LSS"

														0.889-0.892 ResNet50: 0.878-0.883 Efficient1: 0.853-0.876 <u>NPV:</u> VGG19: 0.846-0.849 VGG16: 0.836-0.856 ResNet50: 0.828-0.846 Efficient1: 0.806-0.826 <u>Precision:</u> VGG19: 0.790-0.796 VGG16: 0.790-0.813 ResNet50: 0.774-0.785 Efficient1: 0.722-0.770 <u>Sensitivity:</u> VGG19: 0.857-0.859 VGG16: 0.840-0.869 ResNet50: 0.840-0.861 Efficient1: 0.7832-0.851 <u>Specificity:</u> VGG19: 0.774-0.782 VGG16: 0.770-0.808 ResNet50: 0.750-0.766 Efficient1: 0.682-0.754	
Lehnen et al., 2021 [36]	IVD, nerve root, spinal canal	IVD: disc bulging or herniation, or not Nerve root: compression or not Spinal canal: none, relative or absolute stenosis	146 patients, number of slices (per patient) not explicitly stated	No	No description	Yes	Yes, severe data imbalance for spinal canal stenosis	Test: 100%	-	-	Out-of-sample	Accuracy, NPV, precision, sensitivity, specificity	No	Accuracy: 98.09% NPV: 99.06% Precision: 75.00% Sensitivity: 77.14% Specificity: 98.95%	*"Our major findings revealed that, in a patient cohort with a wide age range, the CNN was highly consistent with the radiologist's expert reading and yielded moderate to high*

															diagnostic sensitivities and specificities for the detection of lumbar degenerative changes ranging from 52–89% and 80–99%, respectively" *"Our results showed that the CNN was capable of processing lumbar MRI acquired independently of the training and test datasets at a different institution, using different MR scanners, different field strengths, and varying sequence parameters, thereby still providing fair to high diagnostic accuracy"*
Lewandrowski et al., 2020 (1) [37]	Central canal, IVD, neural foramina	Central canal: no stenosis, disc bulge without stenosis, disc bulge and stenosis, herniation and stenosis IVD: normal, bulging or herniation Neural foramina: stenosis or not	3,560 patients, 17,800 images, number of slices per patient not explicitly stated	Extract the disc regions from MRI image	Spatial dropouts to minimize overfitting	No description	Mild to severe class imbalance	Training: 80%; test: 20%	No description	No description	Out-of-sample	Accuracy, sensitivity, specificity	No	Accuracy: CSS: 86.2% Disc bulging: 84.9% FS: 81.0% Sensitivity: CSS: 91.1% Disc bulging: 89.4% FS: 72.4% Specificity: CSS: 82.5% Disc bulging: 81.2% FS: 83.1%	*"The model report generated by the network models (the RadBot) was capable of producing a verbally spelled-out report by a specific lumbar spinal level comparable in detail as to what is typically seen by the radiologist in terms of*

															detail and scope"
Lewandrowski et al., 2020 (2) [38]	Central canal, IVD, neural foramina	IVD/central canal: no disc bulge/protrusion/canal stenosis, disc bulge without canal stenosis, disc bulge resulting in canal stenosis, disc herniation/ protrusion/extrusion resulting in canal stenosis Neural foramina: stenosis absent or present	65 patients, 382 images, number of slices per patient not explicitly stated	Extract the disc regions from MRI image	Spatial dropouts to minimize overfitting	No description	Mild class imbalance	Test: 100%	-	-	Out-of-sample	AUC, Cohen's kappa, NPV, precision, sensitivity, specificity	95% CI	AUC: 0.808 (95% CI: 0.760; 0.857) Cohen's kappa: 0.627 NPV: 83.7% Precision: 80.3% Sensitivity: 73.3% Specificity: 88.4%	*"The numbers obtained with these methods suggest that the our AI deep learning network as a diagnostic tool has excellent performance characteristics"*
Lin et al., 2021 [40]	Dural sac measurements (mm and mm^2)	-	143 patients, 895 images, 6 ± 2 axial images per patient	ROI cropping	No	No	No description	Training: 80%; test: 20%	No description	No description	Five-fold cross-validation	Mean absolute error in dural sac cross-sectional area (MAE-DSCA) (mm^2)	No	<u>MAE-DSCA:</u> CARN: 0.96 DE-Net: 0.64 HOG+AKRF: 1.6 HOG+SSVR: 1.28 MF+RF: 1.6 OSBP-Net: 0.32	*"All deep learning-based methods gain results better than the conventional machine learning-based methods. Overall, the proposed OSBP-Net obtains the smallest errors for all indices among all competing methods, which demonstrates the superiority of OSBP-Net against other models."*
Lu et al., 2018 [41]	Neural foramina, spinal canal	Normal, mild, moderate, severe stenosis	4075 patients, number of slices (per patient) not explicitly stated	ROI extraction	Dropout regularization with a dropout rate of 0.2 on the fully connected layers	No description	Moderate to severe class imbalance	Training: 70%; validation: 15%; test: 15%	Yes	Out-of-sample	Out-of-sample	AUC, class average accuracy	Standard deviation	<u>AUC:</u> FS: 0.961 (95% CI: 0.955-0.967) Spinal canal stenosis: 0.983 (95% CI: 0.971-0.992) <u>Class average accuracy</u>	*"This approach enabled us to achieve a high level of performance across the range of stenosis grading classifications for both central*

														(four classes): FS: 67.1 ± 2.2 Spinal canal stenosis: 70.6 ± 2.1 Class average accuracy (three classes): FS: 78.1 ± 0.4 Spinal canal stenosis: 80.4 ± 1.6	*spinal and foraminal locations at each spinal level."*
Pang et al., 2023 [43]	AP diameter and foraminal width measurements (mm)	-	515 patients, 1545 images, 3 axial slices per patient, 1 for each IVD from L3-4 to L5-S1	Resizing, cropping, intensity normalization on preprocessing	No	Random rotation and cropping	No	Training: 70%; validation: 10%; test: 20%	Yes	Out-of-sample	Out-of-sample	Mean absolute error, Pearson correlation coefficient	P-value	MAE: CARN: 0.56* DENet: 0.59* HourglassNet-2-1: 0.87* HourglassNet-4-1: 0.86* HourglassNet-8-1: 0.87* HourglassNet-4-2: 0.85* OSBPNet: 0.55* SGRNet: 0.49 VGG-11: 0.54* Pearson correlation coefficient: CARN: 0.941 DENet: 0.935 HourglassNet-2-1: 0.881 HourglassNet-4-1: 0.881 HourglassNet-8-1: 0.881 HourglassNet-4-2: 0.885 OSBPNet: 0.942 SGRNet: 0.956 VGG-11: 0.944	*"We have presented an accurate SGRNet to achieve spine indices estimation for axial MR images."*
Su et al., 2022 [44]	Central canal, nerve root	Central canal: stenosis grade 0, 1, 2, 3 Nerve root: compression grade 0, 1, 2, 3	Internal dataset: 1015 patients, 15254 images, average of 15 axial per patient	No	No description	Random rotation, cropping	LCCS: severe class imbalance LNRC: severe class imbalance	Internal dataset: Training: 70%; validation: 15%; test: 15% External dataset:	Yes	Out-of-sample	Out-of-sample	Accuracy, AUC, F1 score, Gwet's AC1, precision, sensitivity, specificity	95% CI for Gwet's AC1	**Internal dataset:** Accuracy: LCCS: 87.0 LNRC: 81.2 AUC: LCCS: 0.98 LNRC: 0.95 F1 score:	*"The model proposed in our study achieves good performance for grading (…) LCCS and LNRC at axial MRIs."*

			External dataset: 100 patients, 1273 images, average of 13 per patient					Test: 100%						LCCS: 86.8 LNRC: 80.0 <u>Gwet's AC1:</u> LCCS: 0.86 (95% CI: 0.84-0.87) LNRC: 0.78 (95% CI: 0.76-0.80) <u>Precision:</u> LCCS: 87.0 LNRC: 79.7 <u>Sensitivity:</u> LCCS: 65.2 LNRC: 79.2 <u>Specificity:</u> LCCS: 94.5 LNRC: 92.9 **External dataset:** <u>Accuracy:</u> LCCS: 79.7 LNRC: 74.2 <u>AUC:</u> LCCS: 0.98 LNRC: 0.87 <u>F1 score:</u> LCCS: 83.3 LNRC: 73.4 <u>Gwet's AC1:</u> LCCS: 0.77 (95% CI: 0.75-0.80) LNRC: 0.69 (95% CI: 0.66-0.72) <u>Precision:</u> LCCS: 88.6 LNRC: 74.1 <u>Sensitivity:</u> LCCS: 94.6 LNRC: 68.8 <u>Specificity:</u> LCCS: 86.5 LNRC: 87.6	

Won et al., 2020 [46]	Spinal canal	Stenosis according to Schizas scale (A, B, C, D)	542 patients, 542 images, 1 axial L4-5 slice per patient	No	Global average pooling to alleviate overfitting problem	Random rotation, scaling, flipping, translations	Moderate class imbalance	Training: 80%; validation: 10%; test: 10%	Yes	10-fold cross-validation	10-fold cross-validation	Accuracy, F1 score	P-value	FA model vs. surgeon 1: Accuracy: 83.0 F1 score: 75.4 P-value: 0.8133 FA model vs. surgeon 2: Accuracy: 77.9 F1 score: 74.9 P-value: 0.4420 SA model vs. surgeon 1: Accuracy: 82.8 F1 score: 75.0 P-value: 0.6388 SA model vs. surgeon 2: Accuracy: 78.4 F1 score: 75.6 P-value: 0.3903	*"We achieved comparable performance with previous deep learning based methods for spinal stenosis grading even though direct comparison of the methods on different data sets is difficult."*

Conclusions written in italics were quoted from the article. ACAE: atrous convolution autoencoder; AI: artificial intelligence; AKRF: adjusted K-clustered random forest; AP: anteroposterior; AUC: area under the curve; AUROC: area under the receiver operating characteristic curve; CARN: cascade amplifier network with linear regression model; CI: confidence interval; CNN: convolutional neural network; CSS: central spinal stenosis; DL: deep learning; FCN: fully convolutional network; FS: foraminal stenosis; GAN: generative adversarial network; HOG: histogram of oriented gradients; IDD: intervertebral disc degeneration; IVD: intervertebral disc; LCCS: lumbar central canal stenosis; LNRC: lumbar nerve roots compression; LSS: lumbar spinal stenosis; LSTM: long short term memory; MAE-d_{ap}: mean absolute error in anteroposterior diameter; MAE-DSCA: mean absolute error in dural sac cross-sectional area; MF: multiple features; ML: machine learning; MRI: magnetic resonance imaging; NPV: negative predictive value; OSBP: object-specific bi-path; PE: posterior element; RF: random forest; ROI: region of interest; SGAM: segmentation-guided attention module; SGR: segmentation-guided regression; SSVR: structured support vector regression; TS: thecal sac.

* $p < 0.001$