Original Article

Deep learning for identifying cervical ossification of the posterior longitudinal ligament: a systematic review and meta-analysis

Felix Corr[1,2]^, Dustin Grimm[2]^, Paul Leach[1,3]

[1]Faculty of Medicine and Health Sciences, University of Buckingham, Buckingham, UK; [2]Department of Spine Surgery, Isarklinikum Munich, Munich, Germany; [3]Department of Neurosurgery, University Hospital Wales, Cardiff, UK

Contributions: (I) Conception and design: F Corr, P Leach; (II) Administrative support: F Corr, P Leach; (III) Provision of study materials or patients: F Corr, D Grimm; (IV) Collection and assembly of data: F Corr, D Grimm, P Leach; (V) Data analysis and interpretation: F Corr, D Grimm; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Felix Corr, MD, MSc. Faculty of Medicine and Health Sciences, University of Buckingham, Yeomanry House, Hunter Street, Buckingham, MK18 1EG, UK; Department of Spine Surgery, Isarklinikum Munich, Munich, Germany. Email: Felix.corresponding@gmail.com; felix.corr@live.de.

Background: Ossification of the posterior longitudinal ligament (OPLL) is a significant contributor for unintentional durotomy following anterior spinal surgery, neural compression, and cervical myelopathy. While traditional diagnostic methods like plain radiography are commonly used, they may yield false negatives. The diagnostic accuracy and reliability of artificial intelligence methods for detecting this condition remain largely unexplored. This study aimed to systematically evaluate the performance of deep learning models (DLMs) in diagnosing and predicting cervical OPLL.

Methods: This systematic review assesses the utilization of DLMs in diagnosing and predicting OPLL. Inclusion criteria were defined as the use of DLM for the diagnosis and prediction of cervical OPLL in adult patients. Databases included PubMed, Google Scholar, Cochrane Library, ScienceDirect, and BASE. The risk of bias was assessed using the QUADAS-2 tool.

Results: Seven studies with a pooled sample size of 3,373 patients were included. The pooled accuracy, area under the curve, sensitivity, and accuracy are 0.93, 0.92, 0.88, and 0.9. DLM demonstrated superior diagnostic performance, outperforming human comparator groups in terms of sensitivity (0.86 *vs.* 0.77), specificity (0.98 *vs.* 0.74), and accuracy (0.89 *vs.* 0.76). The meta-analysis with a pooled sample size of 1,016 patients revealed the highest proportion of right-identified OPLL subtypes in the mixed- and continuous subtypes (0.93 and 0.87). Accuracy and sensitivity of DLM were higher in the upper compared to the lower cervical spine.

Conclusions: Despite limitations in methodological variations and deep learning challenges, the findings support integrating these models into diagnostic protocols. Their robust performance suggests potential value in clinical practice, offering improved diagnostic accuracy and enhanced subtype differentiation.

Keywords: Artificial intelligence; cervical spine; deep learning; myelopathy

Submitted Jul 31, 2024. Accepted for publication Dec 31, 2024. Published online Feb 26, 2025.
doi: 10.21037/qims-24-1485
View this article at: https://dx.doi.org/10.21037/qims-24-1485

^ ORCID: Felix Corr, 0000-0002-5365-7511; Dustin Grimm, 0000-0003-2664-3568.

Introduction

Ossification of the posterior longitudinal ligament (OPLL) is characterized by the ectopic formation of calcified tissue along the posterior longitudinal ligament of the cervical spine and represents a significant etiological factor in the development of neural compression, cervical myelopathy, and radiculopathy (1). Among 106 individuals with both OPLL and cervical spinal cord injury (SCI), 88.7% experienced central cord syndrome without an underlying fracture, indicating that OPLL and the resulting cervical stenosis may increase susceptibility to SCI following minor trauma (2).

OPLL poses several challenges. First, anterior cervical approaches in patients with OPLL assimilating into the dura may lead to iatrogenic durotomy, intraoperative SCI, and poor postoperative outcomes (3). Recent findings of a meta-analysis demonstrated that in the context of OPLL, anterior approaches exhibited a higher incidence of postoperative neurological deficits (2.17% compared to 1.11%) and iatrogenic durotomy (3.74% compared to 0.96%) when compared to posterior approaches (4). OPLL masses may induce canal stenosis in anatomical regions beyond the reach of anterior surgical approaches, rendering this surgical approach unsuitable for a subset of patients (5). Similarly, an increasing number of patients are undergoing surgical treatment for cervical myelopathy solely based on magnetic resonance imaging (MRI) findings (5). Consequently, if left unaddressed, cervical OPLL can give rise to significant complications and pose considerable surgical challenges.

Cervical (lateral) radiography has been the initial diagnostic approach for OPLL. While it provides valuable initial information, false negative results may occur up to 48%, especially when OPLL is in its incipient stages or when imaging techniques are suboptimal (6). Computed tomography (CT) scans have gained prominence in recent years in detecting OPLL, offering a more detailed assessment of the ossification extent and morphology (7). Cervical OPLL is estimated at around 2% in Japan, 0.12% in the United States, and 0.1% in Germany when identified through cervical radiography (8). However, when detected through CT, the prevalence rises to 6.3% in Japan and 2.2% in the United States, highlighting the challenge of accurate OPLL diagnosis on plain radiography (9,10). Yet, a significant issue with CT imaging is the radiation exposure, requiring a careful balance between diagnostic precision and patient safety (11).

With advancements in medical technology and the emergence of deep learning models (DLM), particularly convolutional neural networks (CNNs), there has been a growing interest in exploring more sophisticated and accurate diagnostic tools to supplement or potentially enhance the efficacy of traditional radiological imaging (12). CNNs represent a fundamental component of deep learning architectures, particularly tailored for processing and analyzing visual data. As an advanced class of artificial neural networks, CNNs are specifically designed to extract and identify intricate patterns, structures, and features within images (13). The underlying structure of a CNN encompasses multiple layers, including convolutional, pooling, and fully connected layers, which collectively facilitate the automatic learning and extraction of hierarchical representations from input data (14). Utilizing its capability to identify complex visual features and hierarchies in images, the implementation of CNNs shows potential for enhancing the precision and effectiveness of OPLL detection.

This systematic review aimed to evaluate the current evidence surrounding the utilization of deep learning methodologies, particularly CNNs, in the diagnosis and outcome prediction of OPLL using various imaging modalities. We present this article in accordance with the PRISMA-DTA reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1485/rc).

Methods

Eligibility criteria

The inclusion of registered articles was limited to those written in English, conducted on human subjects, and published in peer-reviewed journals. Inclusion criteria for full review were: (I) adult patients; (II) use of DLM; (III) outcome measures (e.g., accuracy, sensitivity, specificity). Exclusion criteria were: (I) animal studies; (II) reviews (Table S1).

Information sources and search strategy

The study was registered in PROSPERO; with the registration number CRD42023445416. The literature search was conducted using the databases PubMed, Google Scholar, Cochrane Library, ScienceDirect, and Bielefeld Academic Search Engine (BASE). Keywords used were “deep learning” and “ossification posterior longitudinal ligament". Search terms were combined with two Boolean operators: AND, OR. The search strategy (Table S2) was peer-reviewed by D.G. and P.L. using the Peer Review of Electronic Search Strategies (PRESS) checklist. No search filters were applied.

Selection and data collection process

The Google Scholar search yielded 106 results, the ScienceDirect search ten, PubMed and BASE six, respectively, and the Cochrane Library one result. In total, 129 articles from databases were identified (as of October 1, 2023). As an additional search strategy references of the selected papers and other reviews were scanned. Titles and abstracts were screened for the inclusion and exclusion criteria by the authors. All authors screened the same publications and discussed the results. If these matched the inclusion criteria, the full-text article was screened for quality using the Oxford Centre of Evidence-Based Medicine (OCEBM) Levels of Evidence Table (Table S3), as well as GRADE scoring (Table S4).

The following information was collected in the consequent data extraction process: (I) author name; (II) publication year; (III) country; (IV) study design; (V) sample size; (VI) DLM; (VII) OPLL subtypes; (VIII) radiological techniques; (IX) outcome measure; (X) results (Table S5). Duplicates were removed. At the end of the selection process, seven articles were included in this review. The evidence will be presented in narrative format, tables, and visual presentation.

Data analysis

Data was collected using Excel Version 16.01 (Microsoft, Redmond, WA, USA). The respective corresponding author was contacted in case of missing data. All articles were critically appraised, and the risk of bias was determined against all the domains of the QUADAS-2 tool by two independent reviewers (F.C. and D.G.) (Table S6). Any disagreements were resolved by consensus after discussion with a third reviewer (P.L.). Definitions of outcome metrics and OPLL subtypes used by each study are shown in Tables S7,S8. A funnel plot was used to visually assess publication bias, complemented by Egger's test ($P<0.05$ indicating significant asymmetry), and the Trim-and-Fill method was applied to adjust for any detected bias. Data preparation, statistical analysis, and forest plot synthesis were performed using the meta package with the R software (version 4.0.4) and GraphPad Prism Version 10.0.3 (GraphPad Software, Inc., San Diego, CA, USA). A meta-analysis was conducted to assess the accuracy of DLM in correctly confirming OPLL subtypes. The R Code is available in Table S9. As the entire dataset required for summarizing receiver operating characteristic (ROC) curves was unavailable, we manually extracted individual ROC graphs, as previously described (15). Descriptive data are presented as means with standard deviations (SD), while outcome data are presented with 95% confidence intervals (CIs). Baseline balance is shown as absolute standardized differences for continuous data, defined as the absolute difference in means divided by the pooled SD.

Data availability

The Supplementary Digital Content accompanying this study contains all pertinent data supporting our findings. Furthermore, a comprehensive dataset employed in the study is openly accessible on the public GitHub repository. For access, please visit the following link: https://github.com/flxcorr/CNN_OPLL. We highly encourage researchers and interested individuals to leverage these resources for their own investigations and analyses.

Results

Study characteristics

In total, 129 studies were initially identified. Duplicates (n=29) were removed. From these, twelve full texts were assessed using our inclusion criteria. A total of seven studies were included in this systematic review. From these, three studies were included in the meta-analysis. The flow chart of data selection is shown in *Figure 1*.

The total pooled sample size of the systematic review was 3,373, and the overall pooled sample size of the meta-analysis was 1,016 patients. The studies included in the analysis were conducted in three countries: Japan (n=5), South Korea (n=1), and Israel (n=1). The publication dates of included studies, sample sizes, and risk of bias analysis can be found in *Figure 2*.

Out of the seven included studies, all were deemed to have a 'low' risk of bias using the QUADAS-2 tool (5,16-21). A scoring table for the risk of bias and applicability concerns as per QUADAS-2 domains is available in Table S8, a graphical summary in *Figure 2F*.

