EFORT open reviews

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

GENERAL ORTHOPAEDICS

Machine and deep learning models for ligament injury recognition: a systematic review and meta-analysis of imaging and novel diagnostic techniques

Guillermo Droppelmann[1], Emilia Varas[2], Joaquín Villagrán[2], Carlos Jorquera[3] and Felipe Feijoo[4]

[1]Clínica MEDS, Santiago, RM, Chile
[2]Facultad de Medicina, Universidad de los Andes, Santiago, RM, Chile
[3]Facultad de Ciencias, Escuela de Nutrición y Dietética, Universidad Mayor, Santiago, RM, Chile
[4]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Correspondence should be addressed to F Feijoo: felipe.feijoo@pucv.cl

- *Purpose:* Diagnosing ligament injuries remains a challenge for musculoskeletal clinicians due to the lack of standardized classification, evaluation, and management protocols. Machine learning (ML) and deep learning (DL) models offer potential to improve diagnostic accuracy. This study aimed to evaluate the diagnostic performance of various ML and DL models in identifying ligament injuries across different medical imaging modalities.
- *Methods:* A meta-analysis was conducted following the PRISMA 2020 checklist. Searches were performed in PubMed, SCOPUS, Web of Science, and the Cochrane Library. Study quality was assessed using the QUADAS-2 tool and Robvis software. Diagnostic performance measures – true positive, true negative, false positive, and false negative – were analyzed. A random-effects model was applied, and heterogeneity and subgroup analyses were conducted. Statistical and graphical analyses were performed using R. The study was registered in PROSPERO (CRD42025646317).
- *Results:* Fifty-nine ML and DL algorithms from 23 studies were analyzed. Pooled sensitivity and specificity were 0.890 (95% CI: 0.829–0.938) and 0.926 (95% CI: 0.820–0.959), respectively. Pooled estimates for PLR, NLR, lnDOR, and AUC were 1,644.37 (95% CI: 73.56–3,215.18), 0.179 (95% CI: 0.095–0.263), 4.130 (95% CI: 3.570–4.700), and 95%, respectively, with $P < 0.001$.
- *Conclusion:* ML and DL models demonstrate high diagnostic accuracy in detecting ligament injuries. Their strong performance supports ongoing integration into clinical practice, offering valuable support for musculoskeletal specialists in image interpretation and diagnosis.

Keywords: artificial intelligence; deep learning; diagnostic; ligament; machine learning

Introduction

Modern society has witnessed the unprecedented rise of artificial intelligence (AI), rapidly transforming numerous industries, including transportation, commerce, telecommunications, agriculture, and forest (1). This technological revolution has profoundly impacted medicine and healthcare. Various medical specialties have benefited from AI-driven tools and innovations, enhancing multiple processes and ultimately improving patient outcomes and quality of life (2).

Published by Bioscientifica Ltd.
www.efortopenreviews.org

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

The field of diagnostics has made significant advancements, emphasizing its role in enhancing precision, speed, and decision-making support for healthcare professionals. Both machine learning (ML) models and deep learning (DL) algorithms have contributed to processing large volumes of data (3). Moreover, their remarkable computational power enables the recognition of patterns in medical images that often go undetected by the human eye, sometimes exceeding the precision of specialists (4). Integrating these tools into hospital settings facilitates early detection of disease, allocation of resources, reduced waiting times, and greater diagnostic accuracy. This transformation transforms traditional medicine, which is based on standardized treatments, to precision medicine tailored to the individual needs of the patient (5).

Among medical specialties, musculoskeletal radiology has been one of the most significantly affected by the development and applications of AI. This is due to its strong connection to diagnostic advancements and its dependence on highly specialized technology (6). In addition, the integration of computer vision elements that enhance human–machine interaction has led to rapid adoption by professionals, administrators, and system users. In particular, the implementation of complex computational models capable of analyzing images with high speed and precision – particularly when the region of interest is accurately delineated – has optimized the diagnostic process by reducing interobserver variability and improving reproducibility and reliability in medical image interpretation (7).

In recent years, the use of AI-based strategies for diagnostic support has increased significantly among musculoskeletal radiologists, driven by the advantages mentioned above. This has led to the application of various models and algorithms aimed at optimizing diagnostic performance in medical image analysis (8). These strategies depend on several key factors, including data availability, computational capabilities, hardware requirements, interpretability and explainability, the required level of precision, and, most importantly, the clinical context in which they are applied. As a result, no single method or model is universally superior; instead, the choice must be carefully tailored to these multiple considerations (9).

One possible approach to implementation is to select protocols and methods based on the required level of complexity (10). Simple but effective strategies include deterministic approaches, such as thresholding, logical rules, or statistical metrics for image evaluation, which do not require data training but rely on proper calibration (11). Another widely used approach involves classic supervised and unsupervised learning methods that do not employ deep neural networks. These models can learn patterns from correctly labeled data without explicit human intervention, relying instead on system-fed input (12). A more advanced strategy is to use hybrid methods

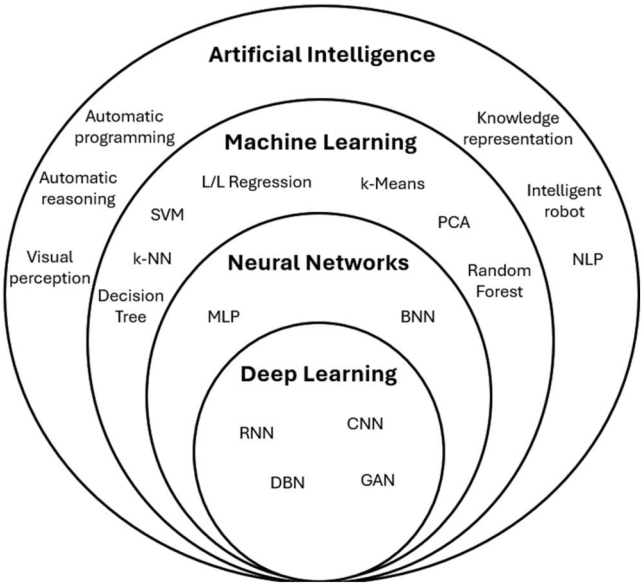

Figure 1

Venn diagram of AI structure.

that combine multiple techniques to improve diagnostic accuracy. These models integrate different approaches to ML and neural networks, leveraging the strengths of each method (13). Finally, the most complex implementation involves convolutional neural networks (CNN) and computer vision models for image processing, which automatically learn by extracting features from individual pixels (14). Figure 1 illustrates the hierarchical relationship of AI.

Ligament injuries remain a key concern in musculoskeletal pathologies due to their high prevalence, disease burden, and significant impact on both general health and quality of life in the general population and athletes (15). In most cases, these injuries result from an initial traumatic event, which, depending on severity, can range from a mild strain to a complete rupture of the affected ligament (16).

The classification of ligament injuries is highly diverse, as it depends on the specific structure being assessed. For example, ultrasound has shown high diagnostic accuracy for ankle sprains (17). In contrast, knee cruciate ligament injuries are easily detected with magnetic resonance imaging (MRI) (18). However, diagnosing ligament injuries in the wrist, hand, and fingers remains controversial (19). Various technological strategies have been developed as tools for detecting the risk of injury in musculoskeletal pathologies. Examples include sensor-based techniques for analyzing movement in sports injuries and portable monitoring systems (20, 21, 22).

Based on our experience, healthcare professionals use different classification systems for these injuries. Some follow a binary approach, distinguishing only between the

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

presence or absence of injury. Others use an ordinal system, categorizing injuries as mild, moderate, or severe based on the degree of ligament distension. Alternatively, a continuous system quantifies damage as a percentage or assesses clinical parameters such as structural alteration and joint stability deterioration.

Regardless of the classification system used, an initial ligament injury often leads to progressive failure over time, increasing the likelihood of recurrent joint instability (23). This condition, in turn, increases the risk of ongoing structural deterioration within the musculoskeletal system, promoting the development of progressive joint damage and, ultimately, secondary osteoarthritis (24). Traditionally, the onset of these injuries has been primarily attributed to trauma. However, recent studies suggest that alterations in specific brain structures involved in motor control and anticipation of biomechanical responses can influence both injury susceptibility and the patient's ability to recover (25). Therefore, early detection and precise diagnostic evaluation are essential for optimizing therapeutic strategies, preventing further damage to affected structures, and improving long-term clinical outcomes.

Previous studies have demonstrated with high precision the effectiveness of these diagnostic strategies in musculoskeletal pathologies affecting the upper extremity (26), and, more recently, in tendinopathies using various DL models and imaging modalities which have shown a high level of precision in detecting abnormalities (27). However, to the best of the authors' knowledge, no studies have yet comprehensively analyzed the range of algorithms available in the scientific literature for the detection and classification of ligament injuries. This systematic review and meta-analysis aims to evaluate the diagnostic performance of multiple ML and DL models in recognizing ligament injuries in different medical imaging modalities.

Methods

Reporting

This meta-analysis was carried out according to the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The 27-item checklist, which ensures comprehensive reporting in the Introduction, Methods, Results, and Discussion sections of a systematic review, was fully verified. This checklist is available at www.prisma-statement.org. Furthermore, the authors voluntarily registered the review on the PROSPERO platform: CRD42025646317.

Research question

This study aimed to evaluate the diagnostic performance of various ML and DL models in recognizing ligament injuries in different medical imaging modalities and novel diagnostic techniques. The research question was structured using the PICOT framework (participants, intervention, comparison, outcome, and time), which is detailed in Table 1.

Table 1 Summary of study components.

Acronym	Component	Explanation
(P)	Population	Patients diagnosed with ligament injury who have any type of diagnostic imaging or novel diagnostic method
(I)	Intervention	Any type of ML and DL model used for diagnostic purposes
(C)	Comparison	Conventional diagnostic method
(O)	Outcome	Evaluate the diagnostic performance of various ML and DL models
(T)	Type of study	Diagnostic study

Search strategy and data sources

The lead author, GD, a specialist in musculoskeletal injuries, selected keywords related to ligament injuries. The validation of diagnostic terminology was performed by an external collaborator, a radiologist with subspecialty expertise in musculoskeletal disorders. Keywords related to ML or DL were identified and validated by coauthor FF, PhD in engineering. Each member of the professional team has over 15 years of experience.

The PubMed search engine confirmed that all selected terms corresponded to medical subject headings (MeSH). The final terms included in this review were: ligaments, ligament, diagnosis, diagnostic imaging, AI, neuronal network (NN), convolutional neural network (CNN), artificial neural network (ANN), new diagnostic method, and novel diagnostic method. Subsequently, two coauthors, EV and JV, conducted a systematic search in selected databases, including: MEDLINE/PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), WOS/Web of Science (https://clarivate.com/), SCOPUS (https://www.scopus.com/home.uri), the Cochrane Library (https://www.cochranelibrary.com/).

Any discrepancies between the reviewers were resolved by a third independent reviewer, CR. The review covered 10 years, from September 2014 to September 2024. Finally, a comprehensive data matrix was generated by simultaneously combining all possible variations of the selected terms. Full access was obtained to all articles included in this study.

Selection criteria

Inclusion criteria

i) complete and published original scientific articles. ii) Studies focusing on ligament injuries in which diagnosis was supported by ML or DL tools. iii) Original scientific

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

articles that included any type of radiological imaging or novel diagnostic method, regardless of the location of the injury. iv) Original scientific articles that incorporate at least one ML or DL model as a diagnostic method. v) Original scientific articles explicitly reporting true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for the calculation of sensitivity and specificity. vi) Original scientific articles explicitly reporting sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). vii) Original scientific articles published in English, Spanish, or Portuguese. viii) Original scientific articles published within the last 10 years, up to September 2024. ix) Human ligaments.

Exclusion criteria

i) scientific articles classified as reviews, letters, conference reports, studies using cadaveric samples, or technical descriptions. ii) Studies focused on medical or technological devices, sensors, virtual reality, or any tangible (hardware) or intangible (software) object that does not incorporate ML or DL models.

Data extraction

A preliminary screening was performed by reviewing titles and abstracts. Original scientific articles in full text that met the predefined selection criteria were then selected, while duplicate manuscripts were removed. A data matrix was created using Microsoft Excel, including the following variables: authors, year of publication, country of origin, number of images used in the validation process, type of imaging modality, clinical diagnostic condition, and computational model used.

The TP, FN, FP, and TN reported for each model were identified. As is customary in diagnostic performance analysis, if an article included multiple models, all reported results were considered, with each model specifically identified by name. In cases in which these metrics were not provided, data on sensitivity, specificity, PPV, and NPV were extracted, along with the respective sample sizes in the validation set. Using all available information, the diagnostic metrics for sensitivity, specificity, precision, and F1 score were calculated. Finally, data extraction was performed manually by two coauthors, EV and JV, under the supervision of a third author, CJ. Any discrepancies were resolved by the lead author.

Ethical approval

Although specific approval from an ethics committee is not required for conducting a systematic review and meta-analysis, this study included only research that adhered to the universal principles outlined in the Declaration of Helsinki. Furthermore, all selected studies had received approval from a scientific ethics committee and ensured the proper execution of informed consent procedures.

Risk of bias (quality) assessment

The QUADAS-2 (quality assessment of diagnostic accuracy studies) guidelines were used to evaluate the quality of the selected articles and identify potential biases that could affect the interpretability of the results. The studies were classified into three levels: low, some concerns, or high (28).

The QUADAS-2 tool consists of four key domains that discuss patient selection, index test, reference standard, patient flow through the study, and timing of index tests and reference standard (28). The coauthors jointly reviewed the patient selection methods and assessed whether the index test was clearly described in terms of application and interpretation. The reference standard was analyzed to ensure that its use was explicitly detailed. Finally, it was verified whether all patients underwent both tests and whether any time interval or intervention could have influenced the results.

The graphs were generated using the Robvis application, developed with the Robvis package in R. The structure of the QUADAS-2 example dataset in Excel was used, and the analysis was performed step by step according to the platform guidelines. This tool is freely available at: https://mcguinlu.shinyapps.io/robvis/.

Statistical analysis: univariate and bivariate methods

To assess the diagnostic performance of the models, TP, FN, FP, and TN were calculated when not explicitly reported, using sensitivity, specificity, PPV, and NPV. Furthermore, the positive likelihood ratio (PLR) and the negative likelihood ratio (NLR) were estimated with their respective 95% confidence intervals (95% CI), considering the relationship between the number of events and the sample size for each model.

To enhance the stability of the results, a logistic transformation followed by an inverse transformation was applied using the Clopper–Pearson method. For better comparison and synthesis of results, the diagnostic odds ratio (DOR) and its logarithmic version (lnDOR) were calculated. Finally, data were visually represented using forest plots. To address the secondary research question, a bivariate analysis was performed, categorizing the model subgroups into three categories according to their architectural complexity.

- **Group 0, (g = 0):** low- and moderate-complexity models.
- **Group 1, (g = 1):** high-complexity models.

For each subgroup, DOR was calculated and visually represented using a forest plot. In addition, diagnostic accuracy metrics were estimated using the area under the curve (AUC), and a summary receiver operating

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

characteristic (SROC) curve was generated to summarize the overall performance of the models.

Heterogeneity analysis

A random effects model was applied due to the significant heterogeneity observed among the selected studies. Variability was estimated using the inverse variance method, assigning a specific weight to each study, and the tau (τ) value was calculated using the DerSimonian–Laird estimator.

The influence of heterogeneity on total variability, compared to random variability between studies, was assessed using the Higgins I^2 statistics, both for the overall data set and within subgroups. Heterogeneity was classified into four levels: minimum: 0–40%; moderate: 30–60%; high: 50–90%; extreme: 75–100%. To assess what proportion of total variability is due to differences between the samples analyzed, Cochran's *Q* test was applied.

Packages and reports

To perform diagnostic precision analyses, the following packages were used in the R statistical environment: ellipse, mada, meta, metafor, mvmeta, mvtnorm, and rmeta. A significance level of $P < 0.05$ was established, and the 95% CI were calculated. The results are reported in three decimal places. All statistical analyses and graphical representations were performed using R statistical software (version 4.1.3).

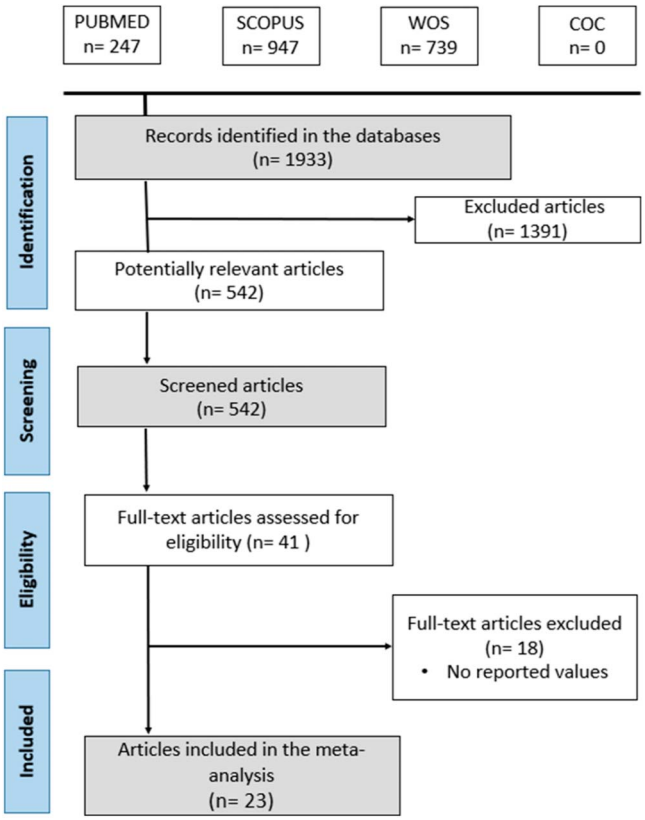

Figure 2

PRISMA 2020 flow diagram.

Results

Search results

The flowchart presented in Fig. 2 follows the PRISMA 2020 guidelines and describes all included studies. The initial search across selected bibliographic databases identified 1,933 scientific articles. After excluding more than 1,300 studies, 542 articles were considered potentially relevant. Through the application of screening and eligibility criteria, the selection was further refined to 23 articles, collectively reporting 59 ML and DL models. The final sample included the model reported with the highest diagnostic performance for incorporation into the meta-analysis.

Studies' features

Among the selected articles, a clear trend emerges: as years progress, the number of published studies increases. This reflects the growing interest of healthcare teams in implementing these strategies to improve the diagnostic accuracy of the musculoskeletal conditions being evaluated. When grouping the articles by continent of origin, Asia emerges as the leader with the highest number of publications, reflecting significant efforts in this region. North America and Europe follow, while South America has the lowest contribution. Among the identified diagnostic methods, medical imaging was the most represented, with a total of 19 articles, focusing mainly on MRI. In addition, four new diagnostic methods were identified. Regarding musculoskeletal conditions, there was a notable interest in studying knee joint disorders, with a particular emphasis on the anterior cruciate ligament. It should be noted that, although CNN was used in a slightly higher number of articles, the distribution of the other implemented models remains relatively homogeneous. For more details, see Table 2. The following section presents the diagnostic metrics for the selected studies. See Table 3.

Risk of bias

The 23 selected articles were rigorously analyzed using the QUADAS-2 tool (see Fig. 3). For further details, refer to Fig. 3. On average, seven studies exhibited a high risk of bias, with the dimension three being the most affected. In contrast, only two studies consistently demonstrated a low risk in all evaluated criteria. The remaining articles

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

Table 2 Summary of studies on ML and DL models for ligament injury diagnosis.

Study	Reference	Country	Diagnostic method	Condition	Model	*g*	TP	FP	TN	FN
1	Astolfi *et al.* (29)	Brazil	MRI	Ankle ligament injury	Random forest	0	578	150	566	102
2	Chang *et al.* (30)	USA	MRI	ACL tear	Resnet + U-net	1	53	0	56	7
3	Cheng *et al.* (31)	China	MRI	ACL tear	SVM model 2	0	75	3	17	13
4	Fang Liu *et al.* (32)	USA	MRI	ACL tear	MRnet	1	48	2	48	2
5	Germann *et al.* (33)	Switzerland	MRI	ACL tear	DCNN	1	223	23	255	11
6	Guha Paul *et al.* (34)	Bangladesh	CT	CSL	MobileNetV2	1	199	0	200	1
7	Ito *et al.* (35)	Japan	X-ray	TPLL	YOLO v4	1	121	46	125	4
8	Jo *et al.* (36)	Korea	MRI	TPLC	InResNetV2	1	41	3	47	9
9	Kanthavel *et al.* (37)	Saudi Arabia	X-ray	Osteoarthritis	DRRL	1	93	127	372	42
10	Kim *et al.* (38)	Korea	X-ray	CPLL	ADA	0	4	1	142	28
11	Liang *et al.* (39)	China	MRI	ACL tear	ResNet	1	179	26	329	96
12	Michael R *et al.* (40)	USA	MRI	ACL tear	FS	0	49	1	150	1
13	Minamoto *et al.* (41)	Japan	MRI	ACL tear	CNN	0	46	7	43	4
14	Mouchotte *et al.* (42)	France	GNRB-MRI	ACL tear	New method	1	78	0	5	5
15	Shemesh *et al.* (43)	Israel	MRI	CPLL	CNN	0	578	102	702	14
16	Tamai *et al.* (44)	Japan	X-ray	CPLL	E-NetB2	1	219	34	209	24
17	Tedesco *et al.* (45)	Ireland	Motion sensor	ACL tear	XGB	0	2,526	1,158	1,978	562
18	Wang *et al.* (46)	China	MRI	Knee imaging	PI protocol	0	14	0	31	1
19	Whiteside *et al.* (47)	USA	Tracking data	UCL	SVM	0	78	26	77	27
20	Yingkai *et al.* (48)	China	MRI	Meniscus injury	C-PCNN	1	548	62	654	132
21	Zhang *et al.* (49)	China	MRI	ACL tear	CNN, 3D DenseNet	1	41	2	37	1
22	Zhang *et al.* (50)	China	MRI	ACL tear	CNN/MGSA	1	520	40	277	80
23	Zhu *et al.* (51)	China	Nomogram	A-spondylitis	SVM-RFE	0	77	17	91	17

contained insufficient or unclear information, preventing a definitive assessment.

Univariate analysis

The studies were included in the meta-analysis, comprising 7,571 events. The pooled sensitivity, estimated using a random-effects model, was 0.895 with a 95% CI of 0.829–0.938, indicating a high overall diagnostic precision between studies. Sensitivity estimates in individual studies showed considerable variability, ranging from 0.125 (95% CI: 0.035–0.299) to 0.995 (95% CI: 0.972–1.000), reflecting differences in study populations, methodologies, and diagnostic criteria. Significant heterogeneity was observed between the studies (I^2 = 92.2%, τ^2 = 1.7224, P < 0.0001), suggesting that variation in sensitivity estimates is not only due to random chance but may be influenced by factors at the underlying level of the study. The high value I^2 indicates substantial inconsistency, warranting further exploration through subgroup analyses or meta-regression to identify potential sources of heterogeneity. Despite this variability, the consistently high sensitivity in most studies underscores the effectiveness of the diagnostic method in diverse clinical settings. Details are shown in Fig. 4.

In the specificity analysis, it included 8,241 events. The pooled specificity, calculated using a random-effects model, was 0.926 with a 95% CI of 0.872–0.959, reflecting high general precision in correctly identifying true negatives. Specificity estimates for individual studies ranged from 0.631 (95% CI: 0.614–0.648) to 1.000 (95% CI: 0.936–1.000), indicating variability in diagnostic performance depending on the context of the study and the methodologies applied.

Substantial heterogeneity was detected between studies (I^2 = 96.1%, τ^2 = 1.7642, P < 0.0001), suggesting that differences in study design, populations, or diagnostic thresholds contributed to the observed variation. The high value I^2 points to considerable inconsistency, highlighting the need for further subgroup analyses or meta-regression to identify potential sources of heterogeneity. Despite this, the consistently high specificity in most studies underscores the reliability of the diagnostic tool in diverse clinical environments. For more details, please see Fig. 5.

The analysis of the PLR yielded a mean value of 1,644.37, indicating a strong ability of the diagnostic test to confirm the presence of disease when the result is positive. The 95% CI for the PLR ranged from 73.56 to 3,215.18, reflecting considerable variability in the performance of the test between different studies or populations of patients. Although the high mean PLR suggests robust diagnostic utility, the wide CI indicates potential inconsistencies that may warrant further investigation to understand the underlying causes of this variation.

The NLR analysis showed a mean value of 0.179, indicating a high ability of the diagnostic test to rule out disease when the result is negative. The 95% CI for the NLR ranged from 0.095 to 0.263, suggesting relatively consistent diagnostic performance between studies. The low mean NLR reflects the test's effectiveness in minimizing false negatives, supporting its reliability as a diagnostic tool in various clinical contexts.

