

Deep learning classification models demonstrate high accuracy and clinical potential in radiograph interpretation in the arthroplasty clinical pathway: A systematic review and meta-analysis

Omar Musbahi^{1,2}  | Amal Ahmed¹ | Thomas Hall³ | Kamila Kutyla¹ |
 Muhayman Sadiq¹  | Justin Cobb¹ | Alex Bottle² | Richard Van Arkel^{1,3} |
 Gareth Jones¹

¹Imperial College London, MSK Lab, London, UK

²Imperial College London, School of Public Health, London, UK

³Department of Mechanical Engineering, Imperial College London, London, UK

Correspondence

Omar Musbahi, Imperial College London, MSK Lab Charing Cross Campus, Laboratory Block, Hammersmith, London W6 8RP, UK.
 Email: o.musbahi19@ic.ac.uk

Funding information

NIHR, Grant/Award Number: 302632

Abstract

Purpose: Imaging is a cornerstone of the osteoarthritis (OA)-arthroplasty clinical pathway: the continuum of care that patients with OA undergo, from initial diagnosis through to arthroplasty and postoperative follow-up. With growing interest from the orthopaedic community, this meta-analysis broadly evaluates the performance of deep learning algorithms in interpreting radiographs and cross-sectional imaging in this pathway. The authors hypothesise that deep learning algorithms will have comparable performance to clinicians when interpreting radiographs, but not cross-sectional imaging, with negligible difference in diagnostic and prognostic tasks.

Methods: Ovid Medline, Ovid Embase, Scopus and Web of Science were searched for studies published between January 1, 2012, and April 1, 2024, evaluating deep learning algorithms for diagnostic and prognostic tasks along the pathway. Eligible studies included those that used established diagnostic or surgical candidacy assessments as ground truth. Study quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 tool, and pooled sensitivity and specificity were determined. Hierarchical summary receiver operating characteristic curves assessed diagnostic performance.

Results: The meta-analysis of artificial intelligence (AI) interpretation included 66 studies, with a pooled sensitivity of 0.88 (95% confidence interval [CI]: 0.81–0.92) and a pooled specificity of 0.91 (95% CI: 0.87–0.94). Sensitivity and specificity values were higher for AI interpretation of radiographs (55 studies) compared to cross-sectional imaging, with

Abbreviations: AI, artificial intelligence; AUC, area under curve; CNN, convolutional neural network; CT, computed tomography; DenseNet, dense convolutional network; DL, deep learning; GAN, generative adversarial network; GMLP, good machine learning practice; HSROC, hierarchical summary receiver operating curve; LSTM, long short-term memory; MRI, magnetic resonance imaging; OA, osteoarthritis; PET, positron emission tomography; PICO, patient/population, intervention, comparison and outcomes; PRISMA, preferred reporting items for systematic reviews and meta-analyses; QUADAS-2, quality assessment of diagnostic accuracy studies 2; ResNet, residual network; RNN, recurrent neural network; SPECT, single photon emission computed tomography; THA, total hip arthroplasty; VGG, visual geometry group; ViT, vision transformers.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Journal of Experimental Orthopaedics* published by John Wiley & Sons Ltd on behalf of European Society of Sports Traumatology, Knee Surgery and Arthroscopy.

no significant difference in performance between diagnostic and prognostic tasks. For clinician interpretation, 11 studies showed a pooled sensitivity of 0.76 (95% CI: 0.64–0.85) and a pooled specificity of 0.79 (95% CI: 0.59–0.90).

Conclusions: This meta-analysis highlights the potential of deep learning algorithms to improve efficiency in OA classification and prognosis in the arthroplasty pathway based on low-to-moderate quality evidence. Although the results are not generalisable, the findings suggest deep learning models have the potential to be adopted in OA treatment pathways, warranting further exploration of its role in patient care.

Level of Evidence: Level III.

KEYWORDS

arthroplasty, artificial intelligence, convolutional neural network, deep learning, radiology

INTRODUCTION

Osteoarthritis (OA) is a condition that affects over 500 million people worldwide, with over 40 million diagnosed in 2019 according to the latest Global Burden of Disease study [119]. End-stage treatment of OA typically requires arthroplasty, and, like the incidence of OA, the number of patients undergoing arthroplasty is also increasing. In 2019, over 450,000 total knee arthroplasties and more than 260,000 total hip arthroplasties were performed in the United States, accounting for over 2% of adults aged 50 and older [95, 112, 148]. From the initial presentation of symptomatic OA to postoperative assessment after arthroplasty, imaging is a key pillar of the OA clinical pathway [30, 57].

Amidst the increasing incidence of OA and demand for arthroplasty, there has been a concomitant rise in imaging volumes—a cornerstone of the OA-arthroplasty clinical pathway. Automated image analysis is emerging as a potential solution for improving efficiency, accuracy and consistency in imaging interpretation across various stages of the pathway [55, 87]. Artificial intelligence (AI) is being deployed in medical image analysis in a variety of clinical settings, showing the potential to serve as a useful and efficient adjunct to human-only interpretation [78]. In the musculoskeletal field, several authors have demonstrated the advantages of AI-based image analysis for many types of tasks, including aiding diagnosis and prognosis [22, 69, 74]. In the OA-arthroplasty clinical pathway, AI has been used for disease severity grading [23, 85], alignment measurement [141, 150], and the detection of postoperative complications [155]. However, despite increasing research into the integration of AI in orthopaedic imaging, to date, there has not been a systematic review that overviews the utility of AI for diagnostic and prognostic purposes on radiographs and cross-sectional imaging in

the arthroplasty pathway. Understanding the utility of AI for these purposes across different imaging modalities is important for guiding for future research and appreciating where AI can best be implemented into clinical practice in the future.