The OCEBM guidance was used to determine each study's evidence level. Six out of seven studies were classified

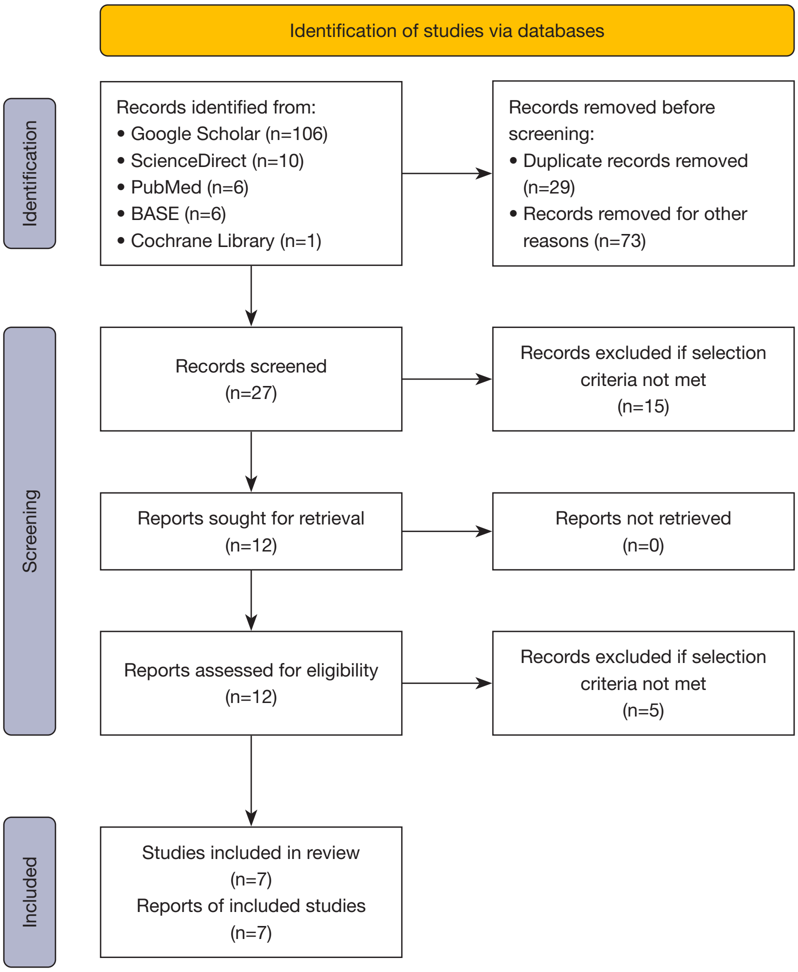

Identification of studies via databases
Identification
Records identified from:
• Google Scholar (n=106)
• ScienceDirect (n=10)
• PubMed (n=6)
• BASE (n=6)
• Cochrane Library (n=1)
Records removed before screening:
• Duplicate records removed (n=29)
• Records removed for other reasons (n=73)
Screening
Records screened (n=27)
Records excluded if selection criteria not met (n=15)
Reports sought for retrieval (n=12)
Reports not retrieved (n=0)
Reports assessed for eligibility (n=12)
Records excluded if selection criteria not met (n=5)
Included
Studies included in review (n=7)
Reports of included studies (n=7)

Figure 1 PRISMA flow chart for the systematic review detailing the database searches, the number of records screened, and the studies included.

as level 3 (5,16,18-21), and one as level 2b (17) (Table S5). The GRADE scoring is shown in Table S6 and showed that all studies scored as moderate. The study characteristics and main findings are demonstrated in *Table 1*.

DLM development

Various studies uniformly defined criteria for including and excluding patients in their investigations. A methodological consensus was observed in the reliance on CT scans to confirm the diagnosis of OPLL (5,16-21). Certain studies further refined patient selection by incorporating specific OPLL thickness criteria as inclusion parameters, while Ogawa *et al.* and Tamai *et al.* prioritized including patients with symptomatic OPLL (20,21).

A recurring exclusion criterion in six out of seven studies was the prior history of cervical surgery (16-21). Additionally, spinal fractures or trauma, tumors, kyphotic deformity, and atlantoaxial subluxation were consistently considered as exclusion criteria in the studies by Chae *et al.* and Tamai *et al.* (16,21). The inclusion and exclusion criteria are summarized in *Table 2*, with a visual representation in *Figure 3*.

Six out of seven studies investigated the detection of OPLL using various imaging techniques. Among them, most included studies utilized plain radiography (16-21), while Shemesh *et al.* focused specifically on MRI (5). The summary of the overall methodology for each study is presented in *Table 3*, with a visual representation in *Figure 3*.

Regarding image preprocessing in plain radiography

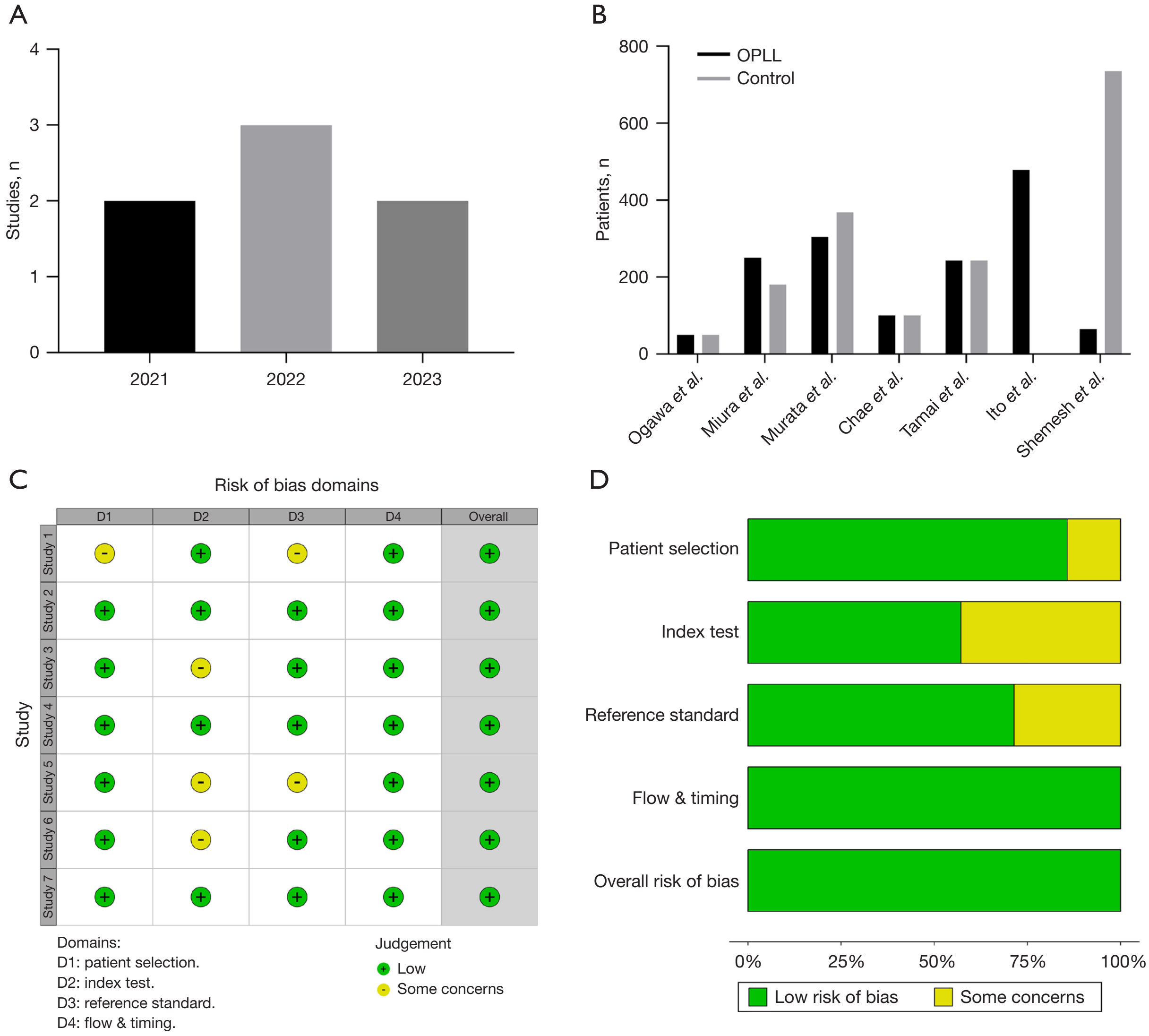


Figure 2 Study characteristics. (A) A bar chart displaying publication dates: 2021 (n=2), 2022 (n=3), 2023 (n=2). (B) A bar chart showing the sample sizes included in each study grouped into OPLL (black) and control (gray) groups. (C) A risk of bias traffic light plot for the seven included studies across the domains of the QUADAS-2 tool. Study 1 = Ogawa *et al.*, Study 2 = Miura *et al.*, Study 3 = Murata *et al.*, Study 4 = Chae *et al.*, Study 5 = Tamai *et al.*, Study 6 = Ito *et al.*, Study 7 = Shemesh *et al.* (D) A risk of bias summary plot presenting the distribution of risk-of-bias judgments across the domains of the QUADAS-2 tool for all included studies, displayed as percentages (%). OPLL, ossification of the posterior longitudinal ligament.

studies, a consistent practice involved converting Digital Imaging and Communications in Medicine (DICOM) images to Joint Photographic Experts Group (JPEG) format, enhancing compatibility with DLMs (5,18-21). Moreover, manual segmentation of OPLL areas was applied to isolate and identify specific regions of interest in the images, thereby improving data interpretability and ensuring its suitability for training DLMs (5,18,21).

In the examined studies, the training phase emerged as a critical component, featuring the widespread application of transfer learning and data augmentation techniques. For instance, Miura *et al.* enhanced the OPLL detection model using the EfficientNetB4 architecture alongside diverse data augmentation methods, fostering versatility in handling various scenarios (18). Similarly, Tamai *et al.* opted for the EfficientNetB2 model to implement data augmentation

Table 1 Study characteristics of the included studies

Author	Year	Design	Radiology	OPLL (N)	Control (N)	Comparator group	Summary
Miura *et al.* (18)	2021	Retrospective	X-ray	250	180	Yes	CNN utilized to distinguish between cervical spondylosis, OPLL, and control lateral radiographs Performance equal/superior to comparator group Accuracy =0.86, sensitivity =0.86, precision =0.87, F1 score =0.87
Murata *et al.* (19)	2021	Retrospective	X-ray	304	368	No	Utilization of RNN for binary classification of OPLL on cervical lateral plain radiography Accuracy =0.989, sensitivity =0.97, specificity =0.994, FP =0.022, FN =0.01, AUC =0.99, CI: 0.97–1.00
Chae *et al.* (16)	2022	Retrospective	X-ray	100*	100*	Yes	DLM able to recognize OPLL on lateral cervical plain radiography Enhancement of diagnostic performance of comparator group Vertebra level: AUC =0.854, patient: AUC =0.851, sensitivity =0.91, specificity =0.69
Ogawa *et al.* (20)	2022	Retrospective	X-ray	50	50	Yes	Accuracy of CNN validated on cervical spine radiographs Excellent diagnostic performance of CNN (AUC =0.924, accuracy =0.9, sensitivity =0.8, specificity =1.0)
Tamai *et al.* (21)	2022	Retrospective	X-ray	243**	243**	Yes	Utilization of CNN for binary classification of OPLL on cervical lateral plain radiography Higher diagnostic performance of CNN compared to comparator group Accuracy =0.88, AUC =0.94, 95% CI: 0.92–0.97, sensitivity =0.9, precision =0.86
Ito *et al.* (17)	2023	Prospective	–	478	–	No	Utilization of CNN for postoperative outcome prediction in patients undergoing surgery with OPLL Accuracy =0.746 for overall complications; accuracy =0.917 for neurological complications
Shemesh *et al.* (5)	2023	Retrospective	MRI	65*	735*	No	Utilization of CNN for binary classification of OPLL on MRI Accuracy =0.98, kappa score =0.917, sensitivity =0.85, specificity =0.98, NPV =0.98, PPV =0.85

*, training *vs.* control. (I) Chae *et al.*: in total, 407 patients were utilized. Of those, 207 patients with cervical OPLL were used in the training-validation set, and 100 patients with and 100 patients without OPLL were utilized in the test set. (II) Shemesh *et al.*: a total of 800 patients were utilized. Of those, 65 patients with cervical OPLL and 735 healthy subjects were included, with 600 in the training group and 200 in the validation group. **, matched controls. AUC, area under the curve; CI, confidence interval; CNN, convolutional neural network; FN, false negative; FP, false positive; NPV, negative predictive value; MRI, magnetic resonance imaging; OPLL, ossification of the posterior longitudinal ligament; PPV, positive predictive value; RNN, residual neural network.

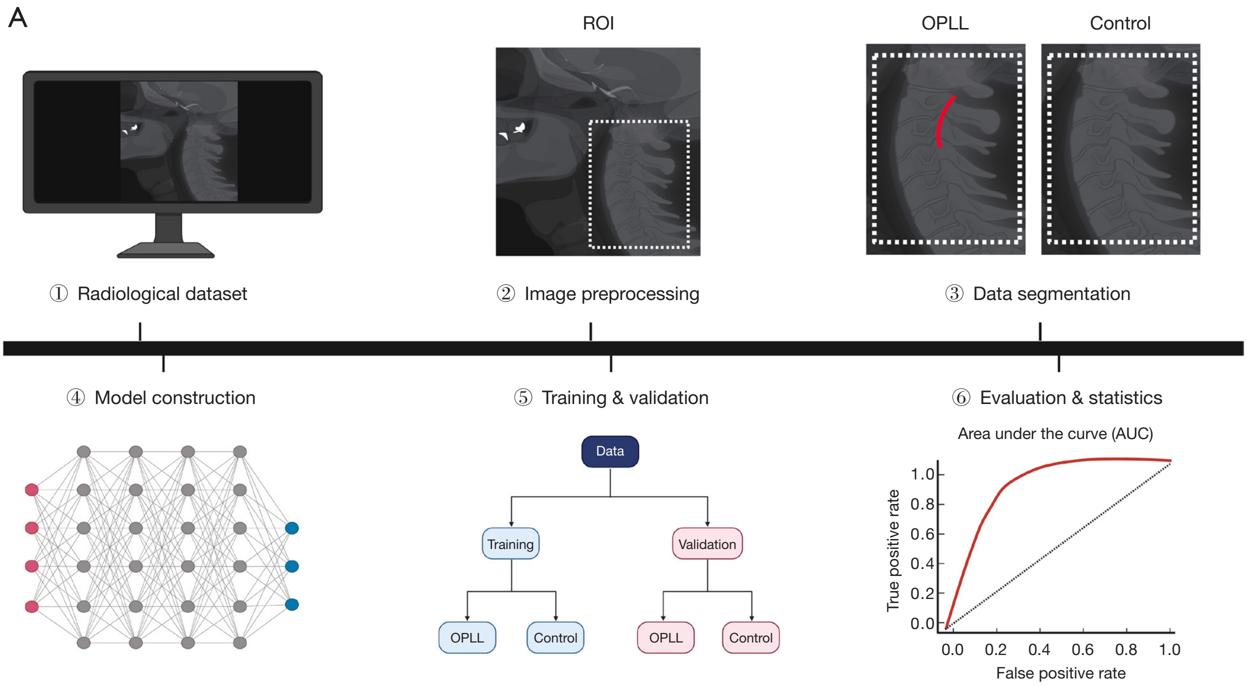
techniques like inversion, equalization, and mix-up (21). Chae *et al.* adopted a two-dimensional (2D) Residual U-net, incorporating rotation, blurring, sharpening, brightness changes, and Gaussian noise during training (16). Shemesh *et al.* introduced randomness in patient samples as a form of data augmentation, ensuring the model was trained on a representative set of scenarios (5).

The composition of the training datasets varied across the

Table 2 Specified inclusion and exclusion criteria of the included studies

Study	Inclusion criteria	Exclusion criteria
Miura *et al.* (18)	OPLL confirmed via CT scans; OPLL confirmed by surgeons; follow-up patients after OPLL diagnosis; control group with CT, MRI, and radiographs	Previous history of cervical surgery; severe kyphotic deformity; atlantoaxial subluxation; foreign body interference; obviously fused vertebrae; invisible C6 or C7 vertebra
Murata *et al.* (19)	OPLL confirmed via CT scans; continuous or mixed-type OPLL; 2000 to 2019	OPLL of segmental type; previous history of cervical surgery. Control group: spondylolisthesis; severe kyphotic deformity; neurological symptoms; ossification of PLL, ALL, nuchal ligament; cervical disc herniation; cervical spondylotic radiculopathy, myelopathy, amyotrophy
Chae *et al.* (16)	OPLL confirmed via CT scans; OPLL >2 mm thickness	Previous history of cervical surgery; fracture/trauma; spinal tumors; infectious spondylitis (cervical)
Ogawa *et al.* (20)	OPLL confirmed via CT scans; symptomatic OPLL; three-way radiography; OPLL >3 mm thickness; control: Surgery for lumbar spine diseases	Congenital malformations of the cervical spine; previous history of cervical surgery
Tamai *et al.* (21)	OPLL confirmed via CT scans; symptomatic OPLL	Previous history of cervical surgery; spinal tumors; fracture/trauma
Ito *et al.* (17)	OPLL confirmed via CT scans; patients >20 years old; confirmed spinal cord compression on MRI, undergoing surgery	Previous history of cervical surgery; comorbidities impairing physical function; OPLL excluded based on previously published criteria
Shemesh *et al.* (5)	OPLL confirmed via CT scans; OPLL confirmed via MRI; OPLL confirmed by surgeons; imaging for clinical indications	No specific exclusion criteria are specified

CT, computed tomography; MRI, magnetic resonance imaging; OPLL, ossification of the posterior longitudinal ligament; PLL, posterior longitudinal ligament; ALL, anterior longitudinal ligament.



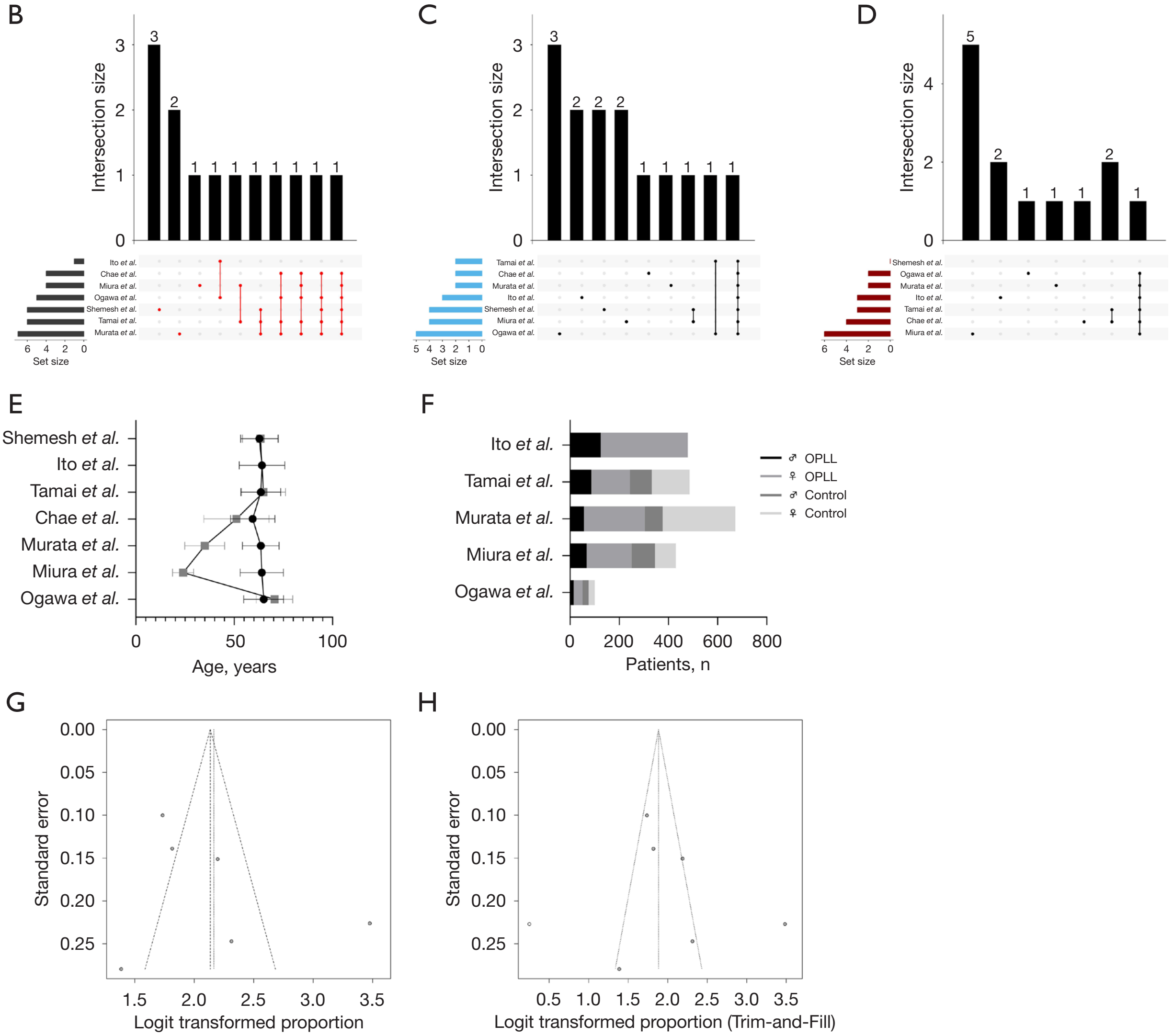


Figure 3 Methodological analysis of the included papers in the systematic review. (A) Consistent methodological workflow for the construction and evaluation of DLMs. 1, radiological datasets are exported from DICOM to JPEG format; 2, image preprocessing, where a ROI of the image (mainly lateral cervical plain radiography) is usually defined; 3, data segmentation, in which the OPLL region and dimensions are segmented using segmentation software (OPLL segmented indicated by a red line), and the images are labeled (OPLL or Control, respectively); 4, model construction of deep learning, notably the convolutional neural network algorithm, which is then [5] trained and validated using OPLL and control groups, and finally [6] evaluated statistically using quantifiable performance measures (created with BioRender.com). (B-D) Three upset plots comparing the similarities in methodology between individual studies. The "Set size" represents the number of each study concerning (B) outcome parameters, (C) inclusion criteria, and (D) exclusion criteria. The intersections are represented by the overlapping areas between the bars. The matrix indicates which combinations of sets have an overlap. Regarding outcome parameters, the highest consensus was noted for sensitivity, specificity, AUC, and accuracy. Regarding inclusion criteria, the identification of OPLL through CT scans emerged as the predominant shared characteristic. For exclusion criteria, a previous history of cervical spine surgery was the most commonly agreed upon factor. (E) A summary plot illustrating the age differences between the OPLL (black) and control (gray) groups for the respective studies. A significant age difference is evident, especially in Chae *et al.*, Murata *et al.*, and Mirura *et al.* (F) Representation of gender distribution between the OPLL and control groups across individual studies. (G) Original and (H) Trim-and-Fill corrected funnel plot that plots every study except for Ito *et al.* The logit-transformed proportion is plotted on the x-axis, and the standard error is plotted on the y-axis. In the corrected funnel plot, one study is imputed (open circle) to address asymmetry observed in the original plot. CT, computed tomography; DICOM, digital imaging and communications in medicine; DLMs, deep learning models; OPLL, ossification of the posterior longitudinal ligament; ROI, region of interest; JPEG, Joint Photographic Experts Group.

Table 3 Detailed methodology of the included studies

Study	Radiography	Methodology
Ogawa *et al.* (20)	Plain radiography	Study participants and input image
		Image preprocessing: DICOM export JPEG images; image composing (centered patch, 448×448 px, resized to 480×480 px; VGG16 architecture model; new classification layer; transfer learning; data augmentation (only training set)
		CNN model construction & performance evaluation: 50 OPLL + 50 control divided into training cases (n=40) and test data (n=10); 80 cases trained; radiograph types: neutral, flexion, extension; probability estimation for each radiograph; optimal threshold decision of probability for OPLL; control group evaluation (blinded); Grad-CAM to visualize areas of diagnostic evidence
		Statistical analysis: sensitivity, specificity, AUC, ROC curve
Miura *et al.* (18)	Plain radiography	Patients: confirmation via physicians (n=2)
		Radiological dataset: lateral cervical radiographs in neutral position
		Image preprocessing: DICOM to JPEG export; manual OPLL segmentation; ROI cropping (2:3)
		Model construction and training: efficientNetB4 model, 3-class classification layer; 380×380 px; training: epochs =100, learning rate =0.1; data augmentation (rotation, width shift, height shift, brightness)
		Performance evaluation: training 200 OPLL cases; validation (n=50); cross-examination through surgeons (n=5), blinded
		Statistical analysis: sensitivity, specificity, accuracy, F1 score
Murata *et al.* (19)	Plain radiography	Study population
		Plain radiography dataset: cervical lateral X-rays only; DICOM export JPEG (224×224 px, 8-bit grayscale)
		Image categorization and preprocessing: categorization into OPLL or control group; review of images
		Development of algorithm: Res-Net12 model using Batch normalization; K-fold cross-validation (k=5); external validation
		Algorithm evaluation: accuracy, sensitivity, specificity, false positive rate, false negative rate, ROC, AUC, confusion matrix
Chae *et al.* (16)	Plain radiography	Patients: 307 patients into validation (n=207) and test set (n=100); additional control set for test set (n=100); multiple radiographs used from same patient for training-validation set, but only one image for test set
		Data labeling and DLM development: manual segmentation of OPLL on radiograph; ground truth read on CT, recording of presence of OPLL lesion for each vertebra; classification of OPLL subtypes; 2D residual U-net with atrous spatial pyramid pooling model; images (288×288 px, stride value 32); initial filter size =32, doubled after each convolutional layer; Adam optimizer (β1 =0.9, β2 =0.999 and eps =1×10^{-8}) and Dice coefficient loss function; Learning rate =0.0003, decay rate =0.9/2,000 steps, epochs =300, batch size =5, dropout rate =0.15; data augmentation (rotation, blurring, sharpening, brightness change, Gaussian noise)
		Evaluation: per-vertebra and per-patient analysis; per-vertebra = C-spine ROI from C2–C7; performance assessment: lowest threshold at OPLL lesion first appeared afterward, increasing the threshold from 0.1 to 0.9 (0.1 increments); observer performance test
		Statistical analysis: AUC, ROC, sensitivity, specificity

Table 3 (*continued*)

Table 3 (*continued*)

Study	Radiography	Methodology
Tamai *et al.* (21)	Plain radiography	Data collection: DICOM to JPEG images (224×224 px)
		Labeling process: segmentation of OPLL area; divided into mask images for ground truth and original image
		Algorithm development: data augmentation (inversion, equalization, brightness, gamma correction, histogram, noise addition, mix-up); efficientNetB2 model; ten-fold cross-validation; 10 groups: training (n=9), validation (n=1), 10× repetition
		Algorithm validation: true positives, false positives, false negatives, true positives, accuracy, sensitivity, specificity, ROC, AUC; sub-analysis institution, OPLL type, OPLL location
		Comparison to surgeons: surgeon (n=4); individual evaluation cervical radiography images (n=50), OPLL (n=25), control (n=25)
Shemesh *et al.* (5)	MRI	Study population: OPLL identification through physicians (n=3)
		Image preprocessing: MRI DICOM files (axial plane, T2w); randomization of patient samples; data augmentation; image segmentation (semiautomatic threshold); review of segmented images; feature extraction (area, perimeter, voxel, OPLL thickness, location, shape, intensity of ossified tissue)
		Algorithm development: VGG16 model; training using extracted features and corresponding labels; data split into training (75%) and validation (25%); review of OPLL patients identified by model by surgeons (n=2); reader study
		Evaluation: sensitivity, specificity, negative predictive value, positive predictive value, accuracy, Cohen's kappa score, confusion matrix
Ito *et al.* (17)	–	Patient selection: multicenter, prospective
		Outcome measures: primary: postoperative complications
		Deep learning-based prediction model: deep neural network; 39 preoperative factors for each patient as input variables; two hidden layers (n=15, and n=10, respectively)
		Performance evaluation: five-fold cross-validation; subgroups divided in equal-sized based on data; data augmentation; training with 100 epochs, mini-batch size 4
		Statistical analysis: comparison to logistic regression analysis; accuracy

AUC, area under the curve; CNN, convolutional neural network; DICOM, Digital Imaging and Communications in Medicine; Grad-CAM, gradient-weighted class activation mapping; JPEG, Joint Photographic Experts Group; MRI, magnetic resonance imaging; OPLL, ossification of the posterior longitudinal ligament; px, pixel; ROC, receiver operating curve; ROI, region of interest; T2w, T2-weighted.

studies, ranging from 80 to 200 OPLL cases. The average size of the study population was 481.9±220.4 (range, 100–800), with 212.9±153.8 (range, 50–478) and 236.9±245.2 (range, 50–717) for the OPLL and Control groups, respectively. The average age of individuals in the OPLL group was 63.2±1.8 (range, 59.4–65), while in the Control group, it was 51.5±18.5 (range, 24–70.5). In total, the count of women and men in the OPLL and Control groups was 424 *vs.* 1,066 and 910 *vs.* 748, respectively. Examining the percentage distribution of genders, there was a higher representation of women in the Control group (47.7%±20.3%; range, 19.8–81.2%) compared to the OPLL group (33.7%±16.5%; range, 18.4–69.2%). The patient characteristics concerning age and gender distribution between the OPLL and Control data sets are presented in *Table 4*.

Evaluation metrics serve as quantitative measures to assess the performance of the DLMs. Across the range of studies examined a diverse set of evaluation metrics has been applied to evaluate diagnostic performance. Notably, sensitivity (5,16,18-21) and specificity (5,16,19-21) emerged as the most commonly employed metrics. The ROC and area under the curve (AUC) curves also appeared recurrently, being utilized in five of the seven studies (16,18-21). Accuracy, providing a fundamental measure of overall correctness in classifications, was consistently assessed in five out of seven studies (5,17-21).

Table 4 Patient characteristics of included studies

Study	OPLL group		Control group		SMD
	Age (years)	Female (%)	Age (years)	Female (%)	
Ogawa *et al.* (20)	65±10.10	30	70.5±9.30	52.5	0.44
Miura *et al.* (18)	64±11.00	26.8	24±5.4	47.2	3.17
Murata *et al.* (19)	63.5±9.3	18.4	35±10.1	19.8	2.26
Chae *et al.* (16)	59.4±11.3	32	51.1±16.6	46	0.66
Tamai *et al.* (21)	63.5±10.1	35.39	64.9±11.2	36.62	0.11
Ito *et al.* (17)	64.1±11.6	25.9	–	–	–
Shemesh *et al.* (5)	62.79±9.49	69.23	63.38±9.2	79.18	0.05
Pooled	63.18	33.96	51.48	46.88	1.01

Data for age are presented as mean ± standard deviation. Female representation is provided as percentages. Missing values are indicated with a dash (–). OPLL, ossification of the posterior longitudinal ligament; SMD, standardized mean difference.

Model performance

All studies assessed the overall performance of the DLM using a variety of metrics that reflect the comprehensive performance. These metrics spanned various aspects of model performance including sensitivity, specificity, accuracy, and the AUC analyses. By considering multiple metrics, all studies provided a comprehensive perspective on the robustness and reliability of the model concerning overall performance (see *Figure 4*).

Four out of seven studies assessed the AUC of the DLM (16,19-21). The pooled AUC is 0.93%±0.06% (ranging from 0.85% to 0.99%). Murata *et al.* reported the highest accuracy (0.99), followed by Tamai *et al.* (0.94), Ogawa *et al.* (0.924), and Chae *et al.* (0.851). The summary of the ROC curves is depicted in *Figure 4A*.

Six out of seven studies evaluated the overall accuracy of the DLM (5,17-21). The pooled accuracy is 0.92±0.05 (ranging from 0.86 to 0.98). Murata *et al.* reported the highest accuracy (0.988), followed by Shemesh *et al.* (0.98), Ito *et al.* (0.917), Ogawa *et al.* (0.9), Tamai *et al.* (0.88), and Miura *et al.* (0.86). The pooled accuracy for plain radiographs alone is 0.91±0.05 (ranging from 0.86 to 0.98).

Six out of seven studies evaluated the overall sensitivity of the DLM (5,16,18-21). The pooled sensitivity is 0.88±0.06 (ranging from 0.8 to 0.97). The highest sensitivity was reported by Murata *et al.* at 0.97, followed by Chae *et al.* (0.91), Tamai *et al.* (0.9), Miura *et al.* (0.86), Shemesh *et al.* (0.85), and Ogawa *et al.* (0.8). The pooled sensitivity for plain radiographs alone is 0.88±0.06 (ranging from 0.86 to 0.97).

Furthermore, five out of seven studies assessed the overall specificity of the DLM (5,16,19-21). The pooled specificity stands at 0.9±0.133 (ranging from 0.69 to 1.0). The highest specificity was reported by Ogawa *et al.* at 1.0, followed by Murata *et al.* (0.994), Shemesh *et al.* (0.98), Tamai *et al.* (0.86), and Chae *et al.* (0.69). The pooled specificity for plain radiographs alone is 0.88±0.15 (ranging from 0.86 to 1.0).

Comparator group differences

In addition to assessing the performance of the DLM, its efficacy was also compared with that of human control groups, represented by surgeons, in terms of accuracy, sensitivity, specificity, and precision. The surgeons varied in experience: Miura *et al.* included board-certified spine surgeons with 11 to 21 years of experience, Ogawa *et al.* had surgeons with over 5 years of general orthopedic or over 10 years of spinal specialization, and Tamai *et al.* included surgeons with 5, 10, 20, and 25 years of spinal surgery experience. An overview of the diverse results is presented in *Table 5*, with the overall outcomes depicted in *Figure 5*.

Pooled results indicate that the CNN model outperformed surgeons in all metrics. The pooled mean sensitivity was 0.86±0.06 (range, 0.8–0.92) for the CNN versus 0.77±0.08 (range, 0.6–0.9) for the comparator group, while the pooled mean specificity was 0.98±0.02 (range, 0.96–1.0) and 0.74±0.13 (range, 0.5–0.9) for the CNN and comparator group, respectively. The pooled mean accuracy was 0.89±0.03 (range, 0.86–0.92) and 0.76±0.06 (range,

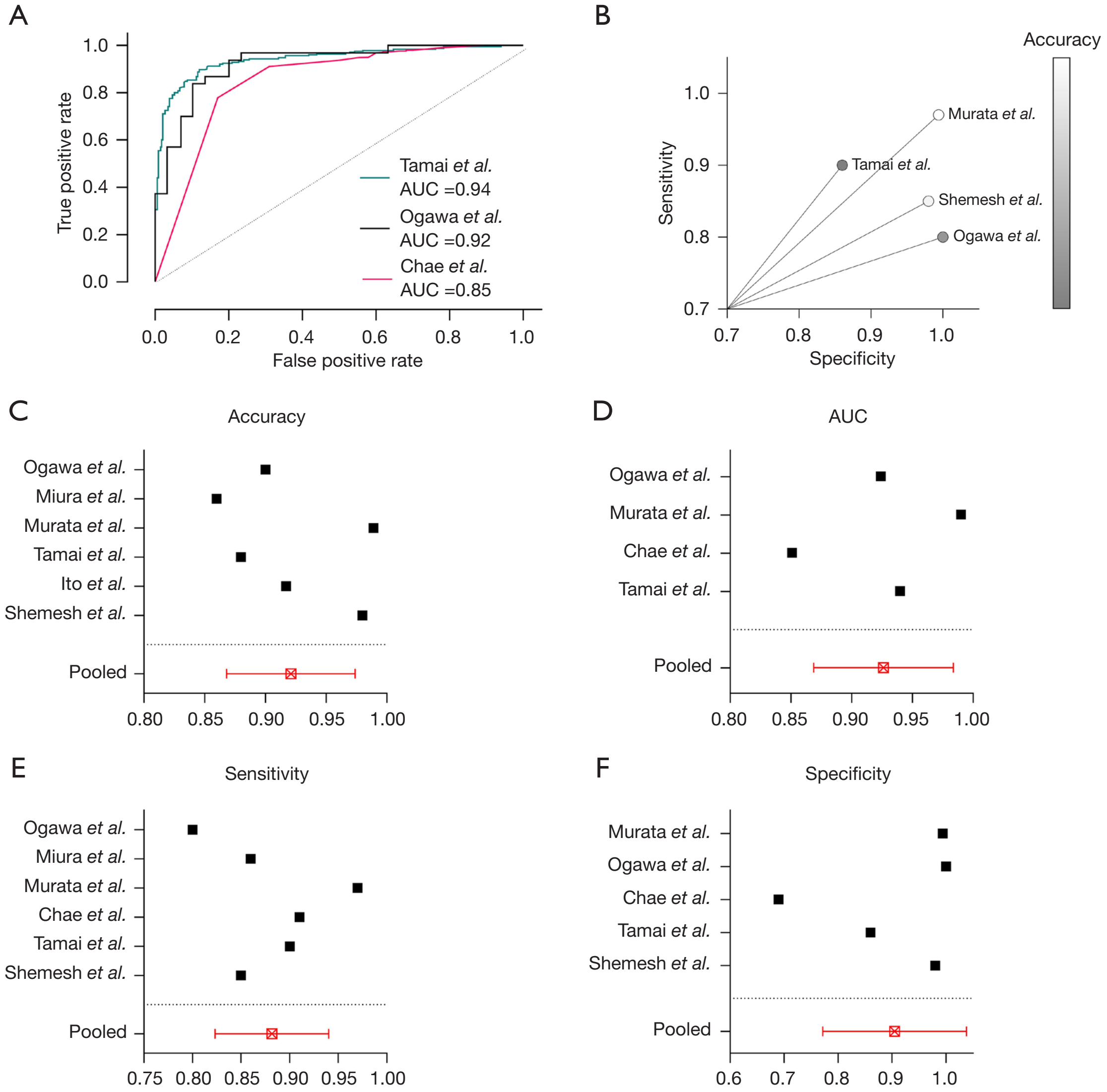


Figure 4 Main outcome parameters of DLM evaluation. (A) Summary of three ROC curves. (B) Scatter plot of primary studies. Specificity is depicted on the x-axis, while sensitivity is on the y-axis. The color-coded bar on the right represents the accuracy of each study. (C) Representation of the mean accuracy of each study, with pooled overall accuracy shown in red, accompanied by standard deviation as error bars. (D) Illustration of the mean AUC, pooled overall AUC depicted in red, and standard deviation as error bars. (E) Presentation of the mean sensitivity, with pooled overall sensitivity shown in red, and standard deviation as error bars. (F) Depiction of the mean specificity, pooled overall specificity presented in red, and standard deviation as error bars. AUC, area under the curve; DLM, deep learning model; ROC, receiver operating characteristic.

0.6–0.8) for the CNN and comparator group, respectively. In addition to sensitivity and specificity precision and F1 scores were assessed in one study, revealing superior values for the CNN model compared to the comparator group (precision: 0.87 *vs.* 0.81±0.03; F1 score: 0.87 *vs.* 0.82±0.02, respectively).

Beyond the direct comparison between the CNN model and the human control group, the CNN model was employed for a second session after a solitary session, during which surgeons utilized the CNN model as an assessment aid and support. An added-on effect in identifying OPLL on plain radiographs was demonstrated, irrespective of the surgeons' experience level. Specifically, the AUC of observers increased to 0.893 (P=0.001) in per-vertebra

Table 5 Summary of accuracy, sensitivity, and specificity among the CNN models and human comparison groups (surgeons)

Study	Group	Accuracy	Sensitivity	Specificity
Miura *et al.* (18)	CNN	0.86	0.86	0.87*
	Surgeon	0.83	0.83	0.82*
	Surgeon	0.83	0.83	0.83*
	Surgeon	0.81	0.81	0.81*
	Surgeon	0.81	0.81	0.84*
	Surgeon	0.76	0.76	0.76*
Ogawa *et al.* (20)	CNN	0.90	0.80	1.0
	Surgeon	0.60	0.60	0.60
	Surgeon	0.70	0.90	0.50
	Surgeon	0.80	0.70	0.90
	Surgeon	0.80	0.80	0.80
	Surgeon	0.75	0.90	0.60
	Surgeon	0.70	0.70	0.70
Tamai *et al.* (21)	CNN	0.92	0.923	0.958
	Surgeon	0.80	0.759	0.857
	Surgeon	0.74	0.714	0.773
	Surgeon	0.76	0.724	0.810
	Surgeon	0.78	0.719	0.889

Asterix (*) for Miura *et al.*, not the specificity but values for precision are reported. CNN, convolutional neural network.

analysis and 0.911 (P<0.001) in per-patient analysis.

CNN model performance among OPLL subtypes

OPLL involves a wide range of pathological subtypes, each distinguished by unique patterns of ligament ossification within the spinal column. Some of the included studies have explored the effectiveness of the CNN model in distinguishing between various OPLL subtypes. Among the seven studies conducted, three specifically examined the accuracy of the CNN model in correctly identifying cases within distinct OPLL subgroups, as depicted in *Figure 6*.

For the continuous subtype, 26 out of 30 subtypes were correctly identified. The pooled proportion of correctly identified cases is 0.87 (range, 0.5–0.83; 95% CI: 0.69–0.95; I^2=5%; P=0.35) (18,20,21). For the segmental subtype, 70 out of 80 subtypes were correctly identified. The pooled proportion of correctly identified cases is 0.82 (range, 0.72–0.85; 95% CI: 0.73–0.89; I^2=36%; P=0.21) (18,21). For the mixed subtype, 129 out of 137 subtypes were correctly identified. The pooled proportion of correctly identified cases is 0.93 (range, 0.8–0.96; 95% CI: 0.82–0.98; I^2=55%; P=0.11) (18,20,21). For the localized subtype, 40 out of 50 subtypes were correctly identified (18,20,21). Overall, the pooled proportion of correctly identified cases is 0.8 (range, 0.5–1.0; 95% CI: 0.67–0.89; I^2=0%; P=0.37). The pooled overall prediction using the random effects model for 302 OPLL subtypes is 0.86 (95% CI: 0.78–0.91).

Furthermore, the performance of the CNN was assessed through the AUC in a per-vertebra analysis of the CNN model (16). The highest values were observed for the continuous subtype (0.897, 95% CI: 0.839–0.956), followed by the mixed subtype (0.881, 95% CI: 0.845–0.916), segmental subtype (0.825, 95% CI: 0.786–0.864), and localized subtype (0.819, 95% CI: 0.637–1.001), thus highlighting the varying degrees of discriminative ability of the CNN model across different OPLL subtypes, with the continuous subtype exhibiting the most robust performance.

In addition to evaluating the standalone performance of the CNN model, the augmentation of observer performance concerning OPLL subtypes was also examined. The observers exhibited enhanced capability in identifying OPLL subtypes when aided by CNN models. The improvement in diagnostic performance was most pronounced in the segmental type (AUC difference 0.087; P=0.002), while the increment was comparatively smaller in the continuous type (AUC difference 0.026; P=0.026) (16).

Anatomical localization of OPLL

OPLL can manifest at various levels along the cervical spine aside from its pathological configuration. The impact of anatomical localization on the DLM performance was assessed by two out of seven studies (16,21) and is depicted in *Figure 7*.

In the analysis of the AUC, superior values were observed in the upper cervical spine at C2 (0.932, 95% CI: 0.862–1.001), C3 (0.904, 95% CI: 0.840–0.968), and the highest scores in C4 (0.906, 95% CI: 0.862–0.949), as opposed to the lower cervical spine starting from C5 (0.865, 95% CI: 0.812–0.918), C6 (0.829, 95% CI: 0.765–0.893), with the lowest scores at C7 (0.582, 95% CI: 0.487–0.677) (16). Notably, the AUC at C7 was significantly poorer compared to the average observer AUC (0.793, 95% CI: 0.675–0.911) with statistical significance (P>0.001) (16).

Similarly, when considering sensitivity, higher values

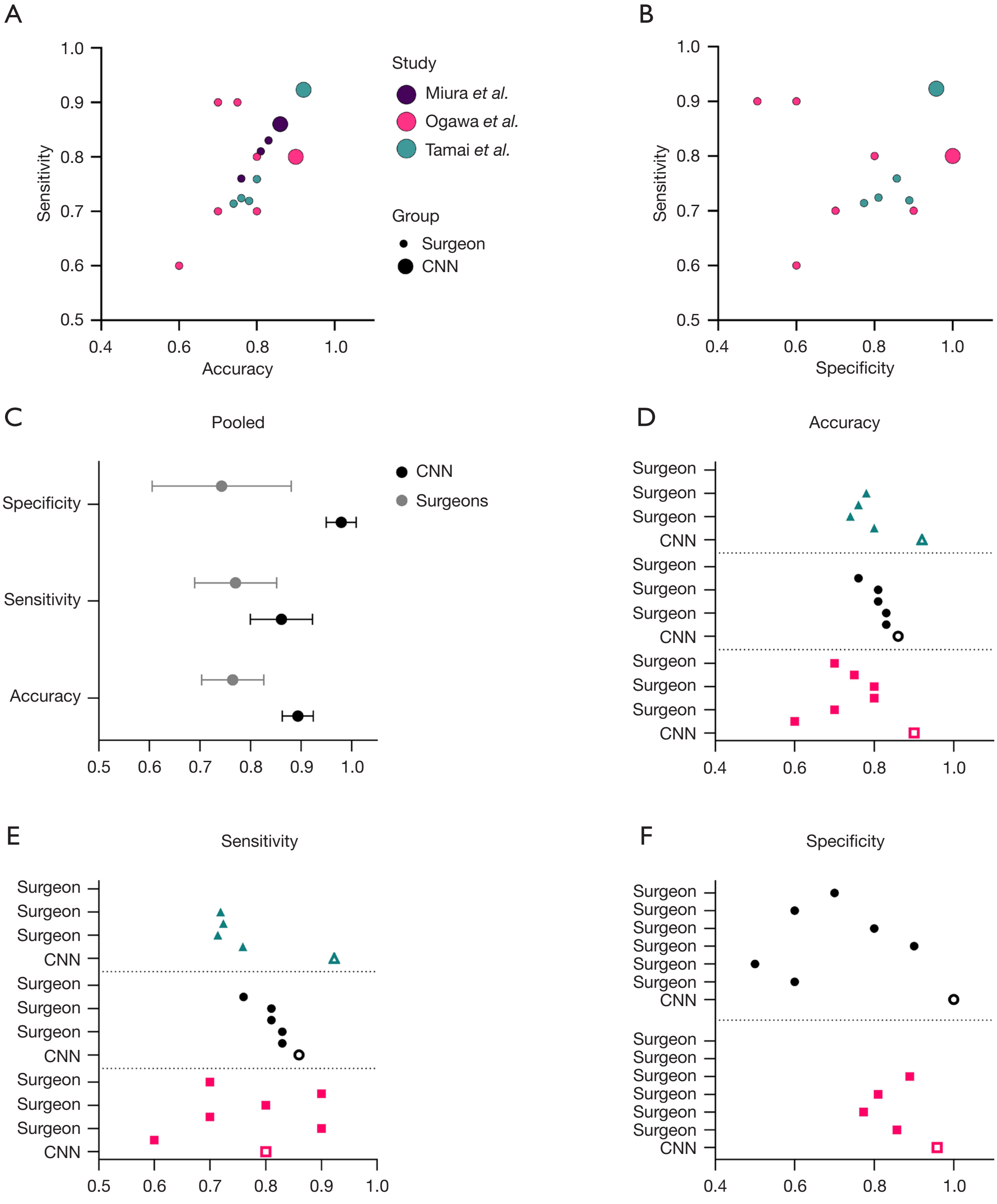


Figure 5 Presentation of DLM overall performance compared to the comparator group. (A) A bubble plot with sensitivity on the y-axis and accuracy on the x-axis. The three studies are color-coded (black = Miura *et al.*, red = Ogawa *et al.*, green = Tamai *et al.*). The performance of the respective DLM is represented by larger objects, while the human comparator groups (surgeons) are depicted as smaller objects. (B) A bubble plot with sensitivity on the y-axis and specificity on the x-axis. (C) The mean pooled sensitivity, specificity, and accuracy of the DLM (black color) and surgeons (grey color) are shown, with standard deviation as error bars. (D) The reported accuracy values for the surgeons' control group (filled objects) versus the DLM (empty objects), with color coding for the studies. (E) The reported sensitivity values for the surgeons' control group (filled objects) versus the DLM (empty objects), with color coding for the studies. (F) The reported specificity values for the surgeons' control group (filled objects) versus the DLM (empty objects), with color coding for the studies. DLM, deep learning model; CNN, convolutional neural network.

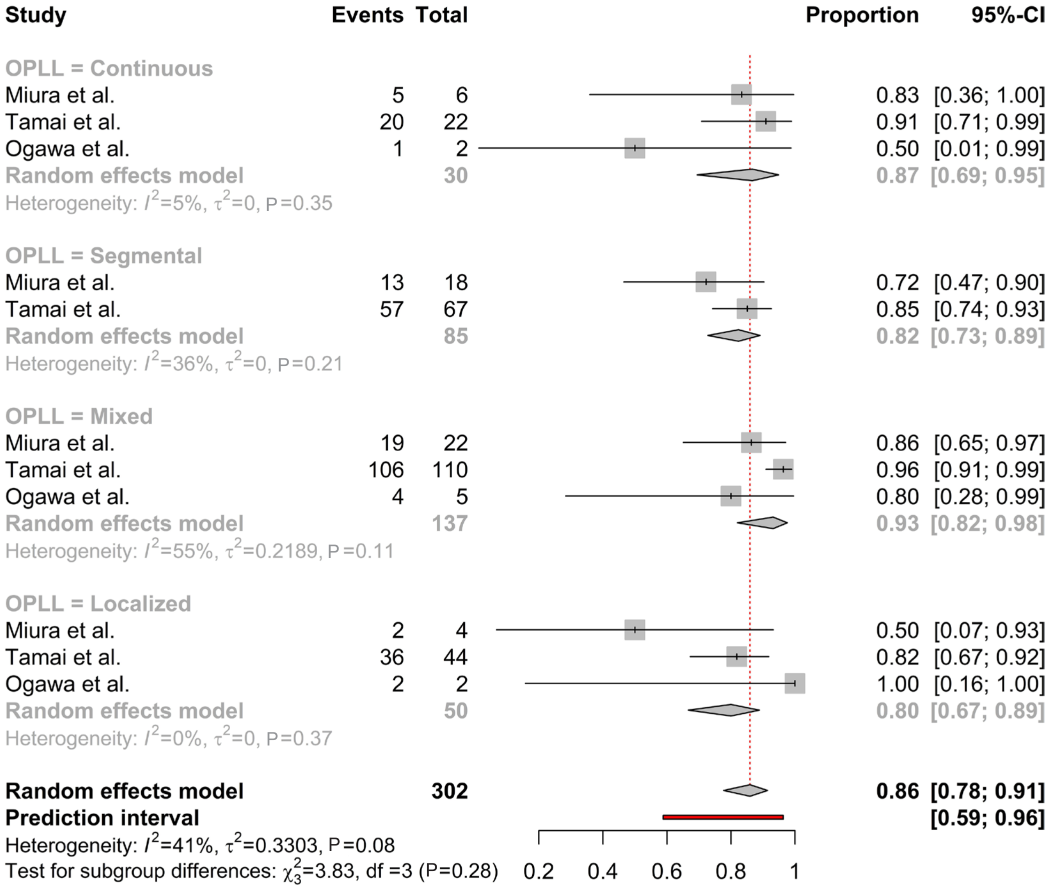


Figure 6 Four forest plots depict the sensitivity of CNN models concerning the OPLL subgroup. Each forest plot features the analyzed OPLL subgroup (continuous, segmental, mixed, and localized) presented in gray, followed by studies evaluating sensitivity values. The size of the gray square in the "Proportion" visually corresponds to the study sample size, and the straight line represents the confidence interval. The diamond at the bottom reflects the overall pooled proportion. Heterogeneity is denoted by the chi-squared statistic (I^2) with associated r^2 and P value. The 95% confidence intervals (CI) are enclosed in square brackets ([ ]). A P value <0.05 is considered significant. Furthermore, each study provides information on the study author with the publication date ("Study"), the total sample size for each study in the respective OPLL subgroup ("Total"), the number of correctly identified cases ("Events") per OPLL subgroup, and the corresponding proportion, e.g., sensitivity ("Proportion"), the test for the significance of the overall effect size as t^2 and P value, and the percentage weighting of each study are presented. CNN, convolutional neural network; OPLL, ossification of the posterior longitudinal ligament; CI, confidence interval.

were evident in the upper to middle cervical spine (defined as C2–C4; sensitivity =0.92) compared to the middle (defined as C5 and C6; sensitivity =0.87) and middle to lower (defined as below C6 level; sensitivity =0.88) (21).

Discussion

Summary of findings

This systematic review evaluates the utilization of DLMs in diagnosing (5,16,18-21) and outcome prediction (17) of OPLL. The results suggest that DLMs show potential for enhancing diagnostic accuracy, particularly in less irradiating imaging modalities such as X-rays, which could reduce the need for CT imaging and its associated radiation exposure. The high levels of accuracy, sensitivity, and specificity observed in both internal and external validations, mainly when applied to plain radiographs of the cervical spine (16,18-21), suggest that DLMs can serve as an effective screening tool for OPLL. Moreover, the robust application of DLMs in recognizing cervical OPLL

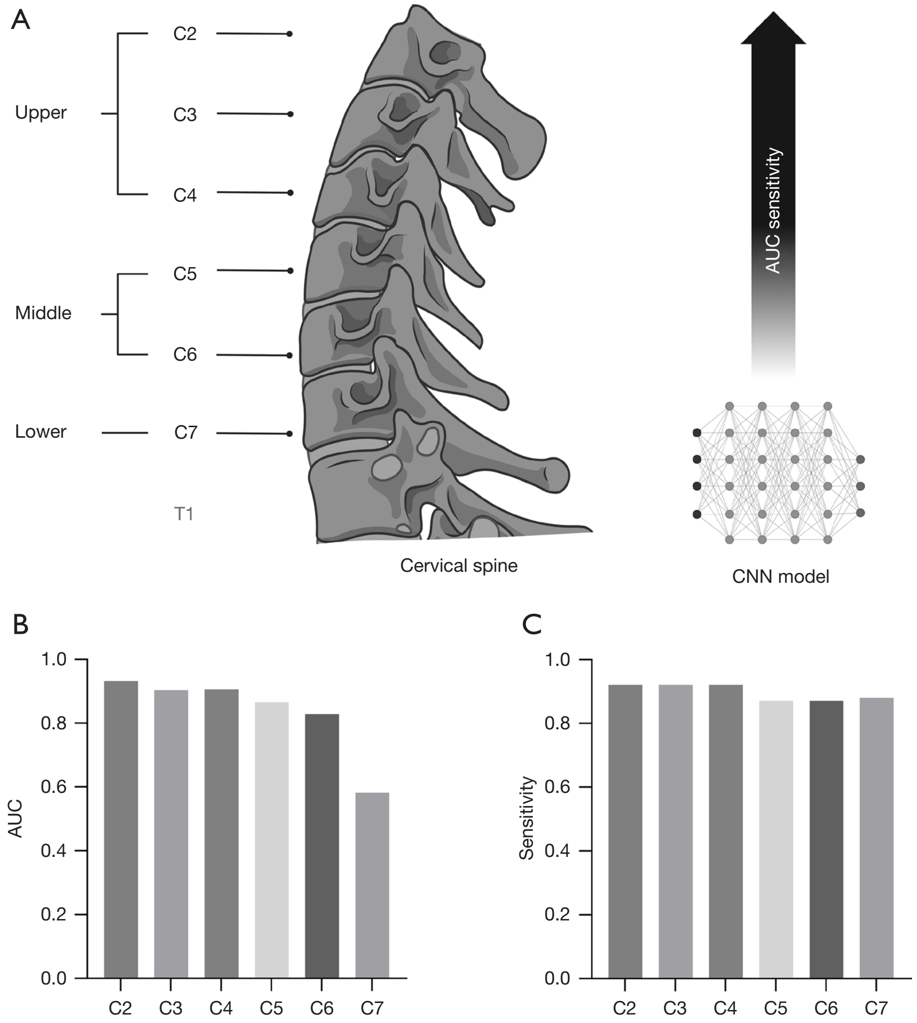


Figure 7 The performance of the DLM depends on the anatomical localization of OPLL. (A) Illustration of the authors' definitions. The cervical vertebrae C2 to C4 correspond to the upper, C5 and C6 to the middle, and below C7 to the lower cervical spine (T1 = thoracic vertebra 1). The AUC and sensitivity of the CNN model increase with higher cervical spine levels. (B,C) Bar plots representing the AUC and sensitivity for each cervical vertebra (region) (created with BioRender.com). AUC, area under the curve; CNN, convolutional neural network; DLM, deep learning model; OPLL, ossification of the posterior longitudinal ligament.

extends beyond lateral plain radiography to include various imaging modalities, including MRI (5). Therefore, while CT remains the primary diagnostic method for conditions like OPLL, the application of deep learning in modalities such as MRI could further streamline diagnosis, offering the added patient benefit of minimizing radiation and reducing reliance on multiple imaging techniques. In addition, consistent outcomes across numerous studies underscore the superior performance of DLMs when compared to human observers (18,20,21), thereby not only enhancing human diagnostic capabilities but also holding promise for advancing risk assessment related to complications associated with cervical OPLL (17).

These results align with the broader trend in medical imaging, where deep learning techniques have shown success in various diagnostic tasks like ophthalmological, respiratory, and breast imaging (22). In neurosurgery, deep learning approaches have been employed especially for glioblastoma prognosis (23), origin identification of spinal metastasis (24), and meningioma classification (25), among others. The ability of DLM to automatically learn and extract complex patterns from images is particularly valuable in cases like OPLL where subtle features may be challenging to identify.

Second, the widespread occurrence of OPLL and the complexities associated with its identification in clinical settings where advanced imaging modalities such as CT or MRI are not readily accessible highlight the potential utility of DLM. The use of cervical radiographs as a screening tool, especially in non-traumatic settings where patients present with local signs or symptoms consistent with cervical root level distribution, offers a practical and potentially cost-effective approach (18,26). However, despite the cost-effectiveness of radiographs, the reliability of diagnostic accuracy, particularly for OPLL, remains suboptimal. Although CT screening has been shown to improve the accuracy of OPLL diagnosis (27,28), its application in large sample cohorts may not be feasible (18,27).

The diverse presentations of OPLL underscore the potential utility of DLMs in aiding diagnosis by identifying patterns that might be challenging to detect through traditional methods. The existence of various morphological subtypes may lead to differences in detection outcomes. The findings of this review demonstrated the best diagnostic results in the continuous and mixed subtypes (16,18,20,21). In Chae *et al.*'s study, both the DLM and the comparator group (radiologists) performed better in accuracy for continuous and mixed subtypes than for segmental and circumscribed types (16). This is consistent with the findings of other studies, which also encountered greater difficulties in segmental and circumscribed types (18,20,21).

Similarly, when utilizing lateral radiographs, diagnostic accuracies are higher for continuous and mixed types (85.7% and 91.7%) compared to the segmental- and circumscribed types (27.3% and 20%) (6). This may be attributed, on one hand, to their longer extension and greater thickness. It has been shown that the mean thickness for the segmental type is 4.3 mm, whereas for the continuous type, it is 9 mm (29). On the other hand, different results indicate that, especially in localized and segmental subtypes, obscurations in the form of osteophytes, facets, and pedicles often exist, influencing diagnostic accuracy (16). Despite the diagnostic challenges posed by the segmental and circumscribed types, the DLM exhibited proficiency in recognizing these subtypes up to approximately 80%. Thus, within the challenging diagnostic context of subtypes, DLMs showcase the potential to comprehensively cover a diverse spectrum of OPLL presentations effectively.

Beyond the pathomorphology of OPLL, the cervical level emerges as a significant factor. Although only a minority of studies (16,21) examined the anatomical height of OPLL, the consistent findings suggest that a higher localization in the cervical spine is linked to greater accuracy and sensitivity than in the lower cervical spine. Challenges may arise in assessing this aspect due to patient body stature variations, including short necks or elevated shoulders (30). Nonetheless, the DLM demonstrated consistent sensitivity in identifying OPLL across various levels, unlike the human comparator group, which tended to overlook OPLL instances in the lower cervical level. Due to the pre-segmentation approach, the DLM can accurately assess OPLL using smaller segments rather than relying on a global perspective. Consequently, the DLM remains unaffected by factors such as the shoulder line, which might pose challenges in evaluating OPLL for the human comparator group (21).

Complication and outcome prediction

In addition to employing DLM for diagnosis and diagnostic support there are other opportunities for their application. Alongside OPLL diagnosis a crucial consideration is the ability to assess complications. While anterior surgical approaches have demonstrated greater efficiency than posterior approaches, they are associated with higher complication rates (4). Therefore, conducting preoperative risk stratification based on various factors would be advantageous to enhance the personalized estimation of complications at the individual patient level.

To address this need, Ito *et al.* developed a DLM trained on clinical, surgical, and radiological factors, which underwent prospective multicentric evaluation. The results revealed that the DLM achieved an accuracy of 74.6% in predicting the overall occurrence of complications and 91.7% accuracy in specifically predicting neurological complications (17).

In contrast, machine learning (ML) models have been applied to predict specific postoperative outcomes in OPLL. Kim *et al.* propose a machine-learning model for predicting postoperative C5 palsy in OPLL (31). The ML algorithm demonstrated the ability to predict C5 palsy, and certain risk variables such as age, anatomical localization height of OPLL, and postoperative shoulder pain showed stronger associations with C5 palsy.

Finally, Maki *et al.* proposed a prognostic ML model for surgical outcomes in OPLL patients (32). The results underscored the feasibility of the ML model, as indicated by the AUC, and its accuracy in predicting minimal clinically important differences in the Japanese Orthopedic Association (JOA) score.

Limitations

The limitations of this study should be considered when interpreting the results. Most studies included in the analysis were retrospective (5,16,18-21), introducing inherent biases. While retrospective studies provide valuable insights, transitioning to prospective evaluations is recommended to enhance the algorithm's reliability. Applying the algorithm to prospective data would offer more accurate, real-time information, reducing recall bias and strengthening its generalizability.

The funnel plot presented in *Figure 3G* should be interpreted with caution. While we chose to include the funnel plot for reporting completeness, it is essential to recognize its limitations. Funnel plots are conventionally more robust and precise tools for analysis when applied to a larger pool of studies, typically exceeding ten. Notably, the study by Ito *et al.* was excluded from the funnel plot due to missing sensitivity data, which is essential for calculating the number of events—a key input for the meta-analysis and the subsequent generation of the funnel plot. The exclusion of this study highlights a limitation of the current analysis and should be considered when interpreting the funnel plot results.

In this case, the observed funnel plot asymmetry is notable, with most studies falling outside the 95% confidence intervals and only two within them. While this could suggest potential publication bias, it is equally plausible that the asymmetry arises from the limited number of studies included, the high degree of heterogeneity (I^2=94.7%), or differences in methodologies, sample sizes, and study designs across the included studies. To further investigate this, Egger's test was conducted and indicated significant asymmetry, reinforcing the possibility of publication bias. However, due to the small number of studies, the statistical power of this test is limited, and the results should be interpreted with caution. To adjust for potential publication bias, the Trim-and-Fill method was employed, which added one hypothetical study to correct for the observed asymmetry. While this adjustment provided a more balanced funnel plot and a revised pooled estimate, it is important to acknowledge that the added study is hypothetical, and the adjustment itself relies on assumptions that may not fully reflect the underlying data. Despite these limitations, the funnel plot was included to transparently present all potential indicators of bias.

Our decision to compute a summary score for assessing the overall study quality using the QUADAS-2 tool reflects a common practice in systematic reviews to provide a concise evaluation. However, it is important to acknowledge that this choice presents particular challenges, primarily due to the inherent subjectivity in summarizing study quality through a scoring system (33,34). Assigning scores to various aspects of the study methodology involves interpretation, and different researchers may attribute different weights to the same criteria. This subjectivity in scoring underscores a broader issue in the existing landscape of risk of bias tools, particularly in their application to studies that explicitly deal with artificial intelligence. Therefore, available tools may not be optimally tailored to capture the nuances and intricacies of methodologies specific to artificial intelligence (AI)-based studies.

Another limitation may be the variability in the levels of experience and specialization among the surgeons involved, which could introduce potential bias in the comparison with the CNNs. We acknowledge that the inclusion of surgeons with varying levels of experience, particularly those without a spinal specialization, may introduce some variability in performance. However, we believe this reflects the real-world clinical scenario, where surgeons with differing backgrounds are involved in diagnosing and treating spinal conditions. The observed differences in performance between the surgeons and the CNNs may therefore be partly attributed to these variations in expertise, which is important to consider when interpreting the strength of our conclusions.

The variations in reporting methodology and results among the studies are notably pronounced. Comprehensive details on model construction, training, and validation, but also outcome data, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN) data, contribute to the transparency and reproducibility of the research. Across the individual studies, we observed differences in the reporting, potentially introducing a degree of distortion to the summarized results. A standardized and detailed reporting approach would enhance the clarity and comparability of studies, facilitating a more accurate understanding of the nuances in model development and validation processes.

Current challenges and future directions

The current challenges in implementing DLM for evaluating OPLL involve several key aspects. Overall, the reliance on human-generated labeled training data to create the algorithm underscores the inherent limitation that AI cannot surpass human abilities, as the ground truth for training is set by humans (21).

The amount of training data poses a challenge, as algorithms are being outperformed when the data are limited. In the study conducted by Chae *et al.*, the DLM demonstrated superior performance compared to the radiologist in the segmental type (with a training-validation set of 122 and a test set of 62). Conversely, radiologists exhibited higher performance in the continuous type (with a training-validation set of 38 and a test set of 16) (16).

Further, the diagnostic performance of the DLM alone may not be statistically significantly better than average observers, as in the study from Chae *et al.* (16). Issues may arise in segmenting lesions, especially when the boundaries of small lesions are not clearly depicted on plain radiographs, leading to inaccuracies in preparing training data and subsequently affecting the overall performance of the DLM (16). Additionally, relying solely on the validation of model performance using data from the same institution as the development dataset raises concerns about potential overfitting (16).

Other challenges include the trimming of disease regions in original images during preprocessing (20), and when high computational requirements are not met by healthcare facility hardware infrastructure (5). False-positive readings may occur, especially in cases of ossified bulging discs (5), and the accuracy of the AI model can be influenced by the institution where radiography was obtained, considering factors such as concentration of radiography, incidence angle, and patient positioning (21).

Therefore, it is essential to interpret these findings cautiously. While this review indicates promising results for DLM, further research and validation in diverse populations and clinical settings are necessary. Moreover, the comparison between CNNs and human observers in this review emphasizes the potential of these technologies but does not diminish the importance of human expertise in clinical decision-making (35).

The high diagnostic performance of DLM in OPLL detection suggests their potential integration into clinical practice as decision-support tools. Radiologists and clinicians could benefit from these tools to enhance accuracy, expedite diagnosis, and minimize radiation exposure (21). Chae *et al.* (16) utilized the DLM in a second session, during which each observer evaluated the images with reference to the results provided by the CNN, demonstrating an additive effect in identifying OPLL on plain radiographs, regardless of the surgeons' experience levels. In the future, DLMs are likely to be implemented primarily as enhancement tools rather than standalone systems, supporting clinicians by improving diagnostic accuracy and efficiency. These models could initially focus on complementing human expertise, serving as decision-support tools that aid in identifying subtle findings and reducing diagnostic variability.

Using advanced technologies like DLMs raises questions about regulatory and ethical considerations (36). Policies need to be developed to govern the integration of artificial intelligence in medical diagnostics, ensuring patient safety (37), data privacy, and adherence to ethical standards (38). In this regard, key areas for policy development include establishing rigorous validation standards to ensure DLMs meet clinical safety requirements, safeguarding patient data in compliance with privacy regulations, addressing potential biases in AI models to ensure fairness and equity, and ensuring that clinicians maintain oversight of AI-driven decisions. Deploying DLMs in research emphasizes the importance of close collaboration between AI developers, clinicians, and regulatory bodies to create robust and effective guidelines for the safe clinical use of AI technologies.

Future research should address the limitations identified, including conducting larger-scale studies with diverse populations, standardizing reporting metrics, and exploring the long-term impact of DLM on patient outcomes. Comparative studies evaluating the cost-effectiveness of DLM-assisted diagnosis compared to traditional methods could also provide valuable insights. Continuous collaboration between medical professionals, researchers, and policymakers will be crucial to navigate the evolving landscape of deep learning applications in medical imaging.

Conclusions

The utilization of DLMs in the identification of cervical OPLL demonstrates robustness, extending its efficacy beyond lateral plain radiography. The consistent findings across numerous studies highlight the DLM's superior overall accuracy, sensitivity, and specificity when compared to human observers. This enhances human diagnostic capabilities and holds promise for risk assessment related to complications. Within the spectrum of OPLL subtypes, the mixed and continuous type is particularly well-identifiable.

Optimal accuracy and sensitivity of OPLL registration are achieved at the upper cervical level. In the future, DLM may serve as a valuable tool in advancing the diagnosis and risk evaluation of cervical OPLL.

Acknowledgments

This work from the author Felix Corr was completed in partial fulfilment of a Master of Science in Neurosurgery from Learna, validated by the University of Buckingham.

Footnote

Reporting Checklist: The authors have completed the PRISMA-DTA reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1485/rc

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1485/coif). F.C. and D.G. were research fellows at Isarklinikum Munich during the conduct of this study. Their positions were purely academic and were not associated with any financial compensation or other benefits. The other author has no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Le HV, Wick JB, Van BW, Klineberg EO. Ossification of the Posterior Longitudinal Ligament: Pathophysiology, Diagnosis, and Management. J Am Acad Orthop Surg 2022;30:820-30.
2. Chikuda H, Seichi A, Takeshita K, Matsunaga S, Watanabe M, Nakagawa Y, et al. Acute cervical spinal cord injury complicated by preexisting ossification of the posterior longitudinal ligament: a multicenter study. Spine (Phila Pa 1976) 2011;36:1453-8.
3. Moghaddamjou A, Fehlings MG. An Age-old Debate: Anterior Versus Posterior Surgery for Ossification of the Posterior Longitudinal Ligament. Neurospine 2019;16:544-7.
4. Kim DH, Lee CH, Ko YS, Yang SH, Kim CH, Park SB, Chung CK. The Clinical Implications and Complications of Anterior Versus Posterior Surgery for Multilevel Cervical Ossification of the Posterior Longitudinal Ligament; An Updated Systematic Review and Meta-Analysis. Neurospine 2019;16:530-41.
5. Shemesh S, Kimchi G, Yaniv G, Harel R. MRI-based detection of cervical ossification of the posterior longitudinal ligament using a novel automated machine learning diagnostic tool. Neurosurg Focus 2023;54:E11.
6. Kang MS, Lee JW, Zhang HY, Cho YE, Park YM. Diagnosis of Cervical OPLL in Lateral Radiograph and MRI: Is it Reliable? Korean J Spine 2012;9:205-8.
7. Singh NA, Shetty AP, Jakkepally S, Kumarasamy D, Kanna RM, Rajasekaran S. Ossification of Posterior Longitudinal Ligament in Cervical Spine and Its Association With Ossified Lesions in the Whole Spine: A Cross-Sectional Study of 2500 CT Scans. Global Spine J 2023;13:122-32.
8. Yoshimura N, Nagata K, Muraki S, Oka H, Yoshida M, Enyo Y, Kagotani R, Hashizume H, Yamada H, Ishimoto Y, Teraguchi M, Tanaka S, Kawaguchi H, Toyama Y, Nakamura K, Akune T. Prevalence and progression of radiographic ossification of the posterior longitudinal ligament and associated factors in the Japanese population: a 3-year follow-up of the ROAD study. Osteoporos Int 2014;25:1089-98.
9. Fujimori T, Le H, Hu SS, Chin C, Pekmezci M, Schairer W, Tay BK, Hamasaki T, Yoshikawa H, Iwasaki M. Ossification of the posterior longitudinal ligament of the cervical spine in 3161 patients: a CT-based study. Spine (Phila Pa 1976) 2015;40:E394-403.
10. Fujimori T, Watabe T, Iwamoto Y, Hamada S, Iwasaki M, Oda T. Prevalence, Concomitance, and Distribution of Ossification of the Spinal Ligaments: Results of Whole Spine CT Scans in 1500 Japanese Patients. Spine (Phila Pa 1976) 2016;41:1668-76.
11. Cao CF, Ma KL, Shan H, Liu TF, Zhao SQ, Wan Y, Jun-Zhang, Wang HQ. CT Scans and Cancer Risks: A

Systematic Review and Dose-response Meta-analysis. BMC Cancer 2022;22:1238.

12. Montagnon E, Cerny M, Cadrin-Chênevert A, Hamilton V, Derennes T, Ilinca A, Vandenbroucke-Menu F, Turcotte S, Kadoury S, Tang A. Deep learning workflow in radiology: a primer. Insights Imaging 2020;11:22.
13. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8:53.
14. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611-29.
15. Gheibi S, Mahmoodzadeh A, Kashfi K, Jeddi S, Ghasemi A. Data Extraction from Graphs Using Adobe Photoshop: Applications for Meta-Analyses. Int J Endocrinol Metab 2019;17:e95216.
16. Chae HD, Hong SH, Yeoh HJ, Kang YR, Lee SM, Kim M, Koh SY, Lee Y, Park MS, Choi JY, Yoo HJ. Improved diagnostic performance of plain radiography for cervical ossification of the posterior longitudinal ligament using deep learning. PLoS One 2022;17:e0267643.
17. Ito S, Nakashima H, Yoshii T, Egawa S, Sakai K, Kusano K, et al. Deep learning-based prediction model for postoperative complications of cervical posterior longitudinal ligament ossification. Eur Spine J 2023;32:3797-806.
18. Miura M, Maki S, Miura K, Takahashi H, Miyagi M, Inoue G, Murata K, Konishi T, Furuya T, Koda M, Takaso M, Endo K, Ohtori S, Yamazaki M. Automated detection of cervical ossification of the posterior longitudinal ligament in plain lateral radiographs of the cervical spine using a convolutional neural network. Sci Rep 2021;11:12702.
19. Murata K, Endo K, Aihara T, Suzuki H, Sawaji Y, Matsuoka Y, Takamatsu T, Konishi T, Yamauchi H, Endo H, Yamamoto K. Use of residual neural network for the detection of ossification of the posterior longitudinal ligament on plain cervical radiography. Eur Spine J 2021;30:2185-90.
20. Ogawa T, Yoshii T, Oyama J, Sugimura N, Akada T, Sugino T, Hashimoto M, Morishita S, Takahashi T, Motoyoshi T, Oyaizu T, Yamada T, Onuma H, Hirai T, Inose H, Nakajima Y, Okawa A. Detecting ossification of the posterior longitudinal ligament on plain radiographs using a deep convolutional neural network: a pilot study. Spine J 2022;22:934-40.
21. Tamai K, Terai H, Hoshino M, Yabu A, Tabuchi H, Sasaki R, Nakamura H. A deep learning algorithm to identify cervical ossification of posterior longitudinal ligaments on radiography. Sci Rep 2022;12:2113.
22. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4:65.
23. Ghanem M, Ghaith AK, Zamanian C, Bon-Nieves A, Bhandarkar A, Bydon M, Quiñones-Hinojosa A. Deep Learning Approaches for Glioblastoma Prognosis in Resource-Limited Settings: A Study Using Basic Patient Demographic, Clinical, and Surgical Inputs. World Neurosurg 2023;175:e1089-109.
24. Duan S, Cao G, Hua Y, Hu J, Zheng Y, Wu F, Xu S, Rong T, Liu B. Identification of Origin for Spinal Metastases from MR Images: Comparison Between Radiomics and Deep Learning Methods. World Neurosurg 2023;175:e823-31.
25. Maniar KM, Lassarén P, Rana A, Yao Y, Tewarie IA, Gerstl JVE, Recio Blanco CM, Power LH, Mammi M, Mattie H, Smith TR, Mekary RA. Traditional Machine Learning Methods versus Deep Learning for Meningioma Classification, Grading, Outcome Prediction, and Segmentation: A Systematic Review and Meta-Analysis. World Neurosurg 2023;179:e119-34.
26. Johnson MJ, Lucas GL. Value of cervical spine radiographs as a screening tool. Clin Orthop Relat Res 1997;(340):102-8.
27. Sasaki E, Ono A, Yokoyama T, Wada K, Tanaka T, Kumagai G, Iwasaki H, Takahashi I, Umeda T, Nakaji S, Ishibashi Y. Prevalence and symptom of ossification of posterior longitudinal ligaments in the Japanese general population. J Orthop Sci 2014;19:405-11.
28. Chang H, Kong CG, Won HY, Kim JH, Park JB. Inter- and intra-observer variability of a cervical OPLL classification using reconstructed CT images. Clin Orthop Surg 2010;2:8-12.
29. Otake S, Matsuo M, Nishizawa S, Sano A, Kuroda Y. Ossification of the posterior longitudinal ligament: MR evaluation. AJNR Am J Neuroradiol 1992;13:1059-67; discussion 1068-70.
30. Abbasi A, Malhotra G. The "swimmer's view" as alternative when lateral view is inadequate during interlaminar cervical epidural steroid injections. Pain Med 2010;11:709-12.
31. Kim SH, Lee SH, Shin DA. Could Machine Learning Better Predict Postoperative C5 Palsy of Cervical

Ossification of the Posterior Longitudinal Ligament? Clin Spine Surg 2022;35:E419-25.

32. Maki S, Furuya T, Yoshii T, Egawa S, Sakai K, Kusano K, et al. Machine Learning Approach in Predicting Clinically Significant Improvements After Surgery in Patients with Cervical Ossification of the Posterior Longitudinal Ligament. Spine (Phila Pa 1976) 2021;46:1683-9.
33. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054-60.
34. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19.
35. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. 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:e271-97.
36. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021;22:122.
37. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform 2020;8:e18599.
38. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Comput Biol Med 2023;158:106848.

Cite this article as: Corr F, Grimm D, Leach P. Deep learning for identifying cervical ossification of the posterior longitudinal ligament: a systematic review and meta-analysis. Quant Imaging Med Surg 2025;15(3):1719-1740. doi: 10.21037/qims-24-1485