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

Table 3 Performance metrics of ML and DL models for ligament injury diagnosis (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

Study	Reference	SE	SP	Accuracy	F1-score
1	Astolfi *et al.* (29)	0.850	0.791	0.819	0.821
2	Chang *et al.* (30)	0.883	1.000	0.940	0.938
3	Cheng *et al.* (31)	0.852	0.850	0.852	0.904
4	Fang Liu *et al.* (32)	0.960	0.960	0.960	0.960
5	Germann *et al.* (33)	0.953	0.917	0.934	0.929
6	Guha Paul *et al.* (34)	0.995	1.000	0.998	0.997
7	Ito *et al.* (35)	0.968	0.731	0.831	0.829
8	Jo *et al.* (36)	0.820	0.940	0.880	0.872
9	Kanthavel *et al.* (37)	0.689	0.745	0.733	0.524
10	Kim *et al.* (38)	0.125	0.993	0.834	0.216
11	Liang *et al.* (39)	0.651	0.927	0.806	0.746
12	Michael *et al.* (40)	0.980	0.993	0.990	0.980
13	Minamoto *et al.* (41)	0.920	0.860	0.890	0.893
14	Mouchotte *et al.* (42)	0.940	1.000	0.943	0.969
15	Shemesh *et al.* (43)	0.976	0.873	0.917	0.909
16	Tamai *et al.* (44)	0.901	0.860	0.881	0.883
17	Tedesco *et al.* (45)	0.818	0.631	0.724	0.746
18	Wang *et al.* (46)	0.933	1.000	0.978	0.966
19	Whiteside *et al.* (47)	0.743	0.748	0.745	0.746
20	Yingkai *et al.* (48)	0.806	0.913	0.861	0.850
21	Zhang *et al.* (49)	0.976	0.949	0.963	0.965
22	Zhang *et al.* (50)	0.867	0.874	0.869	0.897
23	Zhu *et al.* (51)	0.819	0.843	0.832	0.819

The DOR, evaluated in 23 studies, yielded a pooled log DOR of 4.13 with a 95% CI of 3.57–4.70 (see Fig. 6). This result indicates a strong discriminatory power of the diagnostic test in differentiating between diseased and non-diseased individuals. Individual log DOR values ranged from 1.86 (95% CI: 1.45–2.28) to 10.88 (95% CI: 7.68–14.09), demonstrating variability in diagnostic performance between studies. Consistent positive log DOR values across studies suggest reliable diagnostic efficacy, although the observed range indicates that certain study-specific factors, such as population characteristics or methodological differences, may influence diagnostic precision.

Bivariate analysis

When analyzing subgroups based on the variable *g*, the group $g = 0$ exhibited an odds ratio (OR) of 50.33 (95% CI: 16.17–156.64) with $\tau^2 = 2.8078$ and $I^2 = 96\%$, indicating significant heterogeneity within this group. In contrast, the group $g = 1$ showed a higher OR of 112.15 (95% CI: 41.14–305.73), with $\tau^2 = 2.7734$ and $I^2 = 91.3\%$. Despite these differences, the test for subgroup differences yielded a χ^2 value of 1.07 with 1 degree of freedom ($P = 0.2998$), suggesting that there is insufficient evidence to assert significant differences in odds ratios between the two subgroups. The bivariate diagnostic random-effects meta-analysis (see Fig. 7), using the REML estimation method, showed significant fixed-effects coefficients. The intercept for logarithmic transformation sensitivity (tsens) was 2.050 (SE = 0.273, $z = 7.498$, $P < 0.0001$), while the intercept for logarithmic transformation false positive rate (tfpr) was −2.164 (SE = 0.223, $z = -9.698$, $P < 0.0001$). The estimated pooled sensitivity was 0.886 (95% CI: 0.820–0.930), and the false positive rate was 0.103 (95% CI: 0.069–0.151).

The variance components revealed standard deviations between studies of 1.230 for the sensitivity and 0.926 for the false positive rate, with a weak negative correlation of −0.140 between them. The logarithmic likelihood of the

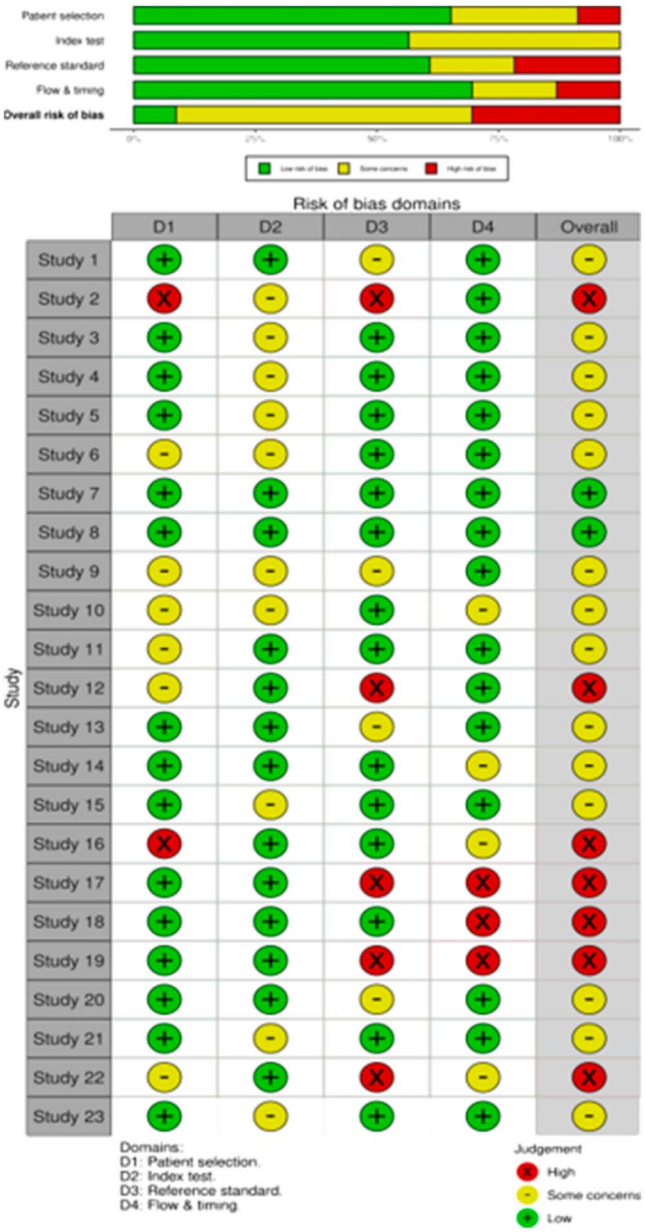

Figure 3

Quality assessment of included studies (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

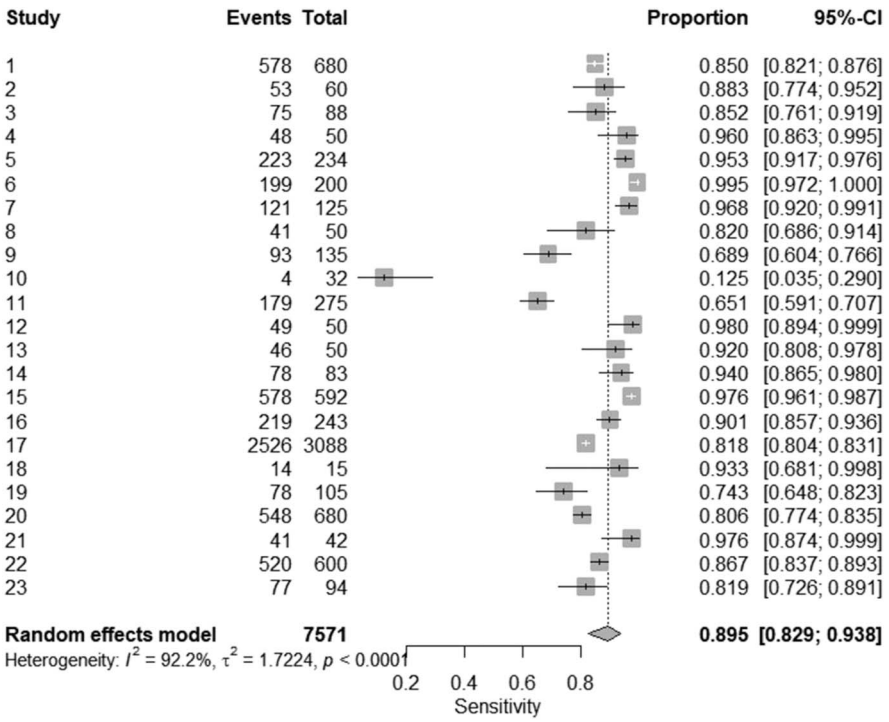

Figure 4

The forest plot of pooled sensitivity (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

model was 47.220, with an AIC of −84.440 and a BIC of −75.296. The area under the curve (AUC) was 0.948, indicating excellent diagnostic performance, while the partial AUC, restricted to observed false positive rates and normalized, was 0.884. For more details, please see Fig. 8.

The heterogeneity estimates varied depending on the method used. The Zhou and Dendukuri approach estimated I^2 at 24.8%, suggesting low heterogeneity (53). However, the Holling sample size-adjusted approach reported heterogeneity ranging from 68% to 95.7%, while the sample size-adjusted approach showed

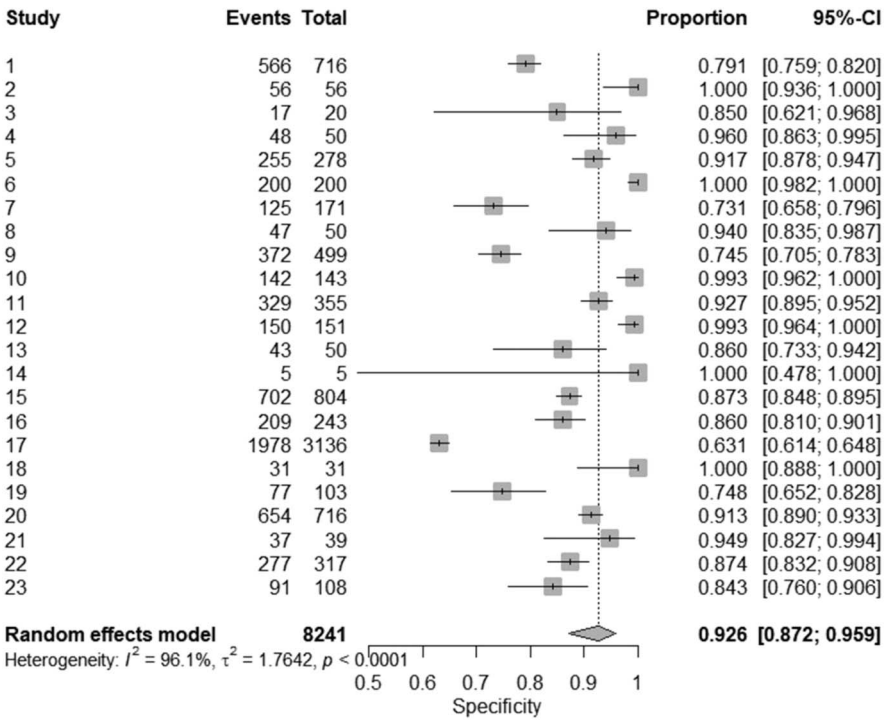

Figure 5

The forest plot of pooled specificity (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

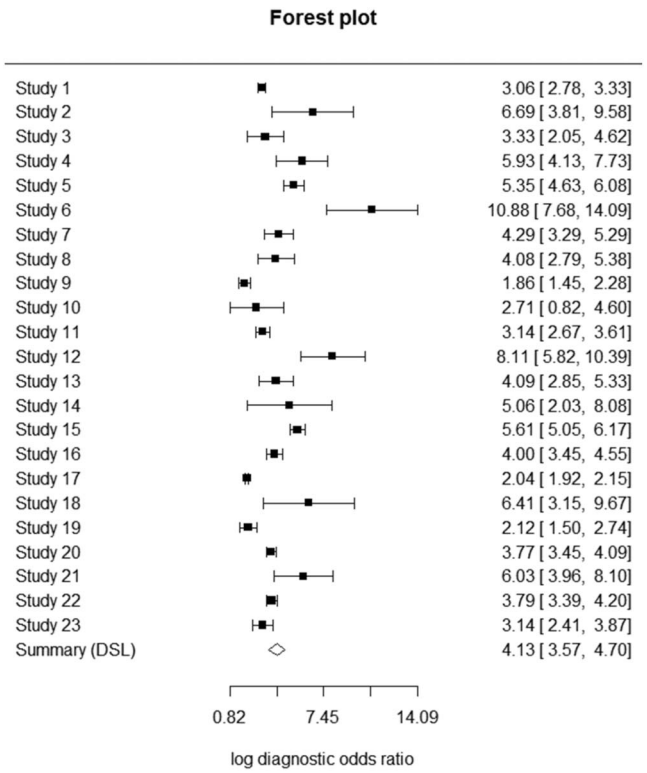

Figure 6

Forest plot of the log DORs (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

considerably lower heterogeneity, between 2.5 and 5.2% (54). These results highlight the strong diagnostic accuracy of the model while also indicating some variability between studies, influenced by methodological approaches and sample size adjustments.

Discussion

This study aimed to evaluate the diagnostic performance of various models of ML and DL to recognize ligament injuries in different medical imaging modalities. The findings support the clinical utility of the evaluated diagnostic tool, highlighting its high precision in identifying both positive and negative cases. Despite the variability observed among the studies, the consistency of its high sensitivity and specificity in different clinical settings reinforces its validity and applicability. However, the significant heterogeneity detected suggests that factors such as differences in study populations, methodological variations, and diagnostic thresholds may have influenced the results. In this context, further subgroup analyses, and meta-regressions are recommended to investigate potential sources of variability. Our findings support the application and usefulness of the ML and DL models in complex clinical settings.

The increasing interest of the clinical and scientific community in integrating AI-based tools into healthcare is evident and continues to grow. Recently, a consortium of 117 professionals from 50 countries reaffirmed that trusting AI in healthcare should follow 30 recommended best practices that cover its entire life cycle (55). In our study, this trend is reflected in the wide range of research utilizing AI to diagnose ligament injuries, highlighting the growing interest among specialists and healthcare teams in adopting these tools for the treatment of musculoskeletal pathology (26, 27). To our knowledge, this is the first study to simultaneously analyze such a diverse set of ML and DL models to detect these conditions while also identifying key challenges that must be addressed.

From a technical perspective, researchers are strongly encouraged to use multiple ML or DL algorithms to enhance the robustness and generalizability of their findings (56). This is achieved by simultaneously exploring and comparing different architectures, allowing their performance to be evaluated for the specific problem at hand. It is essential to recognize that no single model is universally superior; instead, the optimal choice depends on how well a model adapts to the data set and the specific requirements of the study (57). In this study, multiple articles reported the use of several models to identify ligament injuries. Although some studies used a single model (e.g. (35, 41, 43)), others tested up to five models simultaneously (36) or even six (30, 34, 45). One study (48) reported the use of nine models. This trend highlights not only the availability of various architectures for addressing the same problem but also the increasing integration of different ML and DL approaches to optimize diagnostic accuracy in diverse clinical settings. In 2021, a comprehensive guide and checklist for clinical research using AI was introduced, providing detailed recommendations on AI model design and training parameters (58). However, a key challenge remains: establishing a standardized minimum number of models required for diagnostic applications. This would help determine whether a proposed approach truly represents the best available option.

This also requires considering the computational resources necessary for data processing. More complex models demand greater computing power, often relying on graphics processing units or tensor processing units to efficiently handle large datasets and perform advanced calculations (52). In addition, these models typically require long training times, substantial storage, and significant energy consumption. Therefore, careful planning is essential to account for these factors and ensure efficient execution of the project (59). In addition, healthcare professionals must receive training not only in understanding these technical requirements

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

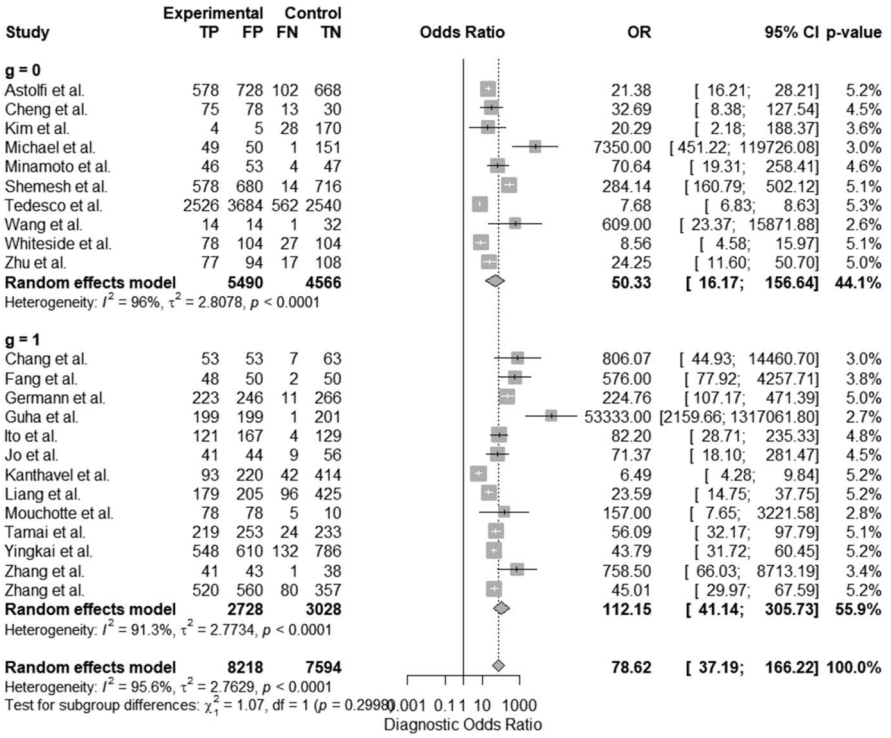

Figure 7

Forest plot of the odds ratios (29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51).

but also in correctly interpreting the models used. This ensures that the results can be applied safely and effectively in clinical practice (60).

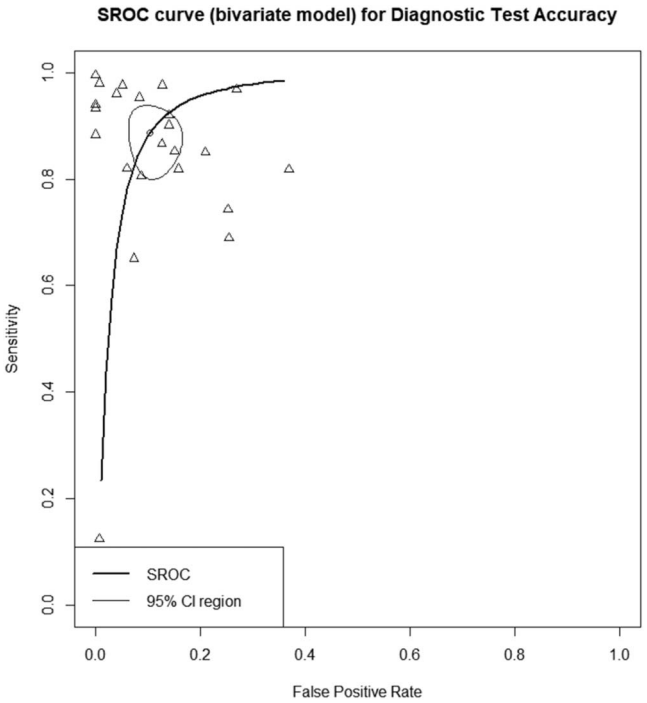

Figure 8

SROC curve (bivariate model) for diagnostic test accuracy.

One of the key challenges in using diagnostic support methods such as ML and DL is training, validation, and testing of models (61). During training, the model learns to classify or make predictions by progressively refining its calculations. This is achieved by repeatedly analyzing a dataset with known (labeled) responses and adjusting its parameters in each iteration. The training algorithm optimizes the predicted values relative to the actual values, and by cycling the dataset multiple times, an epoch is completed, improving accuracy with each pass (62). A critical challenge in this stage is the need for researchers to report the number of epochs used. This helps determine whether a model is undertrained (underfitting) or overfitted (overfitting). Striking the right balance in the number of epochs is essential to ensure optimal predictive performance and generalizability of the results.

The validation process presents a key challenge in the implementation of the model, as it assesses performance during training without modifying the model parameters. This is done using a dataset independent of the one used for training, which is why datasets are typically divided into unequal partitions for these processes. The primary goal of validation is to evaluate the ability of the model to generalize to new data, preventing overfitting (63). A crucial aspect of this stage is the correct segmentation and reporting of the data used for validation. Ensuring a representative sample is essential to accurately capture the variability of the study problem and provide reliable performance assessments.

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

Finally, the testing process represents the final phase of model evaluation and serves as an external validation of the results. The main challenge at this stage is to determine whether the previously obtained results can be applied to a completely new dataset. A common issue in this phase is selecting an appropriate test set that is sufficiently diverse and representative to prevent evaluation biases (64). Ensuring a well-chosen test set is crucial for accurately interpreting diagnostic metrics and confirming that the models can be effectively applied in real clinical settings.

These three phases highlight key considerations that must be addressed when reporting the results and performance of ML and DL models. It is recommended that the authors clearly state the sample size used in each stage (65). If the sample is too small, the model may struggle to capture meaningful relationships, potentially reducing its reliability. Therefore, we suggest reporting the total number of samples and, if possible, the class distribution to ensure transparency.

In addition, it is essential to report the number of data partitions and how they are allocated. For example, specifying the percentage of data assigned to training, validation, and testing (e.g. 70% training, 15% validation, 15% testing) improves reproducibility and clarity. Ensuring that the datasets are sufficiently large and representative is crucial to achieving robust and reliable model performance.

One of the main limitations of this study is that most of the articles analyzed focused on anterior cruciate ligament injuries in the knee, likely due to its high prevalence in the population. However, this emphasis on a single structure may reduce the generalizability of the findings. Future research should consider evaluating a broader range of ligament injuries to ensure the applicability of these models in different clinical scenarios (66). Another potential limitation is the availability of sufficient studies to rigorously assess the diagnostic performance of these models. In previous research, we have emphasized the importance of having a sufficient number of articles for a comprehensive evaluation. However, since the adoption of these technologies in clinical practice has been gradual, relevant data remain limited in some cases. To address this, we have recommended the use of models as units of analysis (26, 27). In this study, we identified a large number of articles and a diverse range of reported models, which allowed us to select only those with the best performance. This selection underscores the need to standardize the evaluation criteria. Specifically, it remains unclear whether researchers should report only the best-performing model, the average of all metrics, or include both the highest and lowest-performing models to assess variability. Standardizing these practices would improve comparability and reliability in future studies.

In addition to standardization, a critical step toward real-world implementation is multicenter external validation. Most of the studies included in this meta-analysis relied on single-center data, often with homogeneous imaging protocols, which may inflate diagnostic performance. Upcoming studies must incorporate external validation across diverse institutions, imaging devices, and patient populations. This approach not only reduces the risk of overfitting but also ensures that models are generalizable and reliable in heterogeneous clinical environments. Recent methodological guidance strongly emphasizes this point, underscoring that multicenter external validation is essential for safe translation of AI into musculoskeletal practice (63, 66, 67, 68). Addressing this gap will be fundamental to moving from proof-of-concept studies to clinically deployable diagnostic tools.

The growing interest among musculoskeletal clinicians in integrating these technologies into clinical practice is evident, and their use will undoubtedly continue to expand. However, for these models to be more clinically valuable, they must be able to identify multiple structures or types of injuries simultaneously. In other words, they should support multi-label classification to detect multiple findings within the same assessment (69). A fundamental challenge in this field is not only achieving high diagnostic accuracy but also ensuring comprehensive monitoring of the structures surrounding the injury site. Enhancing these capabilities will be crucial to improving patient management and optimizing clinical decision making.

Soon, countries must establish a robust national policy on AI (70), providing clear guidelines for a regulatory and legislative framework that governs the use of patient data and information. This policy should go beyond technical and security aspects to safeguard patient rights, ensuring compliance with ethical and legal principles in the application of AI-driven technologies in healthcare. In addition, the role of scientific ethics committees must be strengthened, as they serve as key bodies for health research projects. These committees should incorporate updated guidelines to assess the implications of AI in clinical settings, particularly with regard to algorithm transparency, fairness in outcomes, and potential biases in medical decision making. Ensuring ethical AI implementation will be critical to maintaining trust and equity in healthcare innovation.

Conclusion

This meta-analysis clearly demonstrates that the ML and DL models evaluated improve diagnostic accuracy in clinical settings, particularly when supported by imaging resources or novel diagnostic approaches for ligament injuries. Furthermore, this study highlights the growing diversity of AI-based models, reinforcing the expectation that their application in musculoskeletal pathology will continue to expand. However, the widespread adoption of these technologies will not depend solely on the interest of healthcare professionals in improving clinical

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

performance. It will also require the participation of health system decision makers and policy makers, who must establish robust public policies to facilitate the responsible and effective integration of AI into clinical practice.

Several challenges remain to be addressed. For example, the diversity of available models requires a thorough understanding by healthcare professionals to determine which model best fits the specific clinical problem. In addition, these models must be capable of accurately distinguishing ligament injuries from other musculoskeletal pathologies, minimizing false positives and false negatives, particularly in cases with subtle findings or suboptimal image quality. To mitigate evaluation biases, it is essential to train models in diverse clinical settings to prevent both overestimation and underestimation of diagnostic performance. This requires adjusting for variability in environmental conditions, injury characteristics, anatomical differences, and other relevant factors. A key step toward improving AI-driven diagnostic strategies would be the establishment of open access data repositories by scientific societies. These repositories could facilitate the development of more robust and generalizable models, ultimately enhancing their reliability and effectiveness in clinical practice.

The findings of this study underscore the need to explore additional diagnostic modalities to identify ligament injuries. The results suggest that ML and DL methods hold promise as diagnostic support tools, particularly when integrated with conventional techniques. Their implementation has the potential to improve diagnostic accuracy while offering more cost-effective strategies in clinical practice.

Despite these advancements, several limitations were identified. First, the ability of AI models to simultaneously detect concomitant injuries within the analyzed structures remains limited, as does their capacity for multi-label classification. In addition, most of the reviewed studies have focused on detecting ligament injuries in the lower extremity, particularly ACL injuries. This narrow scope restricts the generalizability of the results and impacts the external validity of the models. Finally, the lack of standardization in reporting training, validation, and test sets remains a challenge, hindering the ability to obtain precise performance metrics. The lack of detailed sample size information for each development phase further complicates the comparative assessment of the diagnostic accuracy of the model. Addressing these limitations is crucial for advancing the reliability and applicability of AI-driven diagnostic tools in musculoskeletal medicine.

ICMJE Statement of Interest

The authors affirm that there is no conflict of interest that could compromise the objectivity or impartiality of the reported study.

Funding Statement

This study was not supported by any specific grant from public, commercial, or nonprofit funding agencies.

Author contribution statement

GD was responsible for conceptualization, software, statistical analysis, and writing the original draft. GD, EV, and JV were responsible for data curation. CJ and FF were responsible for writing review. FF was responsible for supervision and editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

The authors express their sincere gratitude to the MEDS-PUCV Sports Medicine Data Science Center.

References

1 Jan Z, Ahamed F, Mayer W, *et al.* Artificial intelligence for industry 4.0: systematic review of applications, challenges, and opportunities. *Expert Syst Appl* 2023 **216** 119456. (https://doi.org/10.1016/j.eswa.2022.119456)

2 Yu KH, Beam AL & Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018 **2** 719–731. (https://doi.org/10.1038/s41551-018-0305-z)

3 Bhavsar KA, Singla J, Al-Otaibi YD, *et al.* Medical diagnosis using machine learning: a statistical review. *Comput Mater Continua* 2021 **67** 107–125. (https://doi.org/10.32604/cmc.2021.014604)

4 Tzanakou EM. *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence*, pp 1–392. Boca Raton, USA: CRC Press, 2017.

5 Jena OP, Bhushan B & Kose U. *Machine Learning and Deep Learning in Medical Data Analytics and Healthcare Applications*, pp 1–292. Boca Raton, USA: CRC Press, 2022.

6 D'Angelo T, Caudo D, Blandino A, *et al.* Artificial intelligence, machine learning and deep learning in musculoskeletal imaging: current applications. *J Clin Ultrasound* 2022 **50** 1414–1431. (https://doi.org/10.1002/jcu.23321)

7 Jungo A, Meier R, Ermis E, *et al*. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference*, pp 682–690. Granada, Spain: Springer, 2018. September 16-20, 2018, Proceedings, Part I. Notes in Computer Science, vol. **11070**. Cham: Springer; pp 682–690. (https://doi.org/10.1007/978-3-030-00928-1)

8 Tran AQ, Nguyen LH, Nguyen HSA, *et al.* Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. *Front Public Health* 2021 **9** 755644. (https://doi.org/10.3389/fpubh.2021.755644)

9 Teney D, Abbasnejad E, Lucey S, *et al*. Evading the simplicity bias: training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 16761–16772. IEEE/CVF, 2022. (https://doi.org/10.1109/CVPR52688.2022.01626)

10 Hu X, Chu L, Pei J, *et al.* Model complexity of deep learning: a survey. *Knowl Inf Syst* 2021 **63** 2585–2619. (https://doi.org/10.1007/s10115-021-01605-0)

EFORT Open Reviews (2026) **11** 3–16
https://doi.org/10.1530/EOR-2025-0038

11 PlÖtz T. Applying machine learning for sensor data analysis in interactive systems: common pitfalls of pragmatic use and ways to avoid them. *ACM Comput Surv* 2021 **54** 1–25. (https://doi.org/10.1145/3459666)

12 Meena KS & Suriya S. A survey on supervised and unsupervised learning techniques. In *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019*, LA Kumar, LS Jayashree, R Manimegalai (Eds). pp 627–644. Cham: Springer, 2020. (https://doi.org/10.1007/978-3-030-24051-6_58)

13 Begum M & Uddin MS. Analysis of digital image watermarking techniques through hybrid methods. *Adv Multimed* 2020 **2020** 7912690. (https://doi.org/10.1155/2020/7912690)

14 Khan S, Rahmani H, Shah SAA, *et al*. A Guide to Convolutional Neural Networks for Computer Vision. In *Synthesis Lectures on Computer Vision*, G Medioni, S Dickinson (Eds). pp 1–207. San Rafael, CA: Morgan & Claypool, 2018. (https://doi.org/10.2200/S00822ED1V01Y201712COV015)

15 Sebbag E, Felten R, Sagez F, *et al.* The world-wide burden of musculoskeletal diseases: a systematic analysis of the World Health Organization burden of diseases database. *Ann Rheum Dis* 2019 **78** 844–848. (https://doi.org/10.1136/annrheumdis-2019-215142)

16 Provenzano PP, Heisey D, Hayashi K, *et al.* Subfailure damage in ligament: a structural and cellular evaluation. *J Appl Physiol* 2002 **92** 362–371. (https://doi.org/10.1152/jappl.2002.92.1.362)

17 Cao S, Wang C, Ma X, *et al.* Imaging diagnosis for chronic lateral ankle ligament injury: a systemic review with meta-analysis. *J Orthop Surg Res* 2018 **13** 1–14. (https://doi.org/10.1186/s13018-018-0811-4)

18 Li K, Du J, Huang LX, *et al.* The diagnostic accuracy of magnetic resonance imaging for anterior cruciate ligament injury in comparison to arthroscopy: a meta-analysis. *Sci Rep* 2017 **7** 7583. (https://doi.org/10.1038/s41598-017-08133-4)

19 Krastman P, Mathijssen NM, Bierma-Zeinstra SM, *et al.* Diagnostic accuracy of history taking, physical examination and imaging for non-chronic finger, hand and wrist ligament and tendon injuries: a systematic review update. *BMJ Open* 2020 **10** e037810. (https://doi.org/10.1136/bmjopen-2020-037810)

20 Arzehgar A, Seyedhasani SN, Ahmadi FB, *et al.* Sensor-based technologies for motion analysis in sports injuries: a scoping review. *BMC Sports Sci Med Rehabil* 2025 **17** 15. (https://doi.org/10.1186/s13102-025-01063-z)

21 Tan T, Gatti AA, Fan B, *et al.* A scoping review of portable sensing for out-of-lab anterior cruciate ligament injury prevention and rehabilitation. *NPJ Digit Med* 2023 **6** 46. (https://doi.org/10.1038/s41746-023-00782-2)

22 Wang X, Yu H, Kold S, *et al.* Wearable sensors for activity monitoring and motion control: a review. *Biomimetic Intelligence Robotics* 2023 **3** 100089. (https://doi.org/10.1016/j.birob.2023.100089)

23 Hauser RA & Dolan EE. Ligament injury and healing: an overview of current clinical concepts. *J Prolotherapy* 2011 **3** 836–846.

24 Blalock D, Miller A, Tilley M, *et al.* Joint instability and osteoarthritis. *Clin Med Insights Arthritis Musculoskelet Disord* 2015 **8** 15–23. (https://doi.org/10.4137/CMAMD.S22147)

25 Riemann BL & Lephart SM. The sensorimotor system, part II: the role of proprioception in motor control and functional joint stability. *J Athletic Train* 2002 **37** 80.

26 Droppelmann G, Rodríguez C, Jorquera C, *et al.* Artificial intelligence in diagnosing upper limb musculoskeletal disorders: a sys- tematic review and meta-analysis of diagnostic tests. *EFORT Open Rev* 2024 **9** 241–251. (https://doi.org/10.1530/eor-23-0174)

27 Droppelmann G, Rodríguez C, Smague D, *et al.* Deep learning models for tendinopathy detection: a systematic review and meta-analysis of diagnostic tests. *EFORT Open Rev* 2024 **9** 941–952. (https://doi.org/10.1530/eor-24-0016)

28 Whiting PF, Rutjes AW, Westwood ME, *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011 **155** 529–536. (https://doi.org/10.7326/0003-4819-155-8-201110180-00009)

29 Astolfi RS, da Silva DS, Guedes IS, *et al.* Computer-aided ankle ligament injury diagnosis from magnetic resonance images using machine learning techniques. *Sensors* 2023 **23** 1565. (https://doi.org/10.3390/s23031565)

30 Chang PD, Wong TT & Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. *J Digit Imag* 2019 **32** 980–986. (https://doi.org/10.1007/s10278-019-00193-4)

31 Cheng Q, Lin H, Zhao J, *et al.* Application of machine learning-based multi-sequence MRI radiomics in diagnosing anterior cruciate ligament tears. *J Orthop Surg Res* 2024 **19** 99. (https://doi.org/10.1186/s13018-024-04602-5)

32 Liu F, Guan B, Zhou Z, *et al.* Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. *Radiol Artif Intelligence* 2019 **1** 180091. (https://doi.org/10.1148/ryai.2019180091)

33 Germann C, Marbach G, Civardi F, *et al.* Deep convolutional neural network–based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. *Investig Radiol* 2020 **55** 499–506. (https://doi.org/10.1097/rli.0000000000000664)

34 Paul SG, Saha A & Assaduzzaman M. A real-time deep learning approach for classifying cervical spine fractures. *Healthc Analytics* 2023 **4** 100265. (https://doi.org/10.1016/j.health.2023.100265)

35 Ito S, Nakashima H, Segi N, *et al.* Automated detection of the thoracic ossification of the posterior longitudinal ligament using deep learning and plain radiographs. *Biomed Res Int* 2023 **2023** 8495937. (https://doi.org/10.1155/2023/8495937)

36 Jo SW, Khil EK, Lee KY, *et al.* Deep learning system for automated detection of posterior ligamentous complex injury in patients with thoracolumbar fracture on MRI. *Sci Rep* 2023 **13** 19017. (https://doi.org/10.1038/s41598-023-46208-7)

37 Kanthavel R & Dhaya R. Prediction model using reinforcement deep learning technique for osteoarthritis disease diagnosis. *Comput Syst Sci Eng* 2022 **42** 257–269. (https://doi.org/10.32604/csse.2022.021606)

38 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–E425. (https://doi.org/10.1097/bsd.0000000000001295)

39 Liang C, Li X, Qin Y, *et al.* Effective automatic detection of anterior cruciate ligament injury using convolutional neural network with two attention mechanism modules. *BMC Med Imag* 2023 **23** 120. (https://doi.org/10.1186/s12880-023-01091-6)

40 Richardson ML. MR protocol optimization with deep learning: a proof of concept. *Curr Probl Diagn Radiol* 2021 **50** 168–174. (https://doi.org/10.1067/j.cpradiol.2019.10.004)

41 Minamoto Y, Akagi R, Maki S, *et al.* Automated detection of anterior cruciate ligament tears using a deep convolutional neural network. *BMC Muscoskelet Disord* 2022 **23** 577. (https://doi.org/10.1186/s12891-022-05524-1)

42 Mouchotte J, LeBerre M, Cojean T, *et al.* New interpretation of GNRB® knee arthrometer results for ACL injury diagnosis support using machine learning. *Machine Learn Appl* 2023 **13** 100480. (https://doi.org/10.1016/j.mlwa.2023.100480)

43 Shemesh S, Kimchi G, Yaniv G, *et al.* MRI-based detection of cervical ossification of the posterior longitudinal ligament using a novel automated machine learning diagnostic tool. *Neurosurg Focus* 2023 **54** E11. (https://doi.org/10.3171/2023.3.focus2390)

44 Tamai K, Terai H, Hoshino M, *et al.* A deep learning algorithm to identify cervical ossification of posterior longitudinal ligaments on radiography. *Sci Rep* 2022 **12** 2113. (https://doi.org/10.1038/s41598-022-06140-8)

45 Tedesco S, Crowe C, Ryan A, *et al.* Motion sensors-based machine learning approach for the identification of anterior cruciate ligament gait patterns in on-the-field activities in rugby players. *Sensors* 2020 **20** 3029. (https://doi.org/10.3390/s20113029)

46 Wang Q, Zhao W, Xing X, *et al.* Feasibility of AI-assisted compressed sensing protocols in knee MR imaging: a prospective multi-reader study. *Eur Radiol* 2023 **33** 8585–8596. (https://doi.org/10.1007/s00330-023-09823-6)

47 Whiteside D, Martini DN, Lepley AS, *et al.* Predictors of ulnar collateral ligament reconstruction in major league baseball pitchers. *Am J Sports Med* 2016 **44** 2202–2209. (https://doi.org/10.1177/0363546516643812)

48 Ma Y, Qin Y, Liang C, *et al.* Visual cascaded-progressive convolutional neural network (C-PCNN) for diagnosis of Meniscus injury. *Diagnostics* 2023 **13** 2049. (https://doi.org/10.3390/diagnostics13122049)

49 Zhang L, Li M, Zhou Y, *et al.* Deep learning approach for anterior cruciate ligament lesion detection: evaluation of diagnostic performance using arthroscopy as the reference standard. *J Magn Reson Imag* 2020 **52** 1745–1752. (https://doi.org/10.1002/jmri.27266)

50 Zhang M, Huang C & Druzhinin Z. A new optimization method for accurate anterior cruciate ligament tear diagnosis using convolutional neural network and modified golden search algorithm. *Biomed Signal Process Control* 2024 **89** 105697. (https://doi.org/10.1016/j.bspc.2023.105697)

51 Zhu J, Lu Q, Liang T, *et al.* Development and validation of a machine learning-based nomogram for prediction of ankylosing spondylitis. *Rheumatol Ther* 2022 **9** 1377–1397. (https://doi.org/10.1007/s40744-022-00481-6)

52 Kimm H, Paik I & Kimm H. Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. In *Proceedings of the 2021 IEEE 14th International Symposium on Embedded Multicore/ Many-Core Systems-on-Chip (MCSoC)*, pp 312–319. IEEE, 2021. (https://doi.org/10.1109/MCSoC51149.2021.00053)

53 Zhou Y & Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med* 2014 **33** 2701–2717. (https://doi.org/10.1002/sim.6115)

54 Holling H, Böhning W, Masoudi E, *et al*. Evaluation of a new version of I^2 with emphasis on diagnostic problems. *Commun Stat Simulat Comput* 2020 **49** 942–972. (https://doi.org/10.1080/03610918.2018.1489553)

55 Lekadir K, Frangi AF, Porras AR, *et al*. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025 **388** e081554. (https://doi.org/10.1136/bmj-2024-081554)

56 Javed H, El-Sappagh S & Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artif Intell Rev* 2025 **58** 1–107. (https://doi.org/10.1007/s10462-024-11005-9)

57 Alnuaimi AF & Albaldawi TH. An overview of machine learning classification techniques. *BIO Web of Conferences* 2024 **97** 00133 EDP Sciences. (https://doi.org/10.1051/bioconf/20249700133)

58 Olczak J, Pavlopoulos J, Prijs J, *et al.* Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a clinical AI research (CAIR) checklist proposal. *Acta Orthop* 2021 **92** 513–525. (https://doi.org/10.1080/17453674.2021.1918389)

59 Chen C, Zhang P, Zhang H, *et al.* Deep learning on computational-resource-limited platforms: a survey. *Mob Inf Syst* 2020 **2020** 8454327. (https://doi.org/10.1155/2020/8454327)

60 Liu X, Faes L, Kale AU, *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 Digital Health* 2019 **1** e271–e297. (https://doi.org/10.1016/s2589-7500(19)30123-2)

61 Eelbode T, Sinonquel P, Maes F, *et al.* Pitfalls in training and validation of deep learning systems. *Best Pract Res Clin Gastroenterol* 2021 **52** 101712. (https://doi.org/10.1016/j.bpg.2020.101712)

62 Langer M, He Z, Rahayu W, *et al.* Distributed training of deep learning models: a taxonomic perspective. *IEEE Trans Parallel Distr Syst* 2020 **31** 2802–2818. (https://doi.org/10.1109/tpds.2020.3003307)

63 Cabitza F, Campagner A, Soares F, *et al.* The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Progr Biomed* 2021 **208** 106288. (https://doi.org/10.1016/j.cmpb.2021.106288)

64 Sekhon J & Fleming C. Towards improved testing for deep learning. In *Proceedings of the 41st International Conference on Software Engineer-Ing: New Ideas and Emerging Results (ICSE-NIER 2019)*, pp 85–88. IEEE, 2019. (https://doi.org/10.1109/ICSE-NIER.2019.00030)

65 Balki I, Amirabadi A, Levman J, *et al.* Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 2019 **70** 344–353. (https://doi.org/10.1016/j.carj.2019.06.002)

66 Yu AC, Mohajer B & Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiology Artif Intell* 2022 **4** e210064. (https://doi.org/10.1148/ryai.210064)

67 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024 **385** e078378. (https://doi.org/10.1136/bmj-2023-078378)

68 Santos CS & Amorim-Lopes M. Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review. *BMC Med Res Methodol* 2025 **25** 45. (https://doi.org/10.1186/s12874-025-02463-y)

69 Droppelmann G, Tello M, García N, *et al.* Lateral elbow tendinopathy and artificial intelligence: binary and multilabel findings detection using machine learning algorithms. *Front Med* 2022 **9** 945698. (https://doi.org/10.3389/fmed.2022.945698)

70 Scrollini F, Cervantes ME & Mariscal J. En busca de rumbo: el estado de las políticas de inteligencia artificial en América Latina. *Banco Interamericano De Desarrollo BID* 2021. (https://www.empatia.la/wp-content/uploads/2021/11/23NOVPOLICYEMPATIA.pdf)