Amidst strong interest from the orthopaedic community and the first wave of regulatory approvals [35], the current evidence for AI-based image analysis in the OA-arthroplasty clinical pathway was quantitatively appraised. This study aimed to evaluate the efficacy of AI-based algorithms in interpreting radiographs and cross-sectional imaging for diagnostic and prognostic purposes across all stages of the clinical pathway and compare their performance to that of human clinicians. Focusing on deep learning, a class of AI well-suited to image analysis, a broad systematic review summarising the performance of AI across all stages of the clinical pathway was performed. The accuracy and associated biases of the algorithms in the dataset were assessed in the meta-analysis. Furthermore, the efficacy of AI in interpreting radiographs and cross-sectional imaging for diagnostic and prognostic purposes was compared. A comparison between the overall performance of AI-based algorithms and human clinician performance was undertaken. The authors hypothesise that deep learning algorithms will show comparable performance to human clinicians when interpreting imaging for both diagnostic and prognostic purposes on radiographs, but significantly worse performance on cross-sectional imaging in comparison.

METHODS

This study was performed in accordance with the preferred reporting items for systematic review and meta-analysis (PRISMA) guidelines and registered prospectively on PROSPERO CRD42020198878.

Search strategy and selection criteria

In this study, the OA-arthroplasty clinical pathway was defined as the care pathway for patients with OA spanning from initial diagnosis and classification to postoperative care and follow-up. Studies applying deep learning to medical imaging at any stage of the OA-arthroplasty clinical pathway, including the evaluation of preoperative conditions (e.g., OA) and postoperative complications (e.g., aseptic loosening), were searched for. All major joints (hip, knee, ankle, elbow and shoulder) and the most common imaging modalities were considered: plain radiography (X-ray), computed tomography (CT), magnetic resonance imaging (MRI), single-photon emission computed tomography (SPECT), position emission tomography (PET) and scintigraphy.

The use of deep learning for image interpretation was considered to be the intervention within the context of the population, intervention, comparison and outcome (PICO) framework. For the purposes of comparability between studies, only classification algorithms were considered. Contextual examples of classification tasks include diagnosing disease, grading its severity, predicting its progression or outcome, detecting postoperative complications and identifying implant models. Acquisition, preprocessing and segmentation were deemed out of scope as they are evaluated using alternative or inconsistent metrics, and are conceptually different.

The search covered the following databases: Ovid-MEDLINE, Embase, Scopus and Web of Science, with full search terms available as File S1. Where appropriate, search terms were mapped to either medical subject headings (MeSH) or non-MeSH. Further articles were also identified by manually searching the bibliographies, citations and related articles (as recommended by PubMed) of included articles. For contemporaneity, the search only considered studies published between 1 January 2012 and 1 April 2024. The earlier date corresponded to the publication year of AlexNet, which is widely considered a paradigm shift in the performance of Deep Learning algorithms in image classification tasks [67]. Only original research articles were considered eligible for the systematic review, inclusive of conference proceedings, letters, preprints and scientific reports. The search was addressed to English articles, but no articles were discounted due to their language.

Eligibility assessment

Articles were first screened by two independent reviewers based on titles and abstracts (A.A. and K.K.); nonconsensus was resolved by a third reviewer (O.M.). The full text of candidate articles was reviewed before

the final selection. Where the same content formed substantially part of two separate publications by the same authors, only the most developed study was included, for example, the study with the greatest number of subjects. Studies by the same authors but with a substantive change, such as applying the same framework to a different database, were deemed to be distinct and both were included.

Nonhuman and ex vivo studies were excluded. However, no other limitations were placed on the demographics of each included study. Age and sex were not reported given the nature of the study. Only pure approaches to image interpretation using deep learning were considered eligible; hybrid methods that required clinician input to prescore or segment images were excluded. For studies where clinician interpretation results were collected, these interpretations could have come from radiologists, orthopaedic surgeons or both.

Data extraction

The data extracted from selected articles included identifying information such as the title of the study, authors, study type, publication type and digital object identifiers. The study characteristics were also extracted: target anatomy, target condition, purpose, imaging modality, image source, sample size, sample composition (e.g., split between positive-negative indication), train-validation-test split ratios, AI algorithm (neural network architecture), method of acquiring and definition of ground truth and the validation method used.

Performance metrics of the deep learning algorithm and, where possible, comparative performance of the clinician were also investigated. Metrics included specificity, sensitivity, area-under-the-curve (AUC), precision, average precision, F1 score (combines precision and recall scores of a deep learning model to measure its accuracy), Jaccard Index (measures similarity between two sample sets by calculating the ratio of their intersection [similar elements between sets] to their union [the combined size of both sample sets]), Dice coefficient (measures similarity between two sample sets by doubling their intersection term relative to the combined sample size, typically yielding higher values), true positive rate, false positive rate, false negative rate and true negative rate [45, 50]. In this context, sensitivity refers to the performance of AI or clinicians to correctly identify abnormalities on imaging, and specificity refers to the performance of AI or clinicians to avoid false positive diagnoses when interpreting imaging. The Jaccard Index and Dice Score are used to understand similarities between two sample sets: The Jaccard Index is the size of the intersection divided by the size of the union between two sets, and the Dice Score in most cases, these data were

extracted from direct reporting in the article. However, in cases where such data was not explicitly stated, it was ascertained—where possible—from other published information, such as by calculating from confusion matrices or other reported performance metrics. Where there was insufficient information, corresponding authors were contacted to supply the required data. Failing this, the study was excluded from further analysis.

Analysis

A hierarchical summary receiver operating characteristic (HSROC) curve was constructed on STATA using a bivariate random-effects meta-analytic model to pool sensitivity and specificity data from individual studies [105]. This approach accounts for the inherent correlation between sensitivity and specificity and incorporates between-study heterogeneity. The HSROC model estimates key parameters, including accuracy (Λ), threshold (Θ) and asymmetry (β), to characterise the trade-off between sensitivity and specificity across studies. Pooled sensitivity and specificity were derived from the model's fixed and random effects, with studies weighted by their variance (precision). The AUC, reflecting overall diagnostic performance, was computed directly from the HSROC curve, integrating the relationship between sensitivity and specificity across thresholds. This method ensures robust estimates that are less biased by study heterogeneity or imbalances in sample size. HSROC curves from the study results from the following groups were plotted: overall AI performance, AI performance for diagnostic purposes, AI performance for prognostic purposes, AI performance with radiographs, AI performance with cross-sectional imaging, and overall human performance. Using study characteristics such as sample size, image source, algorithm choice, train-test split ratios and ground truth acquisition, a qualitative analysis of best practices was undertaken.

For studies reporting multi-class outputs, the performance metrics were aggregated into a single representative variable, with the method dependent on the output type: semi-continuous or discrete. For semi-continuous (scaled) outputs, the outputs were split between positive and negative indications, with performance metrics recalculated. Using the Kellgren–Lawrence classification of OA as an exemplary case, scores ≤ 2 indicate no, doubtful, or mild evidence of the condition, hence they were aggregated as negative indications. Scores ≥ 3 describe moderate and severe conditions, and hence were aggregated as positive indications. For discrete (categorical) outputs, such as implant models, the numbers of true positives, true negatives, false positives and false negatives were summed across the respective classes.

Where multiple results were reported within the same research article, such as detecting joint space narrowing, osteophytes and subchondral cysts separately, the best- and worst-performing cases were included in the aggregated analysis as separate data points. For subset analyses, individual results were included where appropriate.

The QUADAS-2 framework, a tool used to assess the quality of diagnostic accuracy studies in systematic reviews, was used to evaluate the risk of bias in each study. The framework evaluates studies across four domains: patient selection, index test, reference standard and flow and timing, focusing on the risk of bias and applicability concerns [149].

RESULTS

Search results

5627 peer-reviewed studies were identified, of which 2912 were duplicates. Figure 1 highlights the PRISMA flow diagram. One hundred and thirty-nine studies were included in the systematic review after full-text screening [1–21, 23–29, 31–34, 36–44, 46–49, 51–54, 56, 58–62, 64–66, 68, 70–73, 75–77, 79–83, 86, 88–94, 96–104, 106–111, 113–118, 120–140, 142–147, 151, 152, 154, 156–165] (30 studies from cross-referencing) and 68 (43%) met the criteria for meta-analysis. Radiography ($n = 121$; 88%) and MRI ($n = 18$; 13%) were the primary imaging modalities. Almost all studies deployed convolutional neural networks (CNNs), most commonly off-the-shelf architectures, such as DenseNet (dense convolutional network), ResNet (residual network), EfficientNet, Inception, Inception-ResNet and VGG (visual geometry group) [153, 166]. Most studies used deep learning for disease classification or diagnosis ($n = 105$; 76%), followed by preoperative planning ($n = 12$; 8%) and progression prediction or prognosis ($n = 11$; 8%). The most common application was grading the severity of OA, with 85 studies (62%) targeting Kellgren–Lawrence gradings.

Quality of assessments

The quality assessment of the included studies using the QUADAS-2 tool revealed variability in the risk of bias across domains. For patient selection, 40% of studies were rated as 'low risk', and 35% were rated as 'high risk'. In the reference standard domain, 10% were rated as 'high risk'. Overall, across all domains, 57% of studies had a 'low risk' of bias, and 17% were rated as 'high risk'. These results highlight variability in methodological rigor, particularly in the patient selection domain, which showed the greatest proportion of high-risk studies.

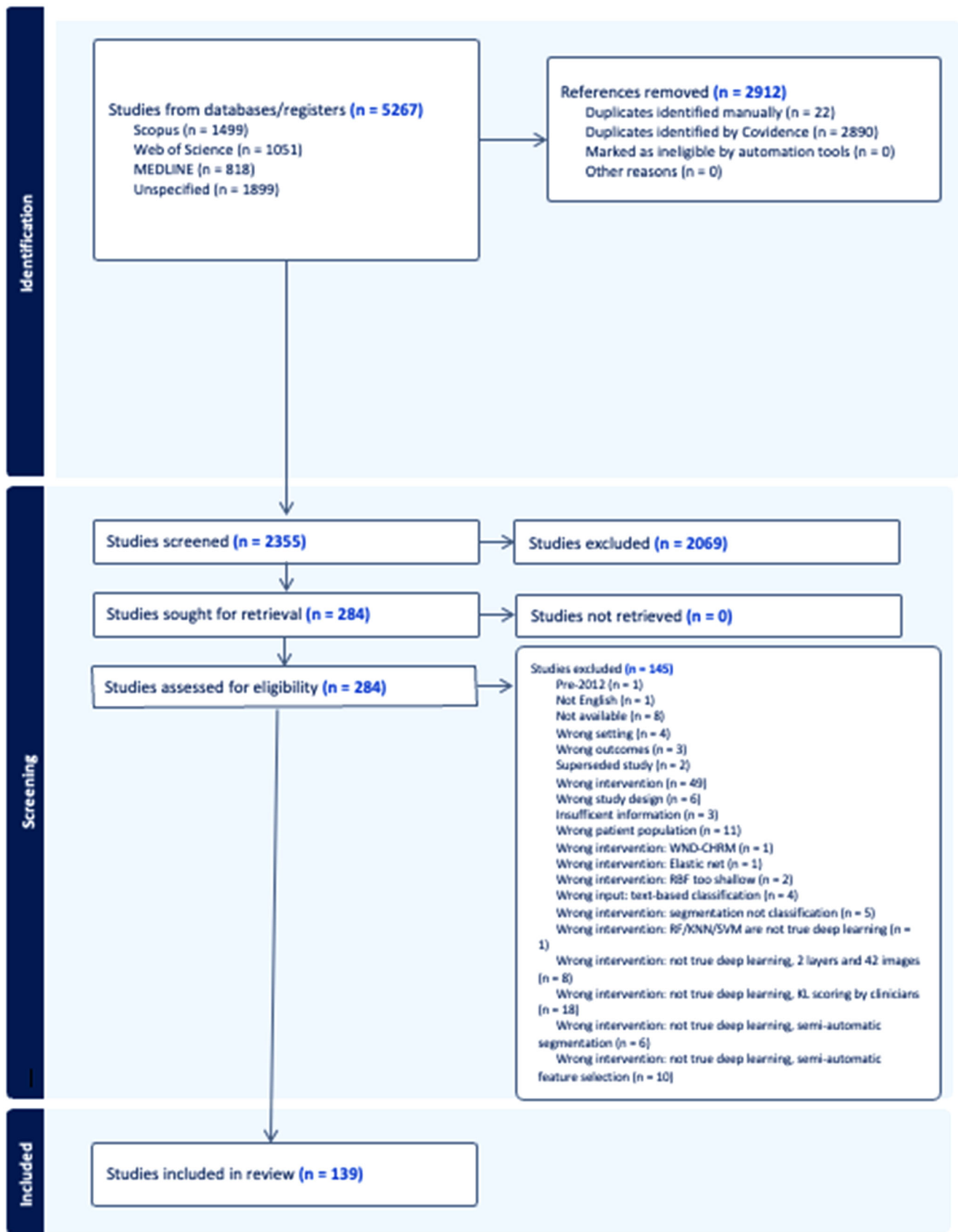


FIGURE 1 Preferred reporting items for systematic reviews and meta-analysis flow diagram.

Pooled analysis

A total of 66 studies met the criteria for meta-analysis of AI-interpretation performance, with the corresponding HSROC curve shown in Figure 2. The studies had a pooled sensitivity of 0.88 (95% CI: 0.81–0.92) and a pooled specificity of 0.91 (95% CI: 0.87–0.94). When assessing AI interpretation performance for diagnostic purposes, the 52 studies had a pooled sensitivity of 0.89 (95% CI: 0.84–0.93) and a pooled specificity of 0.90 (95% CI: 0.86–0.93). The 12 studies assessing AI interpretation performance for prognostic purposes showed a pooled sensitivity of 0.81 (95% CI: 0.43–0.96) and a pooled specificity of 0.93 (95% CI: 0.79–0.98). The HSROC curves for AI interpretation for diagnostic and prognostic purposes is shown in Figures 3 and 4 respectively.

A total of 55 studies investigated the performance of AI interpretation of radiographs (for both diagnostic and prognostic purposes), with the HSROC curve for these

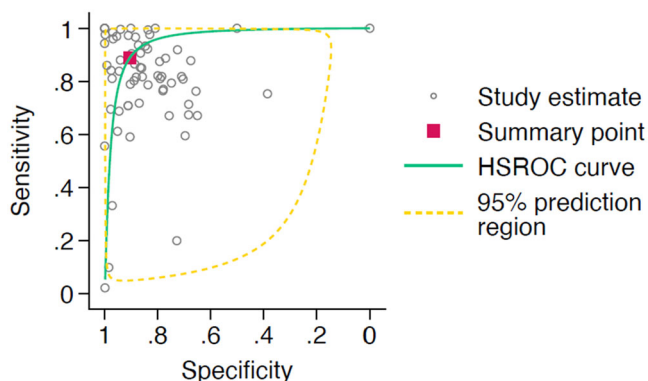


FIGURE 2 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for deep-learning only interpretations.

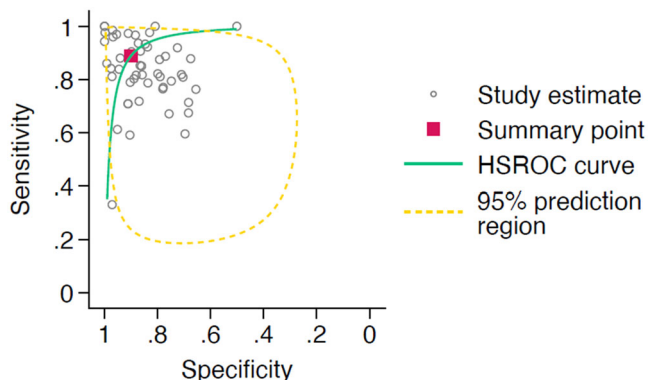


FIGURE 3 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for deep-learning only interpretations for diagnostic purposes.

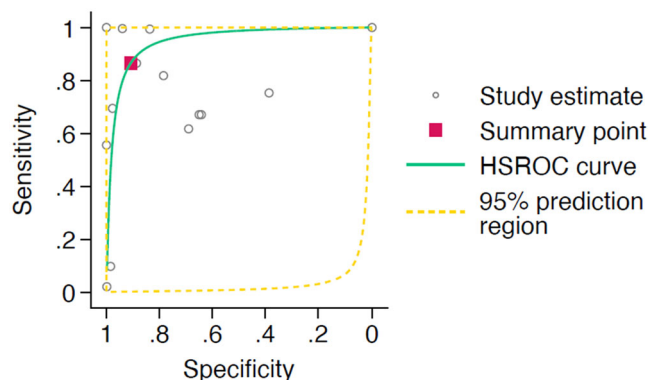


FIGURE 4 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for deep-learning only interpretations for prognostic purposes.

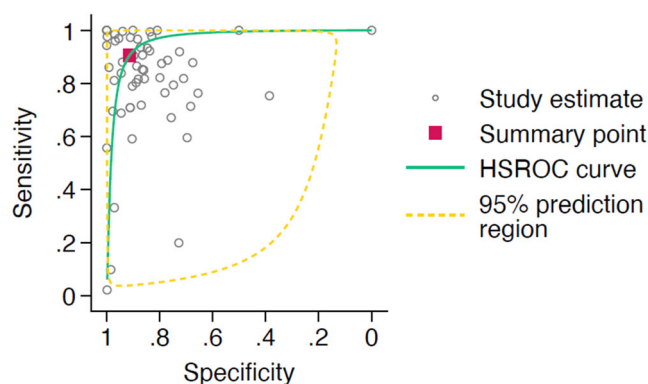


FIGURE 5 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for deep-learning only interpretations of radiographs.

studies shown in Figure 5. These studies had a pooled sensitivity of 0.89 (95% CI: 0.83–0.94) and a pooled specificity of 0.92 (95% CI: 0.88–0.95). Eleven studies investigated the performance of AI interpretation of cross-sectional imaging (for both diagnostic and prognostic purposes), 10 of which specifically looked at MRI interpretation. The studies had a pooled sensitivity of 0.78 (95% CI: 0.74–0.81) and a pooled specificity of 0.83 (95% CI: 0.72–0.90), with the corresponding HSROC shown in Figure 6.

The HSROC curve for the 11 studies that meet the criteria for meta-analysis of clinician-only interpretation performance is shown in Figure 7. Clinicians had a pooled sensitivity of 0.76 (95% CI: 0.64–0.85) and a pooled specificity of 0.79 (95% CI: 0.59–0.90).

In the meta-analysis of diagnostic accuracy, nine studies that made a direct comparison between AI performance and human clinicians alone were included. Among these 'matched' studies, the pooled sensitivity for AI was 0.87 (95% CI: 0.78–0.92), and the pooled specificity was 0.87 (95% CI: 0.76–0.93). In

comparison, human clinicians achieved a pooled sensitivity of 0.75 (95% CI: 0.62–0.84) and a pooled specificity of 0.73 (95% CI: 0.53–0.87). The HSROC curve for these studies is shown in Figure 8.

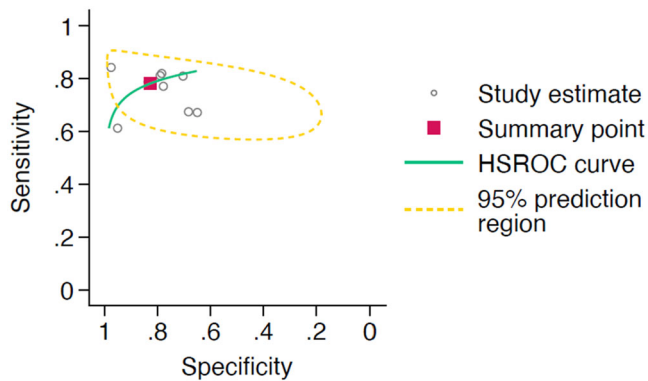


FIGURE 6 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for deep-learning only interpretations of cross-sectional imaging.

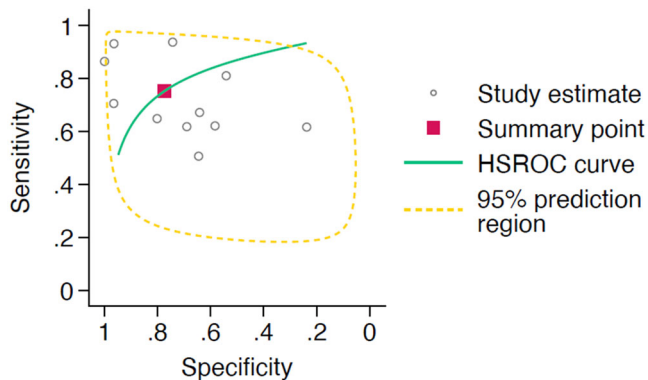


FIGURE 7 Hierarchical summary receiver operator characteristic (HSROC) curve of study estimated sensitivity plotted against specificity for clinician-only interpretations.

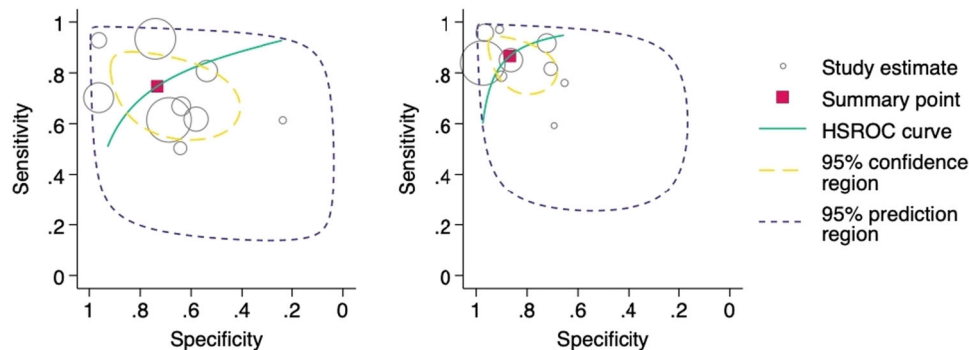


FIGURE 8 Hierarchical summary receiver operator characteristic (HSROC) curves of study estimated sensitivity plotted against specificity for the nine studies directly comparing artificial intelligence (figure on right) and clinician interpretations (figure on left).

DISCUSSION

A high pooled sensitivity and pooled specificity for AI interpretation were revealed in the meta-analysis, albeit with considerable overlap in sensitivity and specificity results for human interpretation. There was also significant overlap in pooled sensitivity and specificity results of AI results for both diagnostic and prognostic purposes. However, the included models showed higher pooled sensitivity and specificity values for AI interpretation of radiographs compared to cross-sectional imaging, with little-to-no overlap in sensitivity and specificity results.

This study's results are consistent with previous reviews evaluating deep learning in specific imaging tasks. Mohammadi et al. previously compared machine learning/AI with clinicians in OA detection across 27 studies, demonstrating a pooled sensitivity and specificity of 0.88 (95% CI: 0.86–0.91) and 0.81 (95% CI: 0.75–0.85) respectively [84]. In a scoping review of 12 studies, Gurung et al. also reported that AI-based tools diagnosed postoperative complications for total hip arthroplasty (THA) with an accuracy of 88.30% [44].

Unlike these two studies, which factored in all aspects of AI for single tasks, this study focused purely on deep learning classification (single task) algorithms and excluded hybrid models to minimise the potential for interpretive bias, thereby allowing for a more objective assessment of the diagnostic performance of deep learning in image interpretation. Despite this, there is variance in the sample size, patient demographics, standardisation of image acquisition, and specific deep learning algorithm being tested across each of the included studies, limiting the generalisability of the results.

A critical element in the evaluation of diagnostic accuracy is the choice of the gold standard against which performance is measured. In this meta-analysis, studies that used established diagnostic criteria, such as the human-derived Kellgren-Lawrence grading system, were included, however, these can be

associated with significant limitations. Owing to the broad nature of the meta-analysis, aggregating different deep learning algorithms throughout different stages of the clinical pathway, the large heterogeneity between studies was accepted. Conversely, with the over-representation of CNNs and OA detection in this dataset, as opposed to models such as vision transformers (ViTs), recurrent neural networks (RNNs), short-term memory networks (LSTMs), generative adversarial networks (GANs) and U-Net variants, caution would be advisable when generalising this study's findings to other architectures and tasks. Several studies in the literature have also highlighted the added value of segmentation and 3D reconstruction (CT, MRI) on DL performance when interpreting imaging. With the exclusion of acquisition, preprocessing and segmentation studies in this analysis, the results are likely skewed. Similarly, there is a skew towards knees, which limits generalisability to the other joints. Therefore, while the study findings broadly support the potential for AI to be adopted as an adjunct to clinical decision-making in the OA-arthroplasty clinical pathway, this study pools results from 138 different datasets across 138 different AI models, limiting the generalisability of the study results, and each application must still be properly validated for clinical use. None of the included models have been widely integrated into existing clinical workflow, limiting the current implications of whether these models truly translate into improved clinical decision-making for orthopaedic surgeons and radiologists [63].

Translating the growing body of AI research into meaningful clinical applications remains a work in progress. Facilitating the adoption of AI in healthcare, 10 guiding principles for good machine learning practice (GMLP) have been jointly developed by the United States, the United Kingdom and Canadian regulators. Similarly, new guidelines for planning and reporting clinical trials involving AI have been released by the SPIRIT-AI and CONSORT-AI [77, 81]. While challenges remain, this study's findings suggest that published AI models demonstrate a promising performance for diagnostic and prognostic purposes in the OA-arthroplasty pathway, warranting further studies to explore its potential for clinical use long term. Studies should handle the challenging ethical issues affiliated with integrating AI into clinical workflows, including confidentiality of patient information and liability in the setting of an incorrect interpretation.

By quantitatively evaluating AI interpretation performance and comparing it to human interpretation, this study provides insight into AI's capacity to improve clinical decision-making, particularly in grading disease severity, planning preoperative alignment and detecting postoperative complications. The findings underscore the need for ongoing validation and refinement of AI models to ensure generalisability across diverse

patient populations and imaging modalities. Ultimately, integrating AI into clinical workflows has the potential to optimise resource utilisation, reduce diagnostic variability and enhance patient outcomes, making it an invaluable adjunct to human expertise in musculoskeletal imaging.

CONCLUSIONS

This systematic review and meta-analysis present low-to-moderate quality evidence of strong performance of deep learning models in the OA-arthroplasty pathway.

The included algorithms showed a stronger performance when interpreting radiographs, and considerable overlap in sensitivity and specificity results when comparing diagnostic and prognostic interpretations. AI interpretation showed significant overlap in sensitivity and specificity results with human clinician interpretation, limiting generalisability of the results. The potential of deep learning extends beyond streamlining clinical pathways, with possible applications in refining surgical decision-making and aiding the detection of post-operative complications. Further validation of specific uses of specific deep learning models on different imaging modalities is required to understand its full impact on clinical practice.

AUTHOR CONTRIBUTIONS

Omar Musbahi: Conceptualisation; literature search and data analysis; drafting and critical revision. **Amal Ahmed:** Literature search and data analysis; drafting and critical revision. **Thomas Hall:** Conceptualisation; literature search and data analysis; drafting and critical revision. **Kamila Kutyla:** Literature search and data analysis; drafting and critical revision. **Muhayman Sadiq:** Literature search and data analysis; drafting and critical revision. **Justin Cobb:** Conceptualisation; drafting and critical revision. **Alex Bottle:** Conceptualisation; drafting and critical revision. **Richard Van Arkel:** Conceptualisation; drafting and critical revision. **Gareth Jones:** Conceptualisation; drafting and critical revision.

ACKNOWLEDGEMENTS

Omar Musbahi is sponsored by NIHR grant (NIHR ID 302632).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

This study analyses the results of 138 scholarly papers sourced from online journals and databases. Articles were accessed from both public domain resources and via institutional access through Imperial College London, the institution at which this review was

undertaken. Accessibility of these papers varies according to reader access privileges and different journal policies. For further details, please see the references section of the manuscript.

ETHICS STATEMENT

The authors have nothing to report.

CLINICAL TRIAL REGISTRATION

PROSPERO registration for systematic reviews—CRD42020198878.

ORCID

Omar Musbahi  <https://orcid.org/0000-0002-4095-0989>

Muhayman Sadiq  <https://orcid.org/0000-0002-2564-5871>

REFERENCES

1. Abdo AA, El-Tarhouni W, Abdulsalam AF, Altajori AB. Estimating the severity of knee osteoarthritis using deep convolutional neural network based on contrast limited adaptive histogram equalization technique. In: International conference on engineering & MIS. Istanbul, Turkey: IEEE; 2022. p. 1–6.
2. Abedin J, Antony J, McGuinness K, Moran K, O'Connor NE, Rebholz-Schuhmann D, et al. Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images. *Sci Rep*. 2019;9:5761.
3. Ahmed N, Saeed M, Aftab M, Mehmood A, Ilyas QM. Knee osteoarthritis detection and classification using transfer learning. In: 3rd International conference on computing and information technology (ICCIIT). Tabuk, Saudi Arabia: IEEE; 2023. pp. 365–9.
4. Ahmed SM, Mstafa RJ. Identifying severity grading of knee osteoarthritis from X-ray images using an efficient mixture of deep learning and machine learning models. *Diagnostics*. 2022;12(12):2939.
5. Akila SM, Imanov E, Almezghwi K. Analysis of knee osteoarthritis grading using deep learning. In: Aliev RA, Kacprzyk J, Pedrycz W, Jamshidi M, Babanli MB, Sadikoglu F. International conference on theory and applications of fuzzy system, and soft computing and artificial intelligence. Baku, Azerbaijan: IEEE; 2023. 435–43.
6. Almansour SHS, Singh R, Alyami SMH, Sharma N, Reshan MSA, Gupta S, et al. A convolution neural network design for knee osteoarthritis diagnosis using X-ray images. *Int J Online Biomed Eng*. 2023;19(7):125–41.
7. Almhdie-Imjabbar A, Nguyen K-L, Toumi H, Jennane R, Lespessailles E. Prediction of knee osteoarthritis progression using radiological descriptors obtained from bone texture analysis and Siamese neural networks: data from OAI and MOST cohorts. *Arthritis Res Ther*. 2022;24(1):66.
8. Al-Rimy BAS, Saeed F, Al-Sarem M, Albarrak AM, Qasem SN. An adaptive early stopping technique for DenseNet169-based knee osteoarthritis detection model. *Diagnostics (Basel)*. 2023;13(11):1903.
9. Alzaid A, Wignall A, Dogramadzi S, Pandit H, Xie SQ. Automatic detection and classification of peri-prosthetic femur fracture. *Int J Comp Assist Radiol Surg*. 2022;17(4):649–60.
10. Antonio PJA, Aldwayne B, Delmo J, Sevilla RV, Angelo D, Ligayo M, et al. Deep transfer network of knee osteoarthritis progression rate classification in MR imaging for medical imaging support system. In: International conference on decision aid sciences and applications, Chiangrai, Thailand. 2022. pp. 285–9.
11. Antony J, McGuinness K, Moran K, O'Connor NE. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In: Perner P, editor. *Lecture notes in computer science*. Cham, Switzerland: Springer; 2017. p. 376–90.
12. Arumugam SR, Balakrishna R, Rajeshram V, Gowr S, Karuppasamy SG, Premnath SP. Prediction of severity of knee osteoarthritis on X-ray images using deep learning. *Conference proceedings*. 2022.
13. Balasubramanian K, Kishore R, Krishnamoorthy GD. Optimal knee osteoarthritis diagnosis using hybrid deep belief network based on salp swarm optimization method. *Concurr Comput Pract Exp*. 2022;34(13):e6913.
14. Bany Muhammad M, Moinuddin A, Lee MTM, Zhang Y, Abedi V, Zand R, et al. Deep ensemble network for quantification and severity assessment of knee osteoarthritis. In: International conference on machine learning and applications. Boca Raton, United States: IEEE; 2019. pp. 951–7.
15. Bayramoglu N, Nieminen MT, Saarakkala S. A lightweight CNN and joint shape-joint space (JS²) descriptor for radiological osteoarthritis detection. In: Perner P, editor. *Lecture notes in computer science*. Springer; 2020. p. 331–45.
16. Bayramoglu N, Nieminen MT, Saarakkala S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. *Int J Med Inform*. 2022; 157:104627. <https://doi.org/10.1016/j.ijmedinf.2021.104627>
17. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network. *J Orthop Res*. 2020;38(7):1465–71.
18. Chang GH, Felson DT, Qiu S, Guermazi A, Capellini TD, Kolachalama VB. Assessment of knee pain from MR imaging using a convolutional Siamese network. *Eur Radiol*. 2020; 30(6):3538–48.
19. Chaugule S, Malemath VS. Osteoarthritis detection using densely connected neural network. *Conference proceedings*. 2022.
20. Chen C-C, Huang J-F, Lin W-C, Cheng C-T, Chen S-C, Fu C-Y, et al. The feasibility and performance of total hip replacement prediction deep learning algorithm with real world data. *Bioengineering*. 2023;10(4):458.
21. Chen C-C, Wu C-T, Chen CPC, Chung C-Y, Chen S-C, Lee MS, et al. Predicting the risk of total hip replacement by using a deep learning algorithm on plain pelvic radiographs: diagnostic study. *JMIR Format Res*. 2023;7:e42788. <https://doi.org/10.2196/42788>
22. Chen K, Stotter C, Klestil T, Nehrer S. Artificial intelligence in orthopedic radiography analysis: a narrative review. *Diagnostics*. 2022;12(9):2235.
23. Chen N, Feng Z, Li F, Wang H, Yu R, Jiang J, et al. A fully automatic target detection and quantification strategy based on object detection convolutional neural network YOLOv3 for one-step X-ray image grading. *Anal Methods*. 2023;15(2): 164–70.
24. Chen P, Gao L, Shi X, Allen K, Yang L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Comput Med Imaging Graph*. 2019;75: 84–92.
25. Cueva JH, Castillo D, Espinós-Morató H, Durán D, Díaz P, Lakshminarayanan V. Detection and classification of knee osteoarthritis. *Diagnostics*. 2022;12(10):2362.
26. Daneshmand M, Panfilov E, Bayramoglu N, Korhonen RK, Saarakkala S. Deep learning based detection of osteophytes in radiographs and magnetic resonance imagings of the knee using 2D and 3D morphology. *J Orthop Res*. 2024;42(7): 1473–81.

27. Deniz C, Segal N, Sodha V, Lynch J, Nevitt M, Felson D, et al. Deep learning analysis of baseline weightbearing computed tomography for predicting radiographic knee osteoarthritis progression. *Osteoarth Cartil*. 2022;30:S300.
28. Dharmani BC, Khatri K. Deep learning for knee osteoarthritis severity stage detection using X-ray images. Conference proceedings. 2023.
29. Domingues JG, Araujo DC, Costa-Silva L, Machado AMC, Machado LAC, Veloso AA, et al. Development of a convolutional neural network for diagnosing osteoarthritis, trained with knee radiographs from the ELSA-Brasil musculoskeletal. *Radiol Bras*. 2023;56(5):248–54.
30. Duong V, Oo WM, Ding C, Culvenor AG, Hunter DJ. Evaluation and treatment of knee pain: a review. *JAMA*. 2023; 330(16):1568–80.
31. E T, Wang C, Cui Y, Nai R, Zhang Y, Zhang X, et al. Automatic diagnosis and grading of patellofemoral osteoarthritis from the axial radiographic view: a deep learning-based approach. *Acta Radiol*. 2023;64(2):658–65.
32. En CZ, Swee TT. Computer-aided knee osteoarthritis classification system using artificial neural network (ANN). *J Med Imaging Health Inform*. 2013;3(4):561–5.
33. Farajzadeh N, Sadeghzadeh N, Hashemzadeh M. IJES-OA Net: a residual neural network to classify knee osteoarthritis from radiographic images based on the edges of the intra-joint spaces. *Med Eng Phys*. 2023;113:103957.
34. Fatema K, Rony MAH, Azam S, Mukta MSH, Karim A, Hasan MZ, et al. Development of an automated optimal distance feature-based decision system for diagnosing knee osteoarthritis using segmented X-ray images. *Heliyon*. 2023; 9(11):e21703. <https://doi.org/10.1016/j.heliyon.2023.e21703>
35. FDA. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. 2024. Available from: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>. Accessed 19 Jan 2025.
36. Feng Y, Liu J, Zhang H, Qiu D. Automated grading of knee osteoarthritis X-ray images based on attention mechanism. In: International conference on bioinformatics and biomedicine (BIBM). Houston, United States: IEEE; 2021. pp. 3791–6.
37. Gebre RK, Hirvasniemi J, van der Heijden RA, Lantto I, Saarakkala S, Leppilahti J, et al. Detecting hip osteoarthritis on clinical CT: a deep learning application based on 2-D summation images derived from CT. *Osteoporos Int*. 2022;33(2): 355–65.
38. Geng EA, Cho BH, Valliani AA, Arvind V, Patel AV, Cho SK, et al. Development of a machine learning algorithm to identify total and reverse shoulder arthroplasty implants from X-ray images. *J Orthop*. 2023;35:74–8.
39. Guan B, Liu F, Haj-Mirzaian A, Demehri S, Samsonov A, Neogi T, et al. Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-month follow-up period. *Osteoarth Cartil*. 2020; 28(4):428–37.
40. Guan B, Liu F, Matthew P, Mirzaian AH, Demehri S, Neogi T, et al. Deep learning approach to predict pain progression in knee osteoarthritis. *Osteoarth Cartil*. 2020;28:S316.
41. Guan B, Liu F, Mizaian AH, Demehri S, Samsonov A, Guermazi A, et al. Deep learning approach to predict pain progression in knee osteoarthritis. *Skeletal Radiol*. 2022;51(2): 363–73.
42. Guarnera F, Rondinella A, Giudice O, Ortis A, