Review

Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis

Tianyi Wang*, MD; Ruiyuan Chen*, MM; Ning Fan*, MD; Lei Zang, MD; Shuo Yuan, MD; Peng Du, MM; Qichao Wu, MD; Aobo Wang, MD; Jian Li, MM; Xiaochuan Kong, MB; Wenyi Zhu, MD

Beijing Chaoyang Hospital, Capital Medical University, Beijing, China

*these authors contributed equally

Corresponding Author:

Lei Zang, MD
Beijing Chaoyang Hospital
Capital Medical University
5 JingYuan Road, Shijingshan District
Beijing, 100043
China
Phone: 86 51718268
Email: zanglei@ccmu.edu.cn

Abstract

Background: Lumbar spinal stenosis (LSS) is a major cause of pain and disability in older individuals worldwide. Although increasing studies of traditional machine learning (TML) and deep learning (DL) were conducted in the field of diagnosing LSS and gained prominent results, the performance of these models has not been analyzed systematically.

Objective: This systematic review and meta-analysis aimed to pool the results and evaluate the heterogeneity of the current studies in using TML or DL models to diagnose LSS, thereby providing more comprehensive information for further clinical application.

Methods: This review was performed under the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines using articles extracted from PubMed, Embase databases, and Cochrane Library databases. Studies that evaluated DL or TML algorithms assessment value on diagnosing LSS were included, while those with duplicated or unavailable data were excluded. Quality Assessment of Diagnostic Accuracy Studies 2 was used to estimate the risk of bias in each study. The MIDAS module and the METAPROP module of Stata (StataCorp) were used for data synthesis and statistical analyses.

Results: A total of 12 studies with 15,044 patients reported the assessment value of TML or DL models for diagnosing LSS. The risk of bias assessment yielded 4 studies with high risk of bias, 3 with unclear risk of bias, and 5 with completely low risk of bias. The pooled sensitivity and specificity were 0.84 (95% CI: 0.82-0.86; I^2=99.06%) and 0.87 (95% CI 0.84-0.90; I^2=98.7%), respectively. The diagnostic odds ratio was 36 (95% CI 26-49), the positive likelihood ratio (LR+) was 6.6 (95% CI 5.1-8.4), and the negative likelihood ratio (LR–) was 0.18 (95% CI 0.16-0.21). The summary receiver operating characteristic curves, the area under the curve of TML or DL models for diagnosing LSS of 0.92 (95% CI 0.89-0.94), indicating a high diagnostic value.

Conclusions: This systematic review and meta-analysis emphasize that despite the generally satisfactory diagnostic performance of artificial intelligence systems in the experimental stage for the diagnosis of LSS, none of them is reliable and practical enough to apply in real clinical practice. Further efforts, including optimization of model balance, widely accepted objective reference standards, multimodal strategy, large dataset for training and testing, external validation, and sufficient and scientific report, should be made to bridge the distance between current TML or DL models and real-life clinical applications in future studies.

Trial Registration: PROSPERO CRD42024566535; https://tinyurl.com/msx59x8k

(J Med Internet Res 2024;26:e54676) doi: 10.2196/54676

KEYWORDS

lumbar spinal stenosis; LSS; machine learning; ML; deep learning; artificial intelligence; AI; diagnosis; spine stenosis; lumbar; predictive model; early detection; diagnostic; older adult;

Introduction

Lumbar spinal stenosis (LSS) is a major cause of pain and disability in older individuals [1]. LSS has become a worldwide public health issue as it is estimated that more than 102 million people are diagnosed with LSS annually, with high incidence in Europe and the United States of America [2,3]. According to the clinical guideline developed by the North American Spine Society, LSS is characterized as a condition of diminished space available for the neural and vascular elements in the lumbar spine, secondary to degenerative changes in the spinal canal [4]. An accurate LSS diagnosis is essential for treatment options and effectiveness. Currently, clinicians diagnose LSS based on a comprehensive evaluation combined with the patient's history, physical examination, and spinal imaging tests such as x-ray, computed tomography (CT), and magnetic resonance imaging (MRI) [1,2]. As a superior radiographic screening tool for soft tissues, MRI plays a crucial role in detecting the presence, classification, and grading of LSS [5-7]. However, detailing numerous information in spinal MRI is time-consuming and repetitive, which causes laborious clinical workloads [7]. Furthermore, existing LSS grading systems are mainly qualitative or semiquantitative, which highly depend on expertise and suffer from high interobserver variations because of the complexity of the spinal canal and foramen [5,6,8-11]. Therefore, more intelligent radiographic diagnostic and grading methods of LSS are warranted.

Machine learning (ML), a subdiscipline of artificial intelligence (AI), has shown great advantages in analyzing medical imaging and predicting outcome decisions [12-14]. ML begins with algorithms trained with a set of data, such as image features, to establish the prediction or diagnosis by extracting and classifying relevant information. More recently, a crucial branch of ML, named deep learning (DL), was standing out rapidly. DL algorithms were designed with multiple processing layers, which can learn more complex image features than traditional ML methods [15]. Although DL is still challenged by the demand for large-scale datasets and the difficulty of interpretation, it owns the incomparable advantage of automatic feature extraction, minimizing the bias by manual intervention [12,14]. In 2016, He et al [16] attempted to use traditional ML (TML) methods based on their newly proposed synchronized superpixel representation model to recognize the presence of radiographic lumbar foraminal stenosis (LFS). Subsequently, increasing studies of TML and DL were conducted in the field of diagnosing and grading LSS and gained prominent results [16-34]. However, most of these studies focus either on algorithm development or clinical validation, causing great variations in experimental settings and incompleteness of evaluation parameters of accuracy and reliability. Hence, a systematic review and meta-analysis were believed to be necessary to evaluate the heterogeneity and provide comprehensive results from these studies. However, to our knowledge, no systematic review and meta-analysis was previously conducted to address this issue.

Therefore, this systematic review and meta-analysis aimed to evaluate the heterogeneity and pool the results of the current studies in using ML or DL models to diagnose LSS, thereby providing more comprehensive information for further clinical application.

Methods

Study Design and Registration

This systematic literature review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and flowchart [35,36] and the PRISMA diagnostic test accuracy checklist (Multimedia Appendix 1) [37]. The protocol for this systematic review was registered in PROSPERO (ID: CRD42024566535). Ethical approval was not required because this systematic literature review focused on retrospective studies.

Search Strategy

This review collected the records from 3 major databases up to October 2023. A second search was performed in February 2024 to complement newly published studies. Those databases include PubMed, Embase, and the Cochrane Library (CENTRAL), which are recommended academic search systems for systematic reviews and meta-analyses [38]. We used the MeSH (Medical Subject Headings) and Emtree headings in several combinations and supplemented them with free text to increase sensitivity. In addition, we searched references contained in the included studies to supplement the relevant literature. An experienced librarian designed and implemented the search strategy. The following MeSH terms were used for PubMed: "Spinal Stenosis," "Intervertebral Disc Degeneration," "Lumbar Vertebrae," "Machine Learning," "Deep Learning," and "Neural Networks, Computer*." The details of the search strategy are stated in Multimedia Appendix 2.

Inclusion and Exclusion Criteria

We included studies that evaluated the assessment value of DL or TML algorithms for diagnosing LSS and that were available in English. The included studies in the meta-analysis should provide or could be reconstructed as a 2×2 confusion matrix from sensitivity, specificity, and precision. Applied statistical, non–artificial intelligence, and general AI methods are not considered DL or TML. Articles with duplicated or unavailable data were excluded. Furthermore, abstracts from protocols, case reports, editorials, and review articles were excluded.

Review Process

A total of 2 reviewers (TW and NF) independently performed an initial screening of the titles and abstracts of the remaining articles to determine potential eligibility after removing duplicates. We reviewed the full texts of the remaining articles and excluded those that did not meet the inclusion criteria. We searched and screened a list of references for all relevant studies and a systematic review of potentially relevant studies. Disagreements were resolved by discussion and by third-party adjudication when necessary. For studies enrolled in systematic review while lack of available data for meta-analysis, an email was sent to the corresponding authors for acquisition of the necessary data.

Data Extraction

A total of 2 reviewers independently extracted, summarized, and tabulated the following data using a standard form: baseline characteristics of studies, including the publication year, study type, model type, algorithms used, LSS classifications, number of participants, validation strategy, imaging modality, and diagnosis criteria of LSS. Any discrepancies in the extracted data were resolved by discussion. For the studies that provided multiple contingency tables based on different classifier algorithms, datasets, LSS types, or label strategies, we assumed these to be independent of each other. For the studies that provided multiple contingency tables based on different preprocessing strategies, we selected the best-performing result. If there was no preprocessing strategy that performed significantly better than the others, we also enrolled each strategy as an individual study and collected the corresponding results. For repeat test results based on the same classifier algorithms, datasets, and so on, we calculated the average values of metrics as the final results.

Quality Assessment

A total of 2 reviewers (TW and NF) used the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2), which is a tool for assessing the quality of primary diagnostic accuracy studies, to independently assess the risk of bias for each eligible study [39]. The QUADAS-2 criteria assessed the risk of bias in 4 domains: patient selection, index test, reference standard, and flow and timing. Any disagreements were resolved by discussion with a third author.

Statistical Analysis

We used the MIDAS module and the METAPROP module [40] of Stata (version 17.0; StataCorp) for statistical analysis. Postestimation procedures for model diagnostic were used to assess heterogeneity using the I^2 statistic. The following metrics were used: 0%-40% (low heterogeneity), 30%-60% (moderate heterogeneity), 50%-90% (substantial heterogeneity), and 75%-100% (considerable heterogeneity). Bivariate mixed-effects logistic regression modeling was conducted, and forest plots were used to compare the sensitivity and specificity of DL or TML models for diagnosing LSS. We used summary receiver operating characteristic (SROC) curves to assess overall diagnostic accuracy. We used the Fagan nomogram to explore the relationship between pretest probability, likelihood ratio (LR), and posttest probability. LR dot plots were divided into 4 quadrants according to the strength of the evidence threshold, which was used to determine DL or TML model exclusion and confirmation. Finally, subgroup analyses were performed to examine whether the estimated sensitivity, specificity, and associated I^2 differed by several moderators when each subgroup included ≥4 datasets.

Results

Study Selection and Characteristics

The initial search identified 934 titles and abstracts, of which 269 were duplicates. After screening, 567 articles were excluded following this study’s inclusion and exclusion criteria. In addition, 98 studies were reviewed for full text, of which 19 and 12 studies were included in the systematic review and meta-analysis, respectively (Figure 1). Table 1 summarizes the characteristics of the studies in the systematic review and meta-analyses, including study type, model type, algorithms used, LSS classifications, number of participants, validation strategy, imaging modality, and diagnosis criteria of LSS. The 19 studies included in the systematic review were published from 2016 to 2024. The 12 studies included in the meta-analysis were all retrospective and included 21 external tests [17,25-27,30,32,34] and 35 internal tests [17,20,22,24,26-28,33,34]. Therefore, the meta-analysis included 56 datasets and completely different data sources. Among the 56 datasets, 32 identified LSS on MRI [17,20,22,24,25,27,28,30,32], 20 on x-ray [26,34], and 4 on CT [33]. Furthermore, 29 datasets have developed and internally tested DL models [17,20,24,26-28,33,34], 6 datasets internally tested TML models [22,28], and 21 datasets externally tested the DL models [17,25-27,30,32,34].

Figure 1. Flowchart depicting PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) search strategy.

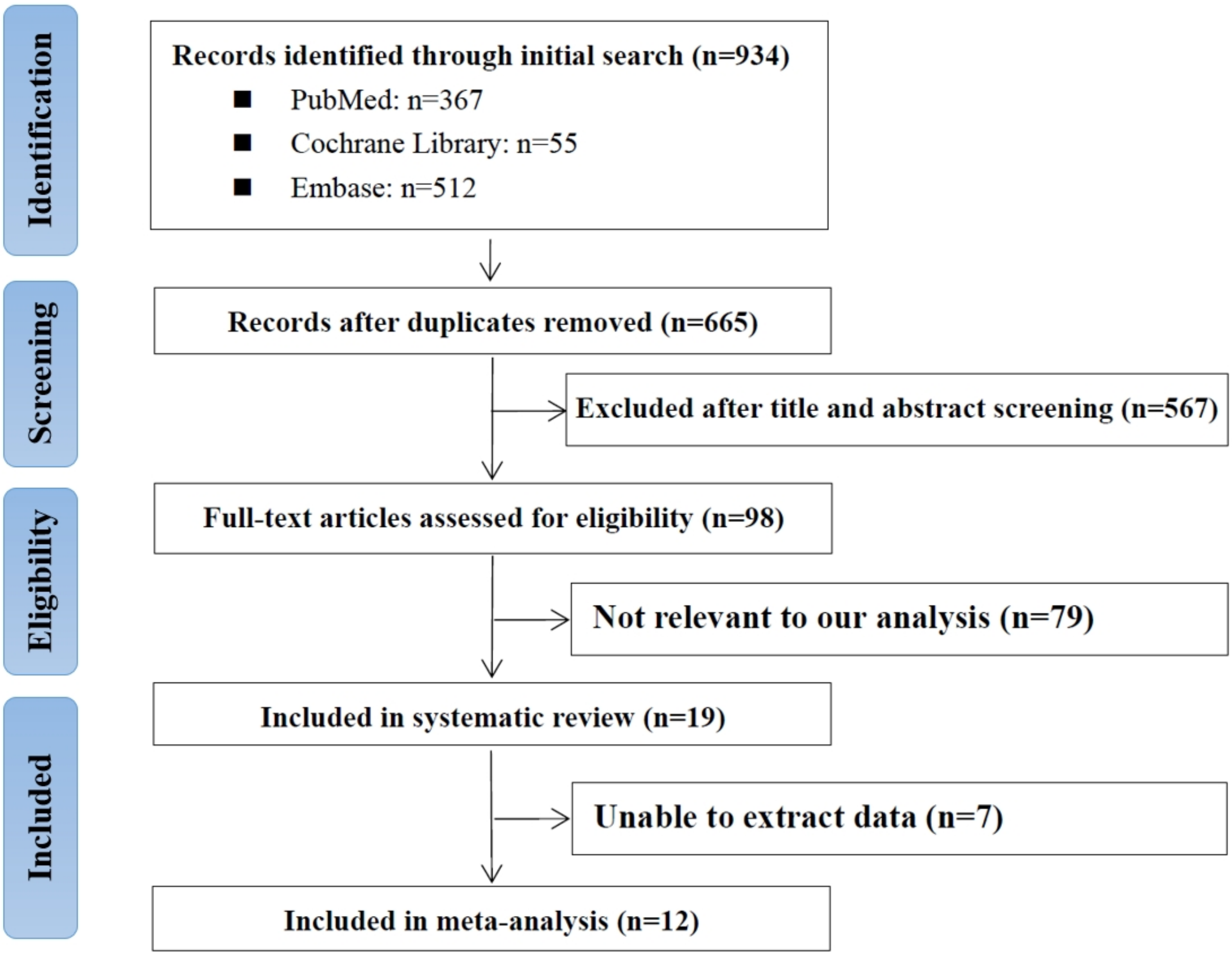

Identification
Records identified through initial search (n=934)
■ PubMed: n=367
■ Cochrane Library: n=55
■ Embase: n=512
Screening
Records after duplicates removed (n=665)
Excluded after title and abstract screening (n=567)
Eligibility
Full-text articles assessed for eligibility (n=98)
Not relevant to our analysis (n=79)
Included
Included in systematic review (n=19)
Unable to extract data (n=7)
Included in meta-analysis (n=12)

Table 1. Characteristics of the included studies in the systemic review and meta-analysis.

Study	Study type	Model type	Algorithms used[a]	LSS[b] type	Number of participants, n	Validation strategy	Imaging modality	Diagnosis criteria
He et al [16]	Development study and internal test	TML[c]	KNN[d] and SVM[e] and LDA[f]	LFS[g]	110	Cross-validation	MRI[h]	Lee et al [5]
Jamaludin et al [18]	Development study and internal test	DL[i]	CNN[j] (SpineNet[k])	LCS[l]	2009	Hold-out validation	MRI	No reference
Zhang et al [19]	Development study and internal test	TML	SVM and Decision Tree	LCS and LFS	600	Hold-out validation	MRI	No reference
Lu et al [21]	Development study and internal test	DL	ResNeXt-50	LCS and LFS	4075	Hold-out validation	MRI	No reference
Han et al [20][m]	Development study and internal test	DL	CNN and FCN and SegNet and DeepLabv3+ and U-Net	LFS	253	Cross-validation	MRI	Lee et al [5]
Huber et al [22][m]	Development study and internal test	TML	Decision Tree	LCS	82	Cross-validation	MRI	Lee et al [9] and Schizas et al [6]
Ishimoto et al [23]	Replication study and internal test	DL	CNN (SpineNet[k])	LCS	971	Hold-out validation	MRI	Lurie et al [8]
Won et al [24][m]	Development study and internal test	DL	VGG	LCS	542	Cross-validation	MRI	Schizas et al [5]
Hallinan et al [17][m]	Development study and internal test and external test	DL	CNN	LCS and LRS[n] and LFS	446/100	Hold-out validation	MRI	Lurie et al [8] and Bartynski and Lin [10]
Lehnen et al [25][m]	External test	DL	CNN (CoLumbo[k])	LCS	146	—[o]	MRI	Lee et al [9]
Grob et al [30][m]	External test	DL	CNN (SpineNet[k])	LCS	882	—	MRI	Lurie et al [8]
Kim et al [26][m]	Development study and internal test and external test	DL	VGG19 and VGG16 and ResNet50 and Efficient1	LCS	4644/199	Cross-validation	X-Ray	Lee et al [9]
Su et al [27][m]	Development study and internal test and external test	DL	ResNet-50	LCS	1015/100	Hold-out validation	MRI	Lee et al [9] and Park et al [11]
Altun et al [28][m]	Development study and internal test	TML and DL	RF and SVM and VGG16 and ResNet and MobileNet and InceptionNet	LSS	1030	Cross-validation	MRI	No reference
Bharadwaj et al [29]	Development study and internal test	TML and DL	Decision Tree and BiTCNN	LCS and LFS	200	Hold-out validation	MRI	Schizas et al [6] and Lee et al [5]
Tumko et al [32][m]	Development study and external test	DL	RegNetY32GF	LCS and LRS and LFS	1635/150	—	MRI	Schizas et al [6]
Shahzadi et al [31]	Development study and internal test	DL	CNN	LRS and LFS	515	Cross-validation	MRI	No reference
Li et al [33][m]	Development study and internal test	DL	VGG11 and ResNet-18	LCS and LRS	236	Hold-out validation	CT	Lurie et al [8] and Bartynski and Lin [10]
Park et al [34][a]	Development study and internal test and extra-internal test and external test	DL	ResNet50 and VGG19 and VGG16 and EfficientNet-B1	LCS	3831/199/100	Cross-validation (validation) and Hold-out validation (Internal test)	X-Ray	No reference

[a] Algorithms for only classifiers.

[b] LSS: lumbar spinal stenosis.

[c] TML: traditional machine learning.

[d] KNN: k-nearest neighbors.

[e] SVM: support vector machine.

[f] LDA: linear discriminant analysis.

[g] LFS: lumbar foraminal stenosis.

[h] MRI: magnetic resonance imaging.

[i] DL: deep learning.

[j] CNN: convolutional neural network.

[k] Name of software.

[l] LCS: lumbar central stenosis.

[m] Studies included in meta-analysis (confusion matrix available or can be reconstructed).

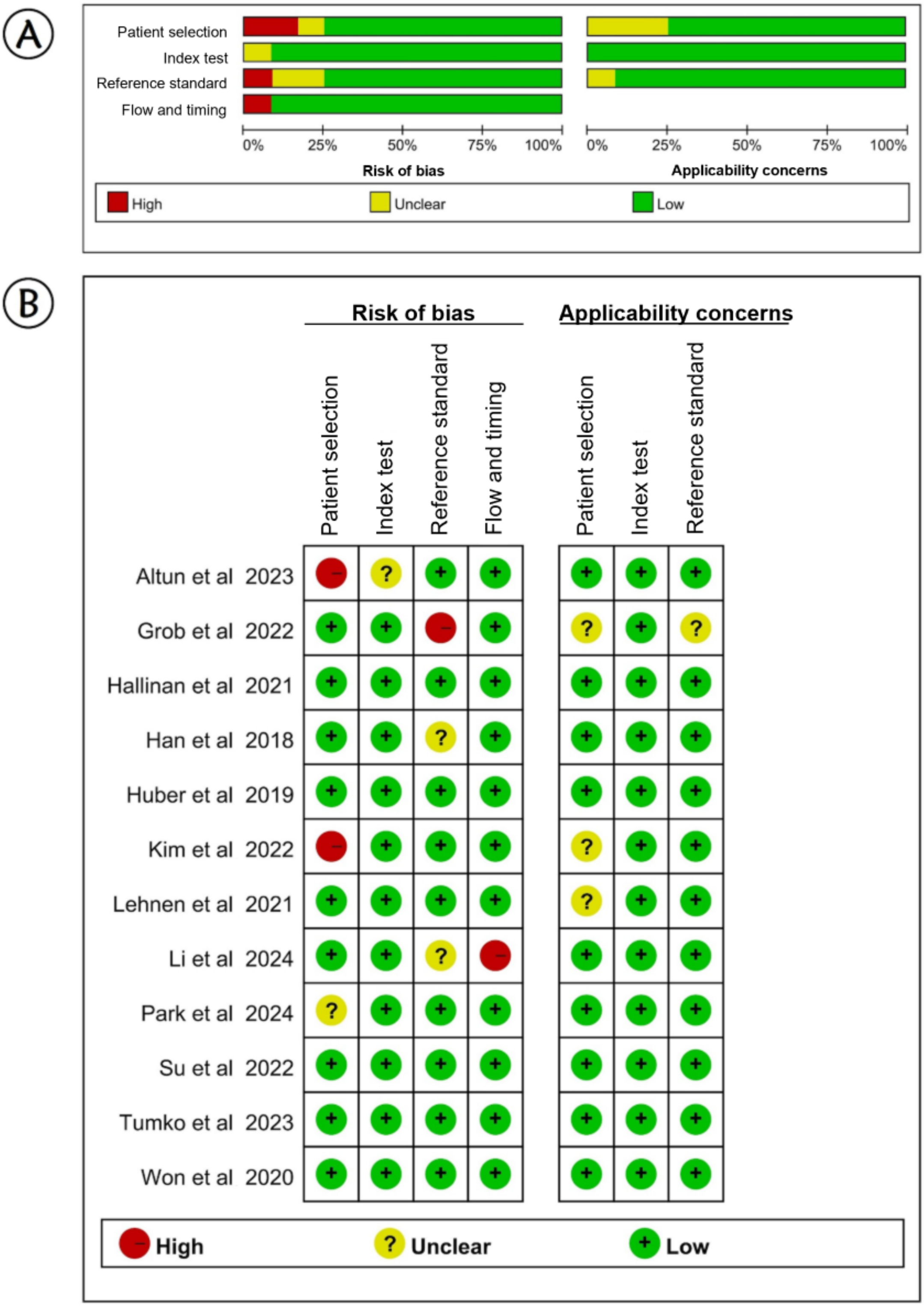
[n] LRS: lateral recess stenosis.

[o] Not applicable.

Methodological Quality

Regarding the QUADAS-2 risk of bias assessment (Figure 2 [17,20,22,24-28,30,32-34]), we revealed 4 studies with a high risk of bias [26,28,30,33], 3 with an unclear risk of bias [20,25,34], and 5 with a completely low risk of bias [17,22,24,27,32]. In particular, 2 of the included studies reported no details of patient selection [26,28], causing a high bias in patient selection. Furthermore, 1 study provided unclear information on how to perform the index test [28], thereby causing an unclear risk of bias. Furthermore, 1 study used the improper reference standard, which was not likely to correctly classify the target condition [30], causing a high risk of bias in the reference standard. Besides, 1 study showed a high risk of bias with regard to flow and timing issues [33].

Figure 2. Methodological assessment by Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). (A) The proportion of risk of bias for all domains and proportion of applicability concerns in three domains. (B) Summary of the risk of bias for each study. Green, yellow, and red circles indicate low, unclear, and high risk of bias, respectively [8,20,22,24-28,30,32-34].



	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Altun et al 2023	−	?	+	+	+	+	+
Grob et al 2022	+	+	−	+	?	+	?
Hallinan et al 2021	+	+	+	+	+	+	+
Han et al 2018	+	+	?	+	+	+	+
Huber et al 2019	+	+	+	+	+	+	+
Kim et al 2022	−	+	+	+	?	+	+
Lehnen et al 2021	+	+	+	+	?	+	+
Li et al 2024	+	+	?	−	+	+	+
Park et al 2024	?	+	+	+	+	+	+
Su et al 2022	+	+	+	+	+	+	+
Tumko et al 2023	+	+	+	+	+	+	+
Won et al 2020	+	+	+	+	+	+	+

− High ? Unclear + Low

Performance of TML and DL Models for LSS

A total of 12 studies with 15,044 patients reported the assessment value of TML or DL models for diagnosing LSS. The pooled sensitivity was 0.84 (95% CI 0.82-0.86; I^2=99.06%), and specificity was 0.87 (95% CI 0.84-0.90; I^2=98.7%; Figure 3). The diagnostic odds ratio was 36 (95% CI 26-49). The SROC curve (Figure 4) revealed that the area under the curve of TML or DL models for diagnosing LSS was 0.92 (95% CI 0.89-0.94), indicating a high diagnostic value.

Figure 3. Forest plots in sensitivity and specificity of traditional machine learning (TML) or deep learning (DL) models. The pooled sensitivity and specificity were 0.84 (95% CI 0.82-0.86) and 0.87 (95% CI 0.84-0.90), respectively.

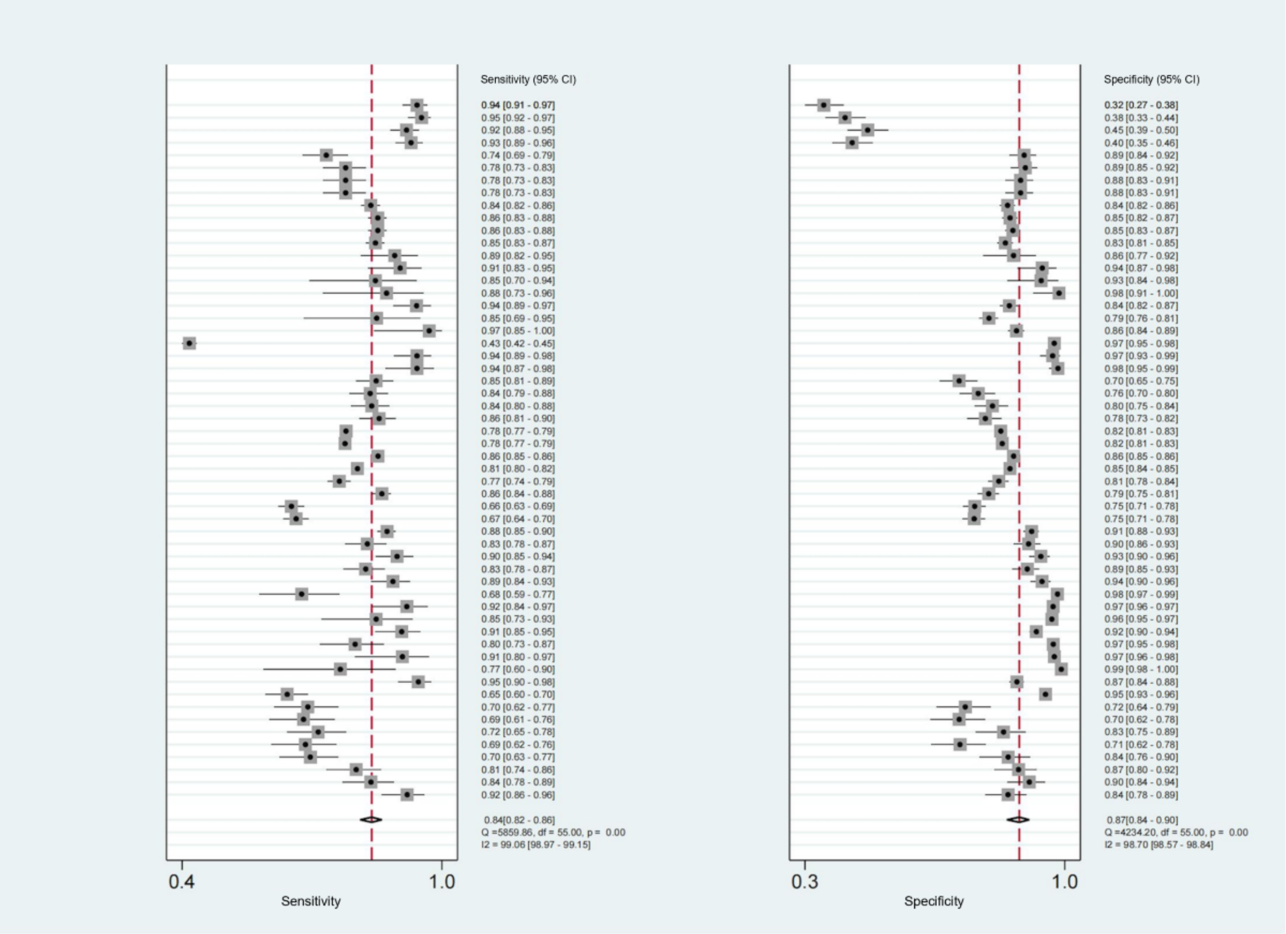
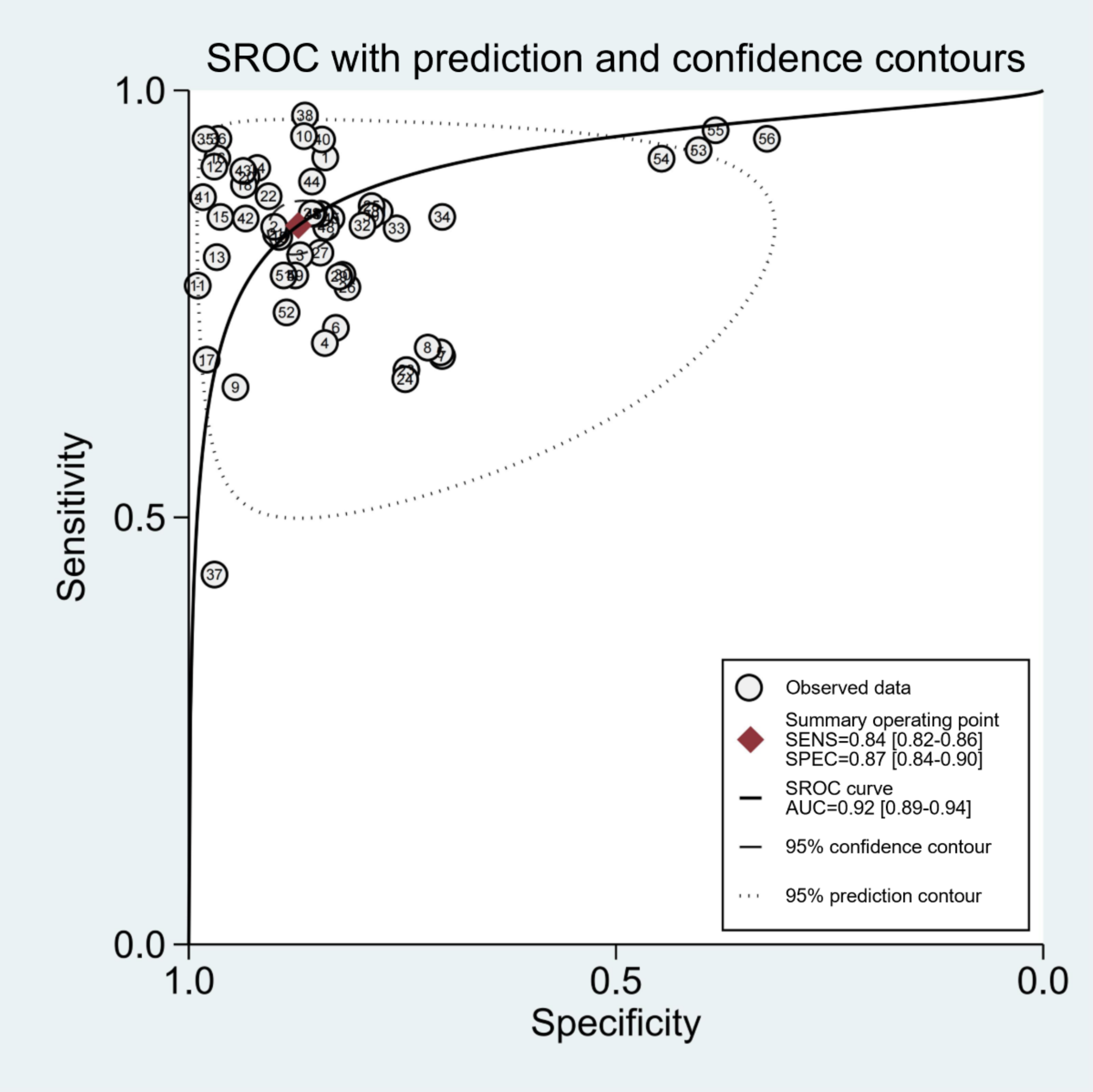


Figure 4. The SROC graph for the studies. The AUC of summary receiver operating characteristic (TML) or deep learning (DL) models for diagnosing LSS was 0.92 (95% CI 0.89-0.94). SROC: summary receiver operating characteristic; AUC: area under the curve; SENS: sensitivity; SPEC: specificity.



We set the pretest probability to 50% based on the pretest probability of disease. At this point, true positives accounted for 87% when patients were diagnosed with LSS by the TML or DL model, and false negatives accounted for 15% when the diagnosis was not LSS (Figure 5). Furthermore, the models showed a positive likelihood ratio (LR+) of 6.6 (95% CI 5.1-8.4) and a negative likelihood ratio (LR–) of 0.18 (95% CI 0.16-0.21), respectively (Figure 5). However, the summary likelihood ratio plot of TML or DL models was in the right lower quadrant (LR+<10 and LR–>0.1: no exclusion or confirmation), and the individual plots were scattered and distributed (Figure 6). The results indicated that although the TML or DL models achieved an acceptable performance generally, it was still insufficient enough for diagnosing or excluding LSS, and the current models suffered from certain performance variations.

Figure 5. Fagan nomogram of traditional machine learning (TML) or deep learning (DL) models for diagnosing lumbar spinal stenosis (LSS). The first column of this nomogram represents the pretest probability, the second column represents the likelihood ratio, and the third shows the posttest probability.

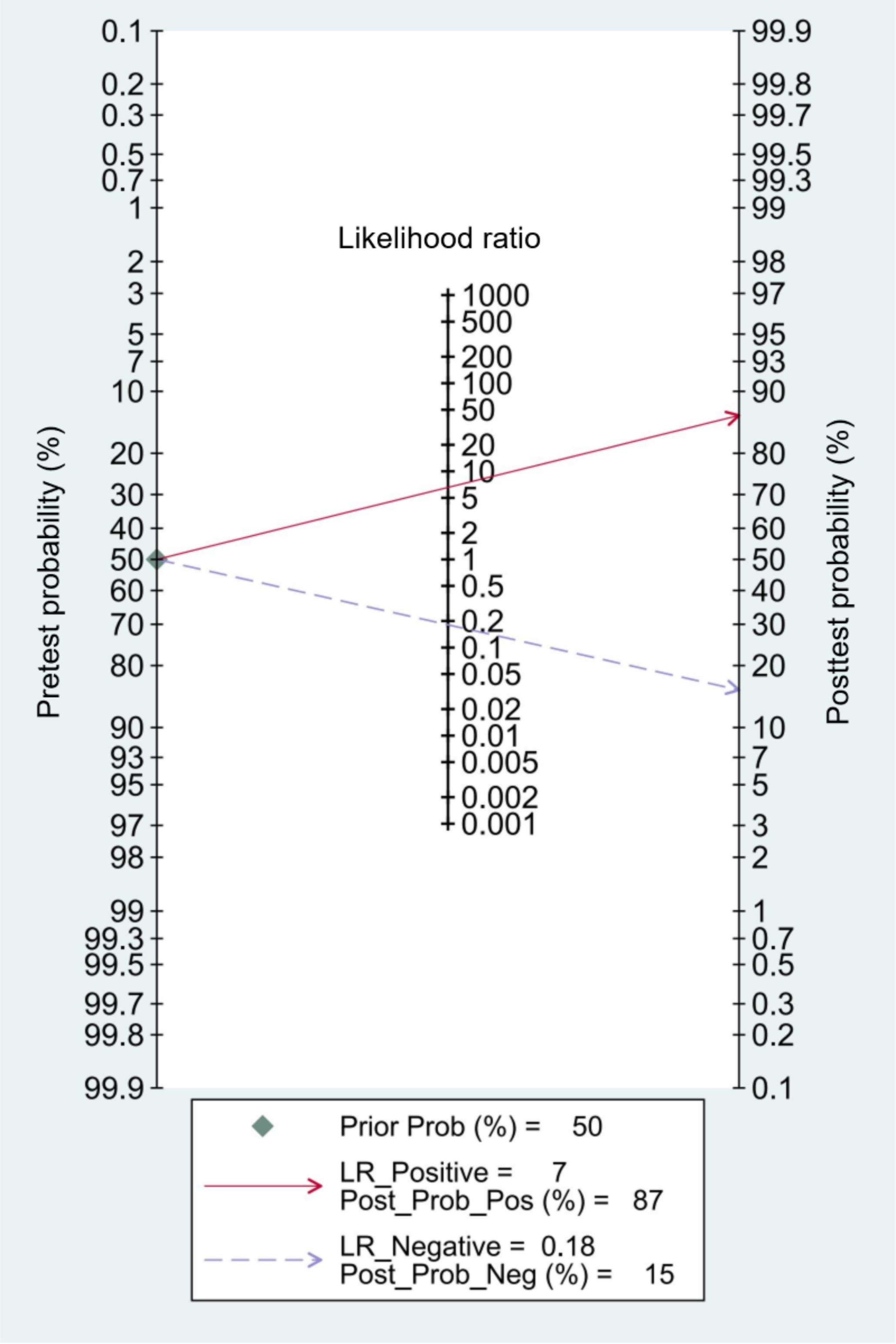
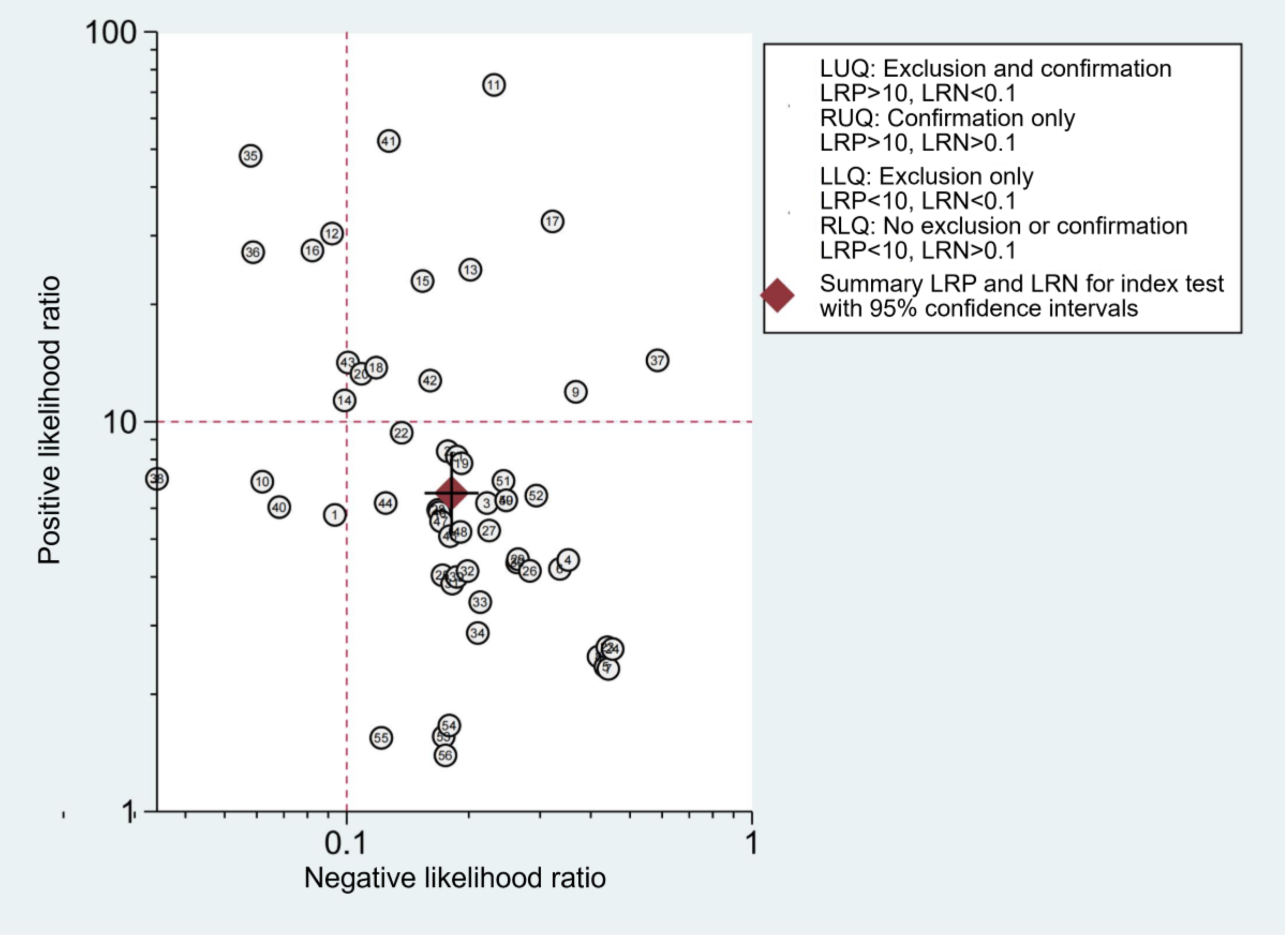


Figure 6. Likelihood ratio (LR) dot plot of traditional machine learning (TML) or deep learning (DL) prediction models. The summary point of TML or DL models was in the right lower quadrant (LR+<10 and LR–>0.1: no exclusion or confirmation). LRN: negative likelihood ratio; LRP: positive likelihood ratio; LUQ: left upper quadrant; RLQ: right lower quadrant; RUQ: right upper quadrant.



In total, 4 studies [17,23,29,32] simultaneously provided the performance of reliability both of observers and TML or DL models, including 3 studies [17,29,32] that performed a direct comparison between their reliabilities based on the same assessment datasets (Table 2).

Table 2. Characteristics of the studies available for the agreement between models and observers and reference standard.

Study	Number of participants, n	Agreement assessment strategy	Control group	Model type	Type of classification	LSS[a] type	Model results	Control group results
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL[b]	Binary	LCS[c]	0.96	0.98/0.98
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Binary	LRS[d]	0.92	0.92/0.95
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Binary	LFS[e]	0.89	0.94/0.95
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LCS	0.82	0.89/0.89
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LRS	0.72	0.71/0.79
Hallinan et al [17]	446	Gwet κ	2 radiologists	DL	Multigrading	LFS	0.75	0.80/0.87
Bharadwaj et al [29]	200	Cohen κ	2 radiologists	DL	Multigrading	LCS	0.54	0.80/0.86
Bharadwaj et al [29]	200	Cohen κ	2 radiologists	TML[f]	Multigrading	LCS	0.80	0.80/0.86
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LCS	0.431	Average 0.372
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LRS	0.315	Average 0.323
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Binary	LFS	0.672	Average 0.596
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LCS	0.310	Average 0.376
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LRS	0.199	Average 0.359
Tumko et al [32]	150	Cohen κ	7 radiologists	DL	Multigrading	LFS	0.637	Average 0.620

[a] LSS: lumbar spinal stenosis.
[b] DL: deep learning.
[c] LCS: lumbar central stenosis.
[d] LRS: lateral recess stenosis.
[e] LFS: lumbar foraminal stenosis.
[f] TML: traditional machine learning.

Subgroup Analysis

We conducted the subgroup analyses in 3 areas, including data partition (internal test or external test), model networks (TML or DL), and image (MRI or x-ray), to effectively understand how the 3 different types affected the performance of the algorithm for LSS assessment (Table 3). The internal test group demonstrated a lower sensitivity (*P*<.01) yet higher specificity (*P*<.01) than the external test group. Besides, the MRI group showed a lower sensitivity (*P*<.01) yet higher specificity (*P*<.01) than the x-ray group. The sensitivity in the DL group achieved 0.85, which was significantly higher than that (0.80) in the TML group (*P*<.01). Meanwhile, the DL group showed a more stable performance on specificity than the TML group (*P*=.04).

Table 3. Results of subgroup analysis.

Categories	Studies, n	Sensitivity (95% CI[a])	*P* value (HBG[b] of sensitivity)	Specificity (95% CI)	*P* value (HBG of specificity)
Data partition			<.001		<.001
Internal test	35	0.83 (0.80-0.86)		0.89 (0.85-0.92)	
External test	21	0.86 (0.82-0.90)		0.85 (0.79-0.91)	
Model networks			<.001		.04
TML[c]	6	0.80 (0.72-0.89)		0.87 (0.77-0.97)	
DL[d]	50	0.85 (0.82-0.87)		0.87 (0.84-0.91)	
Image			<.001		<.001
MRI	32	0.83 (0.79-0.86)		0.91 (0.88-0.93)	
X-ray	20	0.85 (0.82-0.89)		0.77 (0.70-0.84)	

[a] CI: confidence interval.
[b] HBG: heterogeneity between groups.
[c] TML: traditional machine learning.
[d] DL: deep learning.

Discussion

Principal Findings

In recent years, there has been a boom in assessing the diagnosis and grading of LSS by TML or DL methods. After systemically reviewing the available evidence, we revealed that all related studies were published after 2016 and increased annually. It can also be said that TML and DL algorithms have been showing promising potential in this field. To the best of our knowledge, this is the first systematic review and meta-analysis for addressing this issue. Our pooled results showed an overall sensitivity of 0.84 and a specificity of 0.87 for diagnosing LSS by TML or DL models. The area under the SROC was 0.92, indicating a high diagnostic value. Subgroup analysis revealed a better diagnostic performance in internal validation than in external validation, while DL algorithms demonstrated higher sensitivity and specificity than TML algorithms. However, 37% of studies enrolled in the systemic review were unavailable in the meta-analysis, which may have caused a discrepancy between pooled results and reality. Therefore, the results should be interpreted with caution.

A permanent debate focuses on whether the diagnostic performance of ML or DL algorithms surpassed that of clinicians. High-level evidence showed that the performance of AI diagnostic systems is equivalent to health care professionals, and AI-assistance systems improve clinician diagnostic performance [41-43]. However, in the field of LSS, there were few studies designed to directly compare the performance of additional radiologists or orthopedic surgeons with ML or DL algorithms in the same dataset. Hallinan et al [17] developed a DL method for diagnosing different LSS and compared the sensitivity and specificity of the DL model with 2 independent clinicians (a neuroradiologist and a musculoskeletal radiologist) with less than 10 years of experience. The study revealed that the sensitivity of DL in detecting LSS was on par with clinicians in general, with even slightly higher in lumbar central stenosis (LCS) and lateral recess stenosis (LRS), but with lower specificity of DL. It is reasonable because pursuing sensitivity to reduce false-negative results on the premise of maximizing the accuracy and AUC may be an alternative and beneficial method for clinical demands [44]. Compared with the complete replacement of clinicians, AI diagnostic systems are more expected to be assisted screening tools to use in areas with poor medical resources without experts or to reduce the workload of clinicians and missed diagnoses, followed by high-level medical team screening of image marked positive by the automatic diagnosis [45].

Although the general performance of diagnostic models was satisfactory, it was still insufficient enough for diagnosing or excluding LSS according to the summary likelihood ratio plot. Besides, our systemic review and meta-analysis found that ML or DL models showed similar, even slightly lower, sensitivity compared with specificity in general, especially in the MRI modality. There may be several reasons. First, the complexity and variety of pathological structures in individuals with LSS result in no broadly accepted quantitative radiologic evidence for diagnosis, even in expert evaluation [46], which makes automatic detection by MRI difficult. Furthermore, we cannot exclude that the results may be influenced by heterogeneity. Consideration should be taken for developers to optimize models prone to higher sensitivity than specificity for diagnosis and grading of LSS, which may be more beneficial to clinical workflow.

Notably, a consensus of reference standards in determining ML or DL performance for diagnosing LSS has not been reached till now. The reference standards in almost all included studies were labeled by qualitative or semiquantitative expert evaluation, which suffered from considerable heterogeneity due to the different amount, specialties, and years of experience of experts. Huber et al [22] combined texture analysis and decision trees to detect LSS based on the cross-sectional area (CSA) as a quantitative reference standard. However, a CSA of <130mm^2 was not a widely accepted criterion, and it is only appropriate for LCS, while quantitative radiological criteria remain unavailable for diagnosing LRS or LFS [46,47]. More comprehensive and rigorous criteria for reference standards should be developed in future work. In addition, the diagnosis of LSS should combine the imaging findings with history and clinical presentation because LSS is a clinical syndrome, and solely radiographic LSS may be symptom-free [2]. However, the diagnosis criteria in all reviewed studies were only based on radiographic criteria or reports, which means that the current TML or DL models were developed for the diagnosis of radiographic LSS objectively. Yet, it is not said that radiographic evaluation is valueless for LSS. On the one hand, it can provide details in pathological anatomy, which guides further treatment options and surgical approaches. On the other hand, a potentially imperceptible relationship between radiographic characters and clinical LSS may be explored with the help of AI models. Therefore, we suggest attempting to label the data by clinical LSS as golden standards on the premise of model interpretability and eliminating confounding factors. Furthermore, developers can set multiple data types, such as crucial details of patient's history, physical examination, and imaging tests, as inputs to build a multimodal to improve the clinical value of LSS diagnosis and grading by AI approaches [48].

Overall, our meta-analysis revealed a better performance for diagnosing LSS in DL than in TML. Whereas results should be interpreted with caution because of the limited number of enrolled studies on TML in meta-analysis. Only 2 included studies in the systemic review designed a direct comparison of the capability of DL and TML models for diagnosing LSS, yet showed contradictory results. Altun et al [28] found that VGG16 and 3 other DL techniques performed better in addressing the issue of binary LSS classification compared with random forest and support vector machines. Conversely, Bharadwaj et al [29] combined segmentation with DL and TML classifiers to conduct multiclass and binary LSS grading. Both accuracy, AUC, and reproducibility were higher in the TML group [29]. The inconsistency may be attributed to the scale of training data. DL was generally acknowledged as the most outstanding ML technique for automatic medical image analysis [49]. However, DL is restricted to a stronger data dependency compared with any other ML, as it is designed with a more complex architecture [14,15,50]. In particular, there is an extreme need for DL to be trained with a sufficiently large sample set, particularly considering the complexity of spinal MRI. In the 2 studies above, the scales of training samples were more than 5 times higher in the study of Altun et al [28] than that of Bharadwaj et al [29] (927 vs 170). Hence, we recommend that a larger scale dataset for training both TML and DL models is beneficial for exploring their capabilities in diagnosing and grading LSS in order to reduce data overfitting and improve the performance of models.

Any AI diagnostic systems should be clinically oriented instead of technically oriented. This poses a challenge for developers in developing ML or DL models more appropriately for clinical practice rather than more technically challenging. Currently, although promising results in exactitude (accuracy, sensitivity, etc) have been widely reported, other aspects of great clinical significance, like reliability, usability, and safety, were rarely assessed. A good agreement between the model and reference standard can verify the validity and reliability of the model. However, only 3 included studies performed a direct comparison of reliabilities between models and observers [17,29,32], and a generally higher consistency of diagnosis was achieved by clinicians than that of ML or DL systems. Besides, external validation is a valuable approach to validate the generalizability of ML or DL algorithms, testing their capabilities for adapting the differences between initial settings of data collection, imaging tests, and imaging processing with replication or real-world settings [45]. Inspiringly, this meta-analysis showed the performances of external validation were generally on par with that of the internal validation, with better sensitivity but worse specificity. However, the results may be inconclusive because only 37% of studies (7/19) enrolled in this systemic review tested the models by external validation (separate datasets for model validation only) [17,25-27,30,32,34].

Anyway, a gap remains between current ML or DL algorithms for diagnosing and grading LSS with real clinical applications. A recent review highlighted the importance of large-scale and mixed-source datasets, clinician collaboration, and a clear statement of data collection to facilitate DL in clinical applications [15]. Furthermore, only a few software, such as SpineNet (University of Oxford) [18,23,30] and CoLumbo (SmartSoft Ltd) [25], were introduced into public view despite several AI models having been developed in this field. We urged that the exploration of software design may be beneficial to extend the application of AI diagnostic models.

Limitations

Several limitations exist in this systematic review and meta-analysis. First, most of the enrolled studies were conducted under small sample sizes, and only 6 studies (32%) had a sample size >1000 [18,21,26,27,32,34]. However, a large-scale dataset is warranted for both training and validation in AI diagnostic algorithms, especially for DL algorithms [14,15,50]. Second, few models performed external validation to test the reproducibility and extensibility. Thus, the reported performance should be interpreted with caution. Third, only a few studies provided a contingency table, while the incompleteness of reported performance metrics made it difficult to conduct a comprehensive meta-analysis, which a recent systematic review and meta-analysis in the spine field also mentioned [51]. This may cause a discrepancy between pooled results and reality. Finally, the risk of bias in this study was identified by the QUADAS-2, which is more suitable for traditional diagnostic models [52]. A more specific and practical guideline for diagnostic AI models remains under development [53].

Conclusions

This systematic review and meta-analysis emphasize that despite the generally satisfactory diagnostic performance of artificial intelligence systems in the experimental stage for the diagnosis of LSS, none of them is reliable and practical enough to apply in real clinical practice. Further efforts, including optimization of model balance, widely accepted objective reference standards, multimodal strategy, large dataset for training and testing, external validation, and sufficient and scientific report, should be made to bridge the distance between current TML or DL models and real-life clinical applications in future studies.

Authors' Contributions

TW, RC, and NF contributed equally to this work. LZ contributed to the conception and design. TW, RC, and NF contributed to the acquisition of data. TW and NF contributed to the analysis and interpretation of data. TW and NF contributed to drafting the article. RC, NF, SY, PD, QW, AW, JL, XK, and WZ contributed to critically revising the article. LZ reviewed the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for diagnostic test accuracy) checklist.
[DOCX File , 32 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Details of search strategy.
[DOCX File , 13 KB-Multimedia Appendix 2]

References

1. Katz JN, Zimmerman ZE, Mass H, Makhni MC. Diagnosis and management of lumbar spinal stenosis: a review. JAMA. 2022;327(17):1688-1699. [doi: 10.1001/jama.2022.5921] [Medline: 35503342]
2. Lurie J, Tomkins-Lane C. Management of lumbar spinal stenosis. BMJ. 2016;352:h6234. [FREE Full text] [doi: 10.1136/bmj.h6234] [Medline: 26727925]
3. Ravindra VM, Senglaub SS, Rattani A, Dewan MC, Härtl R, Bisson E, et al. Degenerative lumbar spine disease: estimating global incidence and worldwide volume. Global Spine J. 2018;8(8):784-794. [FREE Full text] [doi: 10.1177/2192568218770769] [Medline: 30560029]
4. Kreiner DS, Shaffer WO, Baisden JL, Gilbert TJ, Summers JT, Toton JF, et al. North American Spine Society. An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update). Spine J. 2013;13(7):734-743. [doi: 10.1016/j.spinee.2012.11.059] [Medline: 23830297]
5. Lee S, Lee JW, Yeom JS, Kim KJ, Kim HJ, Chung SK, et al. A practical MRI grading system for lumbar foraminal stenosis. AJR Am J Roentgenol. 2010;194(4):1095-1098. [doi: 10.2214/AJR.09.2772] [Medline: 20308517]
6. 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 (Phila Pa 1976). 2010;35(21):1919-1924. [doi: 10.1097/BRS.0b013e3181d359bd] [Medline: 20671589]
7. Sigmundsson FG, Kang XP, Jönsson B, Strömqvist B. Correlation between disability and MRI findings in lumbar spinal stenosis: a prospective study of 109 patients operated on by decompression. Acta Orthop. 2011;82(2):204-210. [FREE Full text] [doi: 10.3109/17453674.2011.566150] [Medline: 21434811]
8. Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino JA, Kaiser J, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine (Phila Pa 1976). 2008;33(14):1605-1610. [FREE Full text] [doi: 10.1097/BRS.0b013e3181791af3] [Medline: 18552677]
9. Lee GY, Lee JW, Choi HS, Oh K, Kang HS. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. Skeletal Radiol. 2011;40(8):1033-1039. [doi: 10.1007/s00256-011-1102-x] [Medline: 21286714]
10. Bartynski WS, Lin L. Lumbar root compression in the lateral recess: MR imaging, conventional myelography, and CT myelography comparison with surgical confirmation. AJNR Am J Neuroradiol. 2003;24(3):348-360. [FREE Full text] [Medline: 12637281]
11. Park HJ, Kim SS, Lee YJ, Lee SY, Park NH, Choi YJ, et al. Clinical correlation of a new practical MRI method for assessing central lumbar spinal stenosis. Br J Radiol. 2013;86(1025):20120180. [FREE Full text] [doi: 10.1259/bjr.20120180] [Medline: 23426848]
12. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37(2):505-515. [FREE Full text] [doi: 10.1148/rg.2017160130] [Medline: 28212054]
13. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med. 2020;13(1):69-76. [FREE Full text] [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
14. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14. [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
15. Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal. 2022;79:102444. [FREE Full text] [doi: 10.1016/j.media.2022.102444] [Medline: 35472844]
16. He X, Yin Y, Sharma M, Brahm G, Mercado A, Li S. Automated diagnosis of neural foraminal stenosis using synchronized superpixels representation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI. Cham. Springer International Publishing; 2016.
17. Hallinan JTPD, 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):130-138. [doi: 10.1148/radiol.2021204289] [Medline: 33973835]
18. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. Med Image Anal. 2017;41:63-73. [doi: 10.1016/j.media.2017.07.002] [Medline: 28756059]
19. Zhang Q, Bhalerao A, Hutchinson C. Weakly-supervised evidence pinpointing and description. In: Information Processing in Medical Imaging. Cham. Springer International Publishing; 2017.
20. 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. [doi: 10.1016/j.media.2018.08.005] [Medline: 30176546]
21. Lu JT, Pedemonte S, Bizzo B, Doyle S, Andriole KP, Michalski MH, et al. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. 2018. Presented at: Proceedings of the 3rd Machine Learning for Healthcare Conference; August 17-18, 2018; Palo Alto, California. URL: https://proceedings.mlr.press/v85/lu18a.html

22. 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. [doi: 10.1016/j.ejrad.2019.02.023] [Medline: 31005175]
23. 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. [FREE Full text] [doi: 10.1186/s12891-020-3164-1] [Medline: 32164627]
24. 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):804-812. [doi: 10.1097/BRS.0000000000003377] [Medline: 31923125]
25. 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):902. [FREE Full text] [doi: 10.3390/diagnostics11050902] [Medline: 34069362]
26. 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;37(1):104-111. [doi: 10.3171/2021.11.SPINE211136] [Medline: 35061993]
27. 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. [FREE Full text] [doi: 10.3389/fendo.2022.890371] [Medline: 35733770]
28. Altun S, Alkan A, Altun İ. LSS-VGG16: diagnosis of lumbar spinal stenosis with deep learning. Clin Spine Surg. 2023;36(5):E180-E190. [doi: 10.1097/BSD.0000000000001418] [Medline: 36727890]
29. Bharadwaj UU, Christine M, Li S, Chou D, Pedoia V, Link TM, et al. Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI. Eur Radiol. 2023;33(5):3435-3443. [FREE Full text] [doi: 10.1007/s00330-023-09483-6] [Medline: 36920520]
30. 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):2137-2148. [doi: 10.1007/s00586-022-07311-x] [Medline: 35835892]
31. Shahzadi T, Ali MU, Majeed F, Sana MU, Diaz RM, Samad MA, et al. Nerve root compression analysis to find lumbar spine stenosis on MRI using CNN. Diagnostics (Basel). 2023;13(18):2975. [FREE Full text] [doi: 10.3390/diagnostics13182975] [Medline: 37761342]
32. Tumko V, Kim J, Uspenskaia N, Honig S, Abel F, Lebl DR, et al. A neural network model for detection and classification of lumbar spinal stenosis on MRI. Eur Spine J. 2024;33(3):941-948. [doi: 10.1007/s00586-023-08089-2] [Medline: 38150003]
33. Li KY, Weng JJ, Li HL, Ye HB, Xiang JW, Tian NF. Development of a dep-learning model for diagnosing lumbar spinal stenosis based on CT images. Spine (Phila Pa 1976). 2024;49(12):884-891. [doi: 10.1097/BRS.0000000000004903] [Medline: 38112156]
34. Park S, Kim JH, Ahn Y, Lee CH, Kim YG, Yuh WT, et al. Multi-pose-based convolutional neural network model for diagnosis of patients with central lumbar spinal stenosis. Sci Rep. 2024;14(1):203. [FREE Full text] [doi: 10.1038/s41598-023-50885-9] [Medline: 38168665]
35. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
36. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev. 2022;18(2):e1230. [FREE Full text] [doi: 10.1002/cl2.1230] [Medline: 36911350]
37. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
38. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods. 2020;11(2):181-217. [FREE Full text] [doi: 10.1002/jrsm.1378] [Medline: 31614060]
39. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536. [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
40. Nyaga VN, Arbyn M, Aerts M. Metaprop: a stata command to perform meta-analysis of binomial data. Arch Public Health. 2014;72(1):39. [FREE Full text] [doi: 10.1186/2049-3258-72-39] [Medline: 25810908]
41. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297. [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
42. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open. 2021;4(3):e211276. [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1276] [Medline: 33704476]

43. Xue P, Si M, Qin D, Wei B, Seery S, Ye Z, et al. Unassisted clinicians versus deep learning-assisted clinicians in image-based cancer diagnostics: systematic review with meta-analysis. J Med Internet Res. 2023;25:e43832. [FREE Full text] [doi: 10.2196/43832] [Medline: 36862499]
44. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889-897. [FREE Full text] [doi: 10.7326/0003-4819-149-12-200812160-00008] [Medline: 19075208]
45. Lex JR, Di Michele J, Koucheki R, Pincus D, Whyne C, Ravi B. Artificial intelligence for hip fracture detection and outcome prediction: a systematic review and meta-analysis. JAMA Netw Open. 2023;6(3):e233391. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.3391] [Medline: 36930153]
46. Mamisch N, Brumann M, Hodler J, Held U, Brunner F, Steurer J, et al. Lumbar Spinal Stenosis Outcome Study Working Group Zurich. Radiologic criteria for the diagnosis of spinal stenosis: results of a delphi survey. Radiology. 2012;264(1):174-179. [doi: 10.1148/radiol.12111930] [Medline: 22550311]
47. Andreisek G, Deyo RA, Jarvik JG, Porchet F, Winklhofer SFX, Steurer J, et al. LSOS working group. Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. Eur Radiol. 2014;24(12):3224-3232. [FREE Full text] [doi: 10.1007/s00330-014-3346-z] [Medline: 25079488]
48. Ramachandram D, Taylor GW. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. 2017;34(6):96-108. [doi: 10.1109/msp.2017.2738401]
49. Wu JH, Liu TYA, Hsu WT, Ho JHC, Lee CC. Performance and limitation of machine learning algorithms for diabetic retinopathy screening: meta-analysis. J Med Internet Res. 2021;23(7):e23863. [FREE Full text] [doi: 10.2196/23863] [Medline: 34407500]
50. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol. 2020;1213:3-21. [FREE Full text] [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
51. Compte R, Granville Smith I, Isaac A, Danckert N, McSweeney T, Liantis P, et al. Are current machine learning applications comparable to radiologist classification of degenerate and herniated discs and Modic change? A systematic review and meta-analysis. Eur Spine J. 2023;32(11):3764-3787. [FREE Full text] [doi: 10.1007/s00586-023-07718-0] [Medline: 37150769]
52. Zhang Z, Yang L, Han W, Wu Y, Zhang L, Gao C, et al. Machine learning prediction models for gestational diabetes mellitus: meta-analysis. J Med Internet Res. 2022;24(3):e26634. [FREE Full text] [doi: 10.2196/26634] [Medline: 35294369]
53. 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. [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]

Abbreviations

AI: artificial intelligence
CSA: cross-sectional area
CT: computed tomography
DL: deep learning
LCS: lumbar central stenosis
LFS: lumbar foraminal stenosis
LR: likelihood ratio
LRS: lateral recess stenosis
LSS: lumbar spinal stenosis
MeSH: Medical Subject Headings
ML: machine learning
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
SROC: summary receiver operating characteristic
TML: traditional machine learning

Edited by Q Jin; submitted 18.11.23; peer-reviewed by Z Li, Y Wang, S Mao, A Claus; comments to author 31.05.24; revised version received 23.10.24; accepted 11.11.24; published 23.12.24

Please cite as:
Wang T, Chen R, Fan N, Zang L, Yuan S, Du P, Wu Q, Wang A, Li J, Kong X, Zhu W
Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis
J Med Internet Res 2024;26:e54676
URL: https://www.jmir.org/2024/1/e54676
doi: 10.2196/54676
PMID: