Optimizing care in lumbar radiculopathy and neurogenic claudication: from injection to inference, and from clinician to algorithm

Verheijen, E.J.A.

10

Artificial intelligence for segmentation and classification in lumbar spinal stenosis: an overview of current methods

E.J.A. Verheijen, T. Kapogiannis, D. Munteh, J. Chabros,
M. Staring, T.R. Smith, C.L.A. Vleggeert-Lankamp

Eur Spine J 2025

ABSTRACT

Purpose

Lumbar spinal stenosis (LSS) is a frequently occurring condition defined by narrowing of the spinal or nerve root canal due to degenerative changes. Physicians use MRI scans to determine the severity of stenosis, occasionally complementing it with X-ray or CT scans during the diagnostic work-up. However, manual grading of stenosis is time-consuming and induces inter-reader variability as a standardized grading system is lacking. Machine Learning (ML) has the potential to aid physicians in this process by automating segmentation and classification of LSS. However, it is unclear what models currently exist to perform these tasks.

Methods

A systematic review of literature was performed by searching the Cochrane Library, Embase, Emcare, PubMed, and Web of Science databases for studies describing an ML-based algorithm to perform segmentation or classification of the lumbar spine for LSS. Risk of bias was assessed through an adjusted version of the Newcastle-Ottawa Quality Assessment Scale that was more applicable to ML studies. Qualitative analyses were performed based on type of algorithm (conventional ML or Deep Learning (DL)) and task (segmentation or classification).

Results

A total of 27 articles were included of which nine on segmentation, 16 on classification and 2 on both tasks. The majority of studies focused on algorithms for MRI analysis. There was wide variety among the outcome measures used to express model performance. Overall, ML algorithms are able to perform segmentation and classification tasks excellently. DL methods tend to demonstrate better performance than conventional ML models. For segmentation the best performing DL models were U-Net based. For classification U-Net and unspecified CNNs powered the models that performed the best for the majority of outcome metrics. The number of models with external validation was limited.

Conclusion

DL models achieve excellent performance for segmentation and classification tasks for LSS, outperforming conventional ML algorithms. However, comparisons between studies are challenging due to the variety in outcome measures and test datasets. Future studies should focus on the classification task using DL models and utilize a standardized set of outcome measures and publicly available test dataset to express model performance. In addition, these models need to be externally validated to assess generalizability.

INTRODUCTION

Lumbar spinal stenosis (LSS) is a disease defined by narrowing of the spinal or nerve root canal that becomes symptomatic through the compression of neural structures [1]. Classically, LSS causes intermittent neurogenic claudication affecting approximately 11% of older adults in the US and is the most common cause of spinal surgery among this population [2,3,4].

Patients are usually offered surgery if conservative treatment has failed to sufficiently ameliorate symptoms. Surgical candidacy is assessed using radiological imaging in conjunction with clinical history and physical examination [4, 5]. MRI has become the gold standard to determine the severity of stenosis, as it produces detailed images of relevant soft tissues that may contribute to the stenosis [6, 7]. Alternatively, Computed Tomography (CT) may be used in assessing the bony component of stenosis, although delivering less valuable information on true compression of the cauda equina or spinal nerve root. Evaluation of the severity of stenosis can be subjective, with inter-reader variability among radiologists and surgeons [8, 9]. Grading systems to standardize MRI interpretation for the severity of stenosis, such as those proposed by Lee et al. [10], Schizas et al. [11], and Miskin et al. [12], have demonstrated inter-observer metrics ranging from "fair" to "excellent reliability" (Cohen's kappa 0.323–0.702, intraclass correlation coefficient 0.730–0.953), and, hence, have not eliminated variability. In addition, manually assessing MRIs using these grading systems is time-consuming and, thus, not feasible in clinical practice.

Artificial Intelligence (AI) may be a valuable tool to assist clinicians in grading LSS, as it has demonstrated the ability to assess medical images accurately and consistently in other disease areas [13]. Conventional machine learning (ML) or deep learning (DL) architectures can be trained for image analysis either through supervised or through unsupervised learning. In supervised learning, training images are labeled, and this technique is often used for segmentation and outcome prediction [14]. In practice, semi-supervised and weakly-supervised approaches are more common, especially when high-quality labeled data is scarce. Semi-supervised learning combines a limited amount of labeled data with a large amount of unlabeled data, whereas weakly-supervised learning relies on imperfect or imprecise labels when accurate labeling is challenging or costly. In contrast, unsupervised learning models can detect patterns in unlabeled data, which is valuable as image datasets with high-quality labeling are difficult to procure [13]. The conventional ML approach to image analysis usually necessitates the selection of relevant input features to train the algorithm to complete two tasks: segmentation and classification. In segmentation, each

pixel is assigned to a class based on its extracted attributes (Fig. 1) and are then used as inputs for classification, where predictions are generated on the severity of LSS [15]. DL models are a subset of ML that can learn important features from the raw data, obviating the need for extensive feature engineering, do not require the segmentation step before classification and have demonstrated stronger performance than conventional ML before [16, 17] (Fig. 2).

In this systematic review, we describe current conventional ML and DL models for segmentation and classification of LSS, including scoring of their performance. We aim to examine whether AI can be used to improve LSS diagnostics.

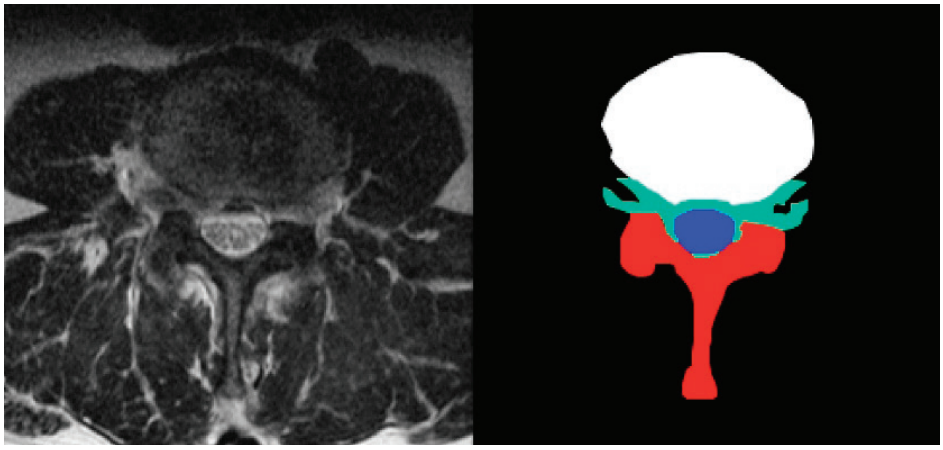

Fig. 1 Example of segmentation of spinal structures in the lumbar spine from an axial T2-weighted MRI image

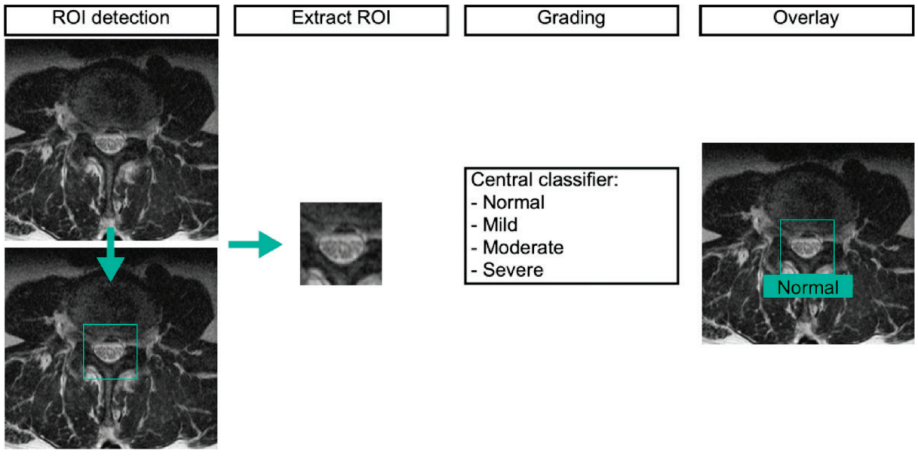

Fig. 2 Example of classification of central spinal stenosis. The algorithm determines the region of interest (ROI) on an axial T2-weighted MRI image, extracts it and decides on the grading of stenosis

METHODS

This systematic review was conducted in accordance with the PRISMA guidelines.

Search and selection

Relevant articles were searched in five databases (Cochrane Library, Embase, Emcare, PubMed, and Web of Science) from inception to 9 February 2023. An expert librarian created a comprehensive search strategy that included strings for studies investigating new or validating existing algorithms for segmentation or classification of LSS (Online Resource 1). All search results were screened by two reviewers (EV and JC) separately based on title and abstract. The remaining full texts were evaluated, and, consequently, screening of references and citation tracking were performed. Any discrepancies were resolved by discussion or consulting a third reviewer (CVL).

Inclusion and exclusion criteria

Studies describing segmentation or classification algorithms for LSS in adults based on conventional ML or DL approaches were considered for eligibility. Segmentation was defined as the ability to label spinal structures on a pixel-level. The placement of a bounding box to extract a region of interest (ROI) was not considered segmentation. For classification, studies were accepted that categorized patients according to severity of stenosis (either binary or multiple classes) or that automated measuring spine indices relevant for classification of LSS. Meeting abstracts, case reports, systematic reviews and meta-analyses were excluded. Only algorithms developed for routine X-ray, CT or MRI were accepted (e.g., excluding non-routine MRI myelography scans). In addition, segmentation studies were required to assess at least two anatomical structures relevant for LSS (e.g., IVD, spinal canal, lateral recess), since LSS is considered a multifactorial disorder. This was also a criterion for classification studies that did not directly classify the degree of stenosis but rather, e.g., degree of disc degeneration and hypertrophy of the ligaments. Only articles written in English and available in full text were included. If an article described a continuation or improvement of previous work by the same author(s), only the most complete work was included.

The aim is to develop a model with the ability to correctly identify and predict classes (discriminate between two or more conditions) with high (spatial/geometric) accuracy and consistent with expert knowledge. Therefore, we defined acceptable outcome measures as those quantifying accuracy (e.g., overall accuracy, F1 score, area under the curve (AUC)), spatial/geometric reliability/error

(e.g., area, Hausdorff distance), and similarity coefficients quantifying agreement with ground truth (e.g., Cohen's kappa, Jaccard index, Dice coefficient).

Risk-of-bias assessment

Risk of bias was assessed by three reviewers separately (EV, TK and DM) using an adjusted version of the Newcastle–Ottawa Quality Assessment Scale [18]. At the time of assessment, an AI-specific risk-of-bias tool had not been developed, although the authors were aware of the work in progress on this matter[19]. The risk-of-bias criteria were adopted to better fit our research aim and be more applicable to studies evaluating computer algorithms (Online Resource 2). Each study could be awarded 0 to 10 points. Studies with a score above seven points were classified as low risk of bias, studies with 5–7 points as intermediate risk of bias and studies with fewer than 5 points as high risk of bias. Differences between the reviewers were resolved during a consensus meeting or with a fourth reviewer (CVL).

Data extraction and analysis

From all included studies, data was collected by two reviewers (TK and DM) on: year of publication, radiological scans (modality, slice orientation, scanning parameters and field strength if applicable), algorithm (type of model, architecture, use of transfer learning, type of loss function and optimization, degree of automation), study comparisons, ground truth, data handling (sample size, preprocessing, augmentation, imbalance, split), validation and testing, outcome measures and results. A third reviewer (EV) verified the final data extraction sheet. In cases where authors presented results for different versions of the same model architecture (e.g., different loss function, varying hyperparameters) or with different thresholds (e.g., varying distance error tolerance), only data were collected for the average of the models or, if the average was not provided, the best performing algorithm, or where the strictest outcome criteria were applied. The heterogeneity among outcome measures precluded pooling of the results, and, therefore, a qualitative analysis was performed. Articles were compared in four categories: conventional ML segmentation, DL segmentation, conventional ML classification and DL classification. Within a category, results were compared that belonged to the same type of outcome measure (i.e., measure of accuracy, spatial/geometric reliability/error or similarity metrics).

RESULTS

Article selection

The initial literature search yielded a total of 661 unique articles. Of those, 616 were excluded and the remaining 29 articles were selected for full-text screening of which 22 were included. After citation tracking an additional 16 studies were screened in full text of which five were accepted. Ultimately, 27 studies were included for this review (Fig. 3) [20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46]. A comprehensive overview of the included studies is provided in Online Resource 3.

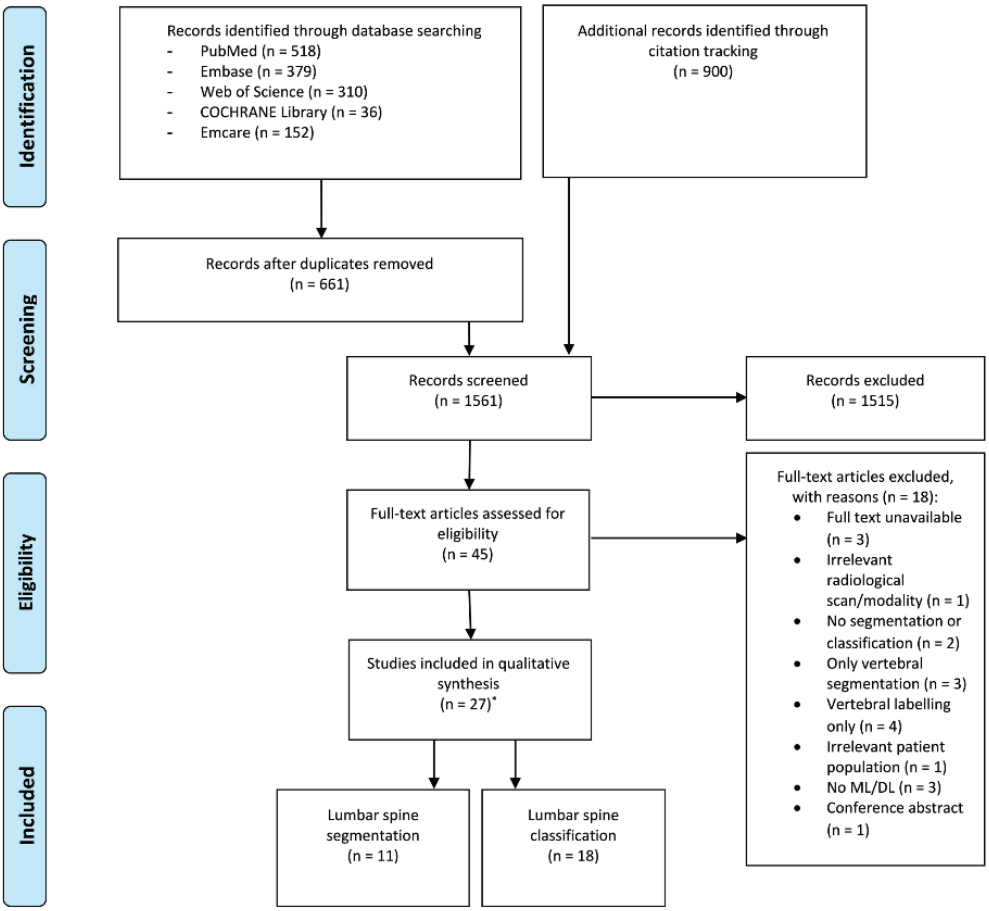

Fig. 3 Flowchart of the article search and selection process.

* Two studies reported on both segmentation and classification and, therefore, appear in both segmentation and classification boxes

Nine articles assessed segmentation, 16 articles assessed classification and two reported on both. Year of publication ranged between 2010 and 2023 with most of the studies being published in 2020 or thereafter. Of the 20 studies published between 2019 and 2023, seven reported on segmentation and 13 reported on classification, whereas, of the six studies published between 2014 and 2018, four reported on segmentation and three reported on classification, demonstrating a shift of focus towards classification challenges (Fig. 4).

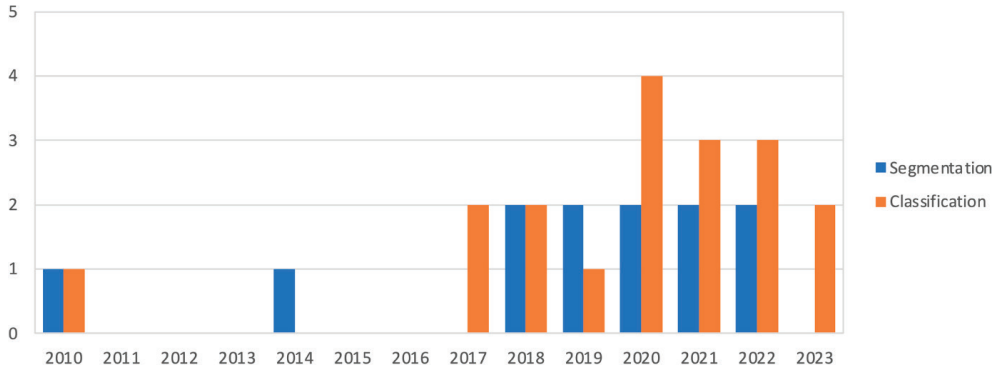

Fig. 4 Number of publications per year focusing on ML/DL algorithms for segmentation or classification of the lumbar spine for LSS. Publications that assessed both appear twice in this graph

Risk-of-bias assessment

Of the 27 studies, three were considered low-risk of bias [26, 27, 31], nineteen were categorized as intermediate-risk of bias [20,21,22, 24, 28, 30, 32,33, 34, 36,37,38,39,40,41,42,43,44, 46] and five were judged as high-risk of bias [23, 25, 29, 35, 45] (Online Resource 4). Risk of bias was higher for segmentation than classification studies on average, but independent of year of publication.

Segmentation algorithms

Segmentation of spinal structures included the IVD, spinal canal, thecal sac (TS), posterior element (PE)/lamina, ligamentum flavum, facet joints, neural foramina, and the area between anterior and posterior vertebrae elements (AAP). Two out of eleven studies employed conventional ML methods, eight used DL, and one applied both techniques. The extracted data are provided in Online Resource 5 and 6.

Conventional machine learning for segmentation

A variety of outcome measures were used to express model performance. None used accuracy, but one study described precision (0.79–0.83) and sensitivity

(0.90–0.92) for disc and dural sac segmentation [25] and another study devised an own metric to express segmentation quality (91.25–98.21%) [35]. Only one study presented spatial metrics reporting Hausdorff distances (7.89–9.41 mm) and surface distances (0.83–0.84) [24]. Two studies presented similarity metrics: one study demonstrated Dice scores ranging 0.83–0.84 [24], while another reported Dice scores of 0.84–0.87 and a Jaccard index between 0.73 and 0.78 [25]. Since Koompairojn et al. received a relatively high risk-of-bias score, the model by Ghosh et al. was considered most reliable [25]. This fully automatic model comprised Histogram of Oriented Gradients (HOG) feature descriptors and random forests (RF) as the classifier with a fixed number of trees (n=100) based on T2-weighted sagittal MRI scans from 50 patients. They used fivefold cross validation and tested on 212 patients.

Deep learning for segmentation

Eight studies performed segmentation using MRI [20, 21, 23, 24, 28, 29, 39, 42] and one used CT scans [45]. Five studies were U-Net based models [23, 24, 29, 39, 45], two studies were SegNet based [20, 42], one study was Generative Adversarial Network (GAN) based [28] and one study tested a standard convolutional neural network (CNN) [21].

An overall accuracy of 85% was reported in one study [47], but a range between 50 and 99% for different spinal structures in another [20]. Four studies provided pixel accuracy with Hou et al. demonstrating the highest accuracy (0.9935) [29]. Other accuracy metrics were only measured in a single study. Only one article reported a geometrical metric achieving a surface distance of 2.71 mm [39]. The most commonly reported performance metrics were similarity coefficients. The Jaccard index was used in five reports with a best overall score of 0.8493 [45], although higher scores were demonstrated for specific spinal structures [20]. When correcting for class frequency, an even higher score was achieved (0.9835) [45]. The Dice coefficient was reported in four studies with a maximum value of 0.9252 [39]. Due to the variety in reported performance metrics it was not possible to determine a single most reliable DL model for segmentation. However, considering (pixel) accuracy, Jaccard index and Dice score, the best performing models were U-Net based [29, 39, 45] or SegNet based [20]. All four compared their model's performance to human expertise using data of at least 120 patients and evaluated their own model (SegNet-TL80 [20], Spine-Seg-2 phase model [29], MANet [39] and DDU-Net [45]. Two were semi-automated [29, 39] and two were fully automated [20, 45].

Classification algorithms

A total of eighteen articles assessing classification algorithms for LSS were published between 2010 and 2023 [22, 26,27,28, 30,31,32,33,34,35,36,37,38, 40, 41, 43, 44, 46]. Two out of eighteen studies described conventional ML algorithms, fifteen used DL and one compared both techniques. The extracted data are provided in Online Resource 7 and 8.

Conventional machine learning for classification

Of the three studies, Koompairojn et al. demonstrated an overall accuracy of 92.66–96.82% for different spinal features [35]. Altun et al. achieved the highest accuracy (0.762) with a Gabor-RF model using one axial image for each of three IVD levels (L3-4 to L5-S1) [22]. Huber et al. evaluated performance through sensitivity, specificity and AUC, and obtained superior results with a decision tree algorithm that used 240 texture analysis features from the dural sac or spinal canal as input (sensitivity: 94.32–94.33%, specificity: 96.53–98.04%, AUC: 0.940–0.962) [30]. The ground truth was established using a reference standard for the cross-sectional area, defined as above 130 mm^2 (non-severe stenosis) or below (severe stenosis). A total of 343 images from 82 patients (max. 5 axial slices per patient, one per IVD level) were used with tenfold cross validation. This model was considered semi-automated since it required manually segmented MRI images.

Deep learning for classification

Sixteen studies described DL models for classification [22, 26,27,28, 31,32,33,34, 36,37,38, 40, 41, 43, 44, 46]. Six studies employed an unspecified CNN [26, 27, 31,32,33, 40], five studies were VGG-based [22, 34, 37, 38, 46], two studies were U-Net based [36, 43], two studies were ResNet-based [41, 44], and one study used a GAN [28]. Models either classified specific spinal features with binary (stenosis or not) or multiclass labels (e.g., normal, mild, moderate, severe stenosis), performed measurements between spinal structures, or classified the image as a whole (binary label).

Out of six studies reporting on accuracy, Lehnen et al. achieved the highest score (98.09%) [36]. Class average accuracy was reported in four studies, with the highest range of values achieved by Jamaludin et al. (0.701–0.947) [32]. AUC was reported the highest by Lu et al. (0.961–0.983) among five studies [41]. Sensitivity and specificity were presented in 9 studies: the highest overall sensitivity was achieved by Altun et al. (0.921) [22], although two studies reported higher values for central canal stenosis specifically (0.922 and 0.946) [27, 44].

Lehnen et al. reported the highest specificity (0.9865) [36]. Negative (NPV) and positive predictive value (PPV) were only reported in three and five studies, respectively. Lehnen et al. achieved an NPV of 0.9906, whereas PPV was highest in the study by Lewandrowski et al. [38], although Kim et al. achieved values between 0.790 and 0.845 depending on the use of neutral, flexion or extension radiographs [34]. F1 score was highest in the study by Su et al., although this was only reported by one other study [44]. Among spatial metrics only the mean absolute error (MAE) for the diameter of the dural sac was reported in two studies with the smallest error obtained by Pang et al. (0.72 mm) [40, 43]. For similarity metrics Cohen's kappa was reported in four studies but highest in the study by Ishimoto et al. (0.75) [31]. Gwet's AC1 statistic, reported in two studies, was highest in the study by Hallinan et al. (0.68–0.96) [27]. In addition, Lin's correlation coefficient was presented in three studies with the best performance reported by Jamaludin et al. (0.88–0.89) [32].

Similar to the results for segmentation models, there was a wide variation in the performance metrics for classification. Moreover, some models achieved high scores for one metric, but lower for other outcome measures. As a result, nine different studies achieved the best score depending on the performance metric. Of the twelve metrics, the best result for four parameters was achieved using U-Net based models [36, 43], for four others unspecified CNN algorithms worked best [27, 31, 32], for two ResNet derived models obtained the best results [41, 44] and for another two metrics VGG-like models performed the best [22, 37]. Two studies used multiple axial and sagittal images (axial: 15–28 slices, sagittal: 15 slices [27, 44], three studies used only one image for each IVD level [22, 31, 43], and 4 studies did not specify the number of slices per patient or per IVD level [32, 36, 37, 41]. Hence, it was not possible to identify one DL algorithm as the most reliable model for the classification task.

DISCUSSION

Through this systematic review, an overview is presented of the currently developed conventional ML and DL techniques for segmentation and classification tasks of LSS. It has been demonstrated that there is a wide spectrum of models for these aims, mostly based on MRI scans. In addition, there is a preponderance of DL models among the studies that were published more recently. Although the paucity of comparable performance metrics posed a challenge for comparisons, DL models, especially U-Net based, yielded better results than conventional ML algorithms for the segmentation task. For classification purposes, DL networks were superior to conventional ML solutions, specifically U-Net and

'standard' CNN algorithms performed the best for most of the performance metrics, obviating the need for a preceding segmentation step. Overall, the models demonstrated results nearing human performance which holds promise to support physicians in the diagnosis of LSS.

LSS can occur centrally around the thecal sac, in the lateral recess or at the neural foramen. Therefore, the ideal classification model would include all these structures. Three out of the nine best performing DL classification algorithms labelled the central canal, lateral recess and neural foramen (or lateral recess and neural foramen combined as 'nerve root') for stenosis [27, 36, 44] and one model labelled the whole image [22]. In general, central and lateral recess stenosis are best assessed using axial MRI images, whereas the neural foramina can be most optimally viewed on sagittal scans. Since the studies by Hallinan et al. and Lehnen et al. included both planes, they can be considered the most comprehensive models for classification of LSS [27, 36], and, as a result, hold the most clinical value. However, both models have only been externally validated to a limited extent. Hallinan et al. tested their model on an external dataset of 100 patients from Saudi Arabia following training and testing with patient data from a hospital in Singapore. Lehnen et al. validated the commercially available CoLumbo model originating from Bulgaria with imaging from their institution in Germany. Yet, the generalizability of these models could be further substantiated by additional validation studies.

The heterogeneity among the performance metrics reported precluded a quantitative analysis of the results for both segmentation and classification algorithms. There was no subset of metrics that was consistently reported in every study, creating a challenge for appropriate comparisons between the studies. To overcome this problem in future studies, reporting on ML algorithms needs to be standardized, and recommendations have been made to guide authors in this process [48, 49]. Additionally, more specific guidelines are being developed which may include a combination of performance metrics that should be reported for every ML model [19]. For segmentation algorithms, we propose that accuracy and the spatial error in mm should be reported at a minimum. For classification models, authors should report accuracy, the confusion matrix and area under the receiver operating characteristic curve [50]. Furthermore, a publicly available test dataset should be used to improve comparability between studies and assess generalizability. Ideally, such a dataset would contain multi-vendor MRI images from different medical centers across the world with high-quality labelling. To our knowledge, only one publicly available dataset with corresponding ground truth data is currently available containing axial and sagittal MRI images from 515 patients obtained across several international institu-

tion [51]. It was used by four studies from this review [20, 29, 42, 43], however, this dataset only contains segmentation ground truth data for one axial scan of the last three IVDs [52]. Radiologists' readings reporting on the presence of central or foraminal stenosis are available [53], but, ideally, classification labels should be derived from a consensus amongst spine-dedicated radiologists and surgeons. Furthermore, the scanning parameters are homogenous which does not reflect the heterogeneity of scans in clinical practice, and no demographic information is provided. Other datasets are available but with limited size and without appropriate ground truth data [54]. Furthermore, authors should clearly state the sample sizes and inter- and intra-observer variability when comparing with human performance. Notably, most studies in this review did not provide details on data or code sharing which defies transparency and developments in AI research [49].

This systematic review is limited by the databases that were searched since relevant articles may have been published in non-medical journals. Furthermore, the lack of consistency in reported outcome metrics precluded quantitative analyses.

Future studies should focus on DL models for classification tasks since they outperform conventional ML methods and obviate the need for segmentation. Furthermore, future studies should validate previously published models in addition to developing new ones. Of the 27 articles in this review, only seven discussed external validation [26, 27, 31, 34, 36, 38, 44]. It is essential that published algorithms are tested on external datasets for assessment of generalizability using a standardized set of performance metrics and a publicly available test dataset with high-quality ground truth data. Additionally, authors should consider the implementability of their classification algorithms, as the ultimate goal is to create a useful model for daily practice. Several models required manual pre-processing of the input data, and, therefore, will not be viable for the high volume of imaging in clinical practice.

This systematic review provides an overview of current conventional ML and DL methods for LSS segmentation and classification. We have elucidated the wide range of developed models and the tendency of DL methods to perform better than conventional ML models obviating the need for a segmentation step before classification. It is essential that guidelines are developed for reporting of performance metrics and that researchers focus on validation of (current) models on external datasets. Primarily for classification, DL models have great potential to improve diagnostics and aid clinicians in LSS identification.

Online Resources can be reviewed online:
https://doi.org/10.1007/s00586-025-08672-9

REFERENCES

1. Katz JN, Harris MB. Clinical practice. Lumbar spinal stenosis. N Engl J Med. 2008;358(8):818-25.
2. Katz JN, Zimmerman ZE, Mass H, Makhni MC. Diagnosis and Management of Lumbar Spinal Stenosis: A Review. JAMA. 2022;327(17):1688-99.
3. Deyo RA, Mirza SK, Martin BI, Kreuter W, Goodman DC, Jarvik JG. Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. Jama. 2010;303(13):1259-65.
4. Melancia JL, Francisco AF, Antunes JL. Chapter 35 - Spinal stenosis. In: Biller J, Ferro JM, editors. Handbook of Clinical Neurology. 119: Elsevier; 2014. p. 541-9.
5. Zileli M, Crostelli M, Grimaldi M, Mazza O, Anania C, Fornari M, et al. Natural Course and Diagnosis of Lumbar Spinal Stenosis: WFNS Spine Committee Recommendations. World Neurosurg X. 2020;7:100073.
6. Seo J, Lee JW. Magnetic Resonance Imaging Grading Systems for Central Canal and Neural Foraminal Stenoses of the Lumbar and Cervical Spines With a Focus on the Lee Grading System. Korean J Radiol. 2023;24(3):224-34.
7. Bozzo A, Marcoux J, Radhakrishna M, Pelletier J, Goulet B. The role of magnetic resonance imaging in the management of acute spinal cord injury. J Neurotrauma. 2011;28(8):1401-11.
8. Winklhofer S, Held U, Burgstaller JM, Finkenstaedt T, Bolog N, Ulrich N, et al. Degenerative lumbar spinal canal stenosis: intra- and inter-reader agreement for magnetic resonance imaging parameters. Eur Spine J. 2017;26(2):353-61.
9. Speciale AC, Pietrobon R, Urban CW, Richardson WJ, Helms CA, Major N, et al. Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area. Spine (Phila Pa 1976). 2002;27(10):1082-6.
10. Guen YL, Joon WL, Hee SC, Kyoung-Jin O, Heung SK. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. Skeletal Radiology. 2011;40(8):1033-9.
11. Schizas C, Theumann N, Burn A, Tansey R, Wardlaw D, Smith FW, 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):1919-24.
12. Miskin N, Isaac Z, Lu Y, Makhni MC, Sarno DL, Smith TR, et al. Simplified Universal Grading of Lumbar Spine MRI Degenerative Findings: Inter-Reader Agreement of Non-Radiologist Spine Experts. Pain Medicine. 2021;22(7):1485-95.
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500-10.
14. Aljuaid A, Anwar M. Survey of Supervised Learning for Medical Image Processing. SN Computer Science. 2022;3(4):292.
15. Seo H, Badiei Khuzani M, Vasudevan V, Huang C, Ren H, Xiao R, et al. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. Med Phys. 2020;47(5):e148-e67.
16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60-88.

17. Wang H, Zhou Z, Li Y, Chen Z, Lu P, Wang W, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from (18)F-FDG PET/CT images. EJNMMI Res. 2017;7(1):11.
18. Wells G, Shea B, O'Connell J. The Newcastle-Ottawa Scale (NOS) for Assessing The Quality of Nonrandomised Studies in Meta-analyses. Ottawa Health Research Institute Web site. 2014;7.
19. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
20. Al-Kafri AS, Sudirman S, Hussain A, Al-Jumeily D, Natalia F, Meidia H, et al. Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks. IEEE Access. 2019;7:43487-501.
21. Al-Kafri AS, Sudirman S, Hussain AJ, Al-Jumeily D, Fergus P, Natalia F, et al., editors. Segmentation of Lumbar Spine MRI Images for Stenosis Detection Using Patch-Based Pixel Classification Neural Network. 2018 IEEE Congress on Evolutionary Computation (CEC); 2018 8-13 July 2018.
22. Altun S, Alkan A, Altun İ. LSS-VGG16: Diagnosis of Lumbar Spinal Stenosis With Deep Learning. Clin Spine Surg. 2023;36(5):E180-e90.
23. Altun S, Alkan A. LSS-net: 3-dimensional segmentation of the spinal canal for the diagnosis of lumbar spinal stenosis. International Journal of Imaging Systems and Technology. 2022;33(1):378-88.
24. Gaonkar B, Villaroman D, Beckett J, Ahn C, Attiah M, Babayan D, et al. Quantitative Analysis of Spinal Canal Areas in the Lumbar Spine: An Imaging Informatics and Machine Learning Study. AJNR Am J Neuroradiol. 2019;40(9):1586-91.
25. Ghosh S, Chaudhary V. Supervised methods for detection and segmentation of tissues in clinical lumbar MRI. Computerized Medical Imaging and Graphics. 2014;38(7):639-49.
26. Grob A, Loibl M, Jamaludin A, Winklhofer S, Fairbank JCT, Fekete T, et al. External validation of the deep learning system "SpineNet" for grading radiological features of degeneration on MRIs of the lumbar spine. Eur Spine J. 2022;31(8):2137-48.
27. Hallinan J, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI. Radiology. 2021;300(1):130-8.
28. Han Z, Wei B, Mercado A, Leung S, Li S. Spine-GAN: Semantic segmentation of multiple spinal structures. Med Image Anal. 2018;50:23-35.
29. Hou C, Zhang W, Wang H, Liu F, Liu D, Chang J. A semantic segmentation model for lumbar MRI images using divergence loss. Applied Intelligence. 2023;53(10):12063-76.
30. Huber FA, Stutz S, Vittoria de Martini I, Mannil M, Becker AS, Winklhofer S, et al. Qualitative versus quantitative lumbar spinal stenosis grading by machine learning supported texture analysis-Experience from the LSOS study cohort. Eur J Radiol. 2019;114:45-50.
31. Ishimoto Y, Jamaludin A, Cooper C, Walker-Bone K, Yamada H, Hashizume H, 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.
32. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, Battié MC, 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):1374-83.
33. Jamaludin A, Kadir T, Zisserman A. SpineNet: Automated classification and evidence visualization in spinal MRIs. Med Image Anal. 2017;41:63-73.
34. Kim T, Kim YG, Park S, Lee JK, Lee CH, Hyun SJ, et al. Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs. J Neurosurg Spine. 2022:1-8.
35. Koompairojn S, Hua K, Hua K, Srisomboon J. Computer-Aided Diagnosis of Lumbar Stenosis Conditions. Proc SPIE. 2010.
36. Lehnen NC, Haase R, Faber J, Rüber T, Vatter H, Radbruch A, et al. Detection of Degenerative Changes on MR Images of the Lumbar Spine with a Convolutional Neural Network: A Feasibility Study. Diagnostics (Basel). 2021;11(5).
37. Lewandrowsk IK, Muraleedharan N, Eddy SA, Sobti V, Reece BD, Ramírez León JF, et al. Feasibility of Deep Learning Algorithms for Reporting in Routine Spine Magnetic Resonance Imaging. Int J Spine Surg. 2020;14(s3):S86-s97.
38. Lewandrowski KU, Muraleedharan N, Eddy SA, Sobti V, Reece BD, Ramírez León JF, et al. Reliability Analysis of Deep Learning Algorithms for Reporting of Routine Lumbar MRI Scans. Int J Spine Surg. 2020;14(s3):S98-s107.
39. Li H, Luo H, Huan W, Shi Z, Yan C, Wang L, et al. Automatic lumbar spinal MRI image segmentation with a multi-scale attention network. Neural Comput Appl. 2021;33(18):11589-602.
40. Lin L, Tao X, Yang W, Pang S, Su Z, Lu H, et al. Quantifying Axial Spine Images Using Object-Specific Bi-Path Network. IEEE Journal of Biomedical and Health Informatics. 2021;25(8):2978-87.
41. Lu J-T, Pedemonte S, Bizzo B, Doyle S, Andriole K, Michalski M, et al. DeepSPINE: Automated Lumbar Vertebral Segmentation, Disc-level Designation, and Spinal Stenosis Grading Using Deep Learning2018.
42. Natalia F, Meidia H, Afriliana N, Young JC, Yunus RE, Al-Jumaily M, et al. Automated measurement of anteroposterior diameter and foraminal widths in MRI images for lumbar spinal stenosis diagnosis. PLoS One. 2020;15(11):e0241309.
43. Pang C, Su Z, Lin L, Lin G, He J, Lu H, et al. Automated measurement of spine indices on axial MR images for lumbar spinal stenosis diagnosis using segmentation-guided regression network. Med Phys. 2023;50(1):104-16.
44. Su ZH, Liu J, Yang MS, Chen ZY, You K, Shen J, et al. Automatic Grading of Disc Herniation, Central Canal Stenosis and Nerve Roots Compression in Lumbar Magnetic Resonance Image Diagnosis. Front Endocrinol (Lausanne). 2022;13:890371.
45. Tang H, Pei X, Huang S, Li X, Liu C. Automatic Lumbar Spinal CT Image Segmentation With a Dual Densely Connected U-Net. IEEE Access. 2020;8:89228-38.
46. Won D, Lee HJ, Lee SJ, Park SH. Spinal Stenosis Grading in Magnetic Resonance Imaging Using Deep Convolutional Neural Networks. Spine (Phila Pa 1976). 2020;45(12):804-12.
47. Kafri ASA, Sudirman S, Hussain AJ, Al-Jumeily D, Fergus P, Natalia F, et al., editors. Segmentation of Lumbar Spine MRI Images for Stenosis Detection Using Patch-

Based Pixel Classification Neural Network. 2018 IEEE Congress on Evolutionary Computation (CEC); 2018 8-13 July 2018.
48. Huang C, Li SX, Caraballo C, Masoudi FA, Rumsfeld JS, Spertus JA, et al. Performance Metrics for the Comparative Analysis of Clinical Risk Prediction Models Employing Machine Learning. Circ Cardiovasc Qual Outcomes. 2021;14(10):e007526.
49. Stevens L, Mortazavi B, Deo R, Curtis L, Kao D. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circulation: Cardiovascular Quality and Outcomes. 2020;13.
50. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Scientific Reports. 2022;12(1):5979.
51. Sudirman S, Ala Al-Kafri, Natalia F, Meidia H, Afriliana N, Al-Rashdan W, et al. Lumbar Spine MRI Dataset. Version 2. Mendeley Data. 2019. doi:10.17632/k57fr854j2.2.
52. Sudirman S, Ala Al-Kafri, Natalia F, Meidia H, Afriliana N, Al-Rashdan W, et al. Label Image Ground Truth Data for Lumbar Spine MRI Dataset. Version 2. Mendeley Data. 2019. doi:10.17632/zbf6b4pttk.2.
53. Sudirman S, Ala Al-Kafri, Natalia F, Meidia H, Afriliana N, Al-Rashdan W, et al. Radiologists Notes for Lumbar Spine MRI Dataset. Version 2. Mendely Data. 2019. doi:10.17632/s6bgczr8s2.2.
54. SpineWeb: Collaborative Platform for Research on Spine Imaging and Image Analysis. Datasets. 2018. Available from: http://spineweb.digitalimaginggroup.ca/Index.php?n=Main.Datasets. [Accessed on: 29-January-2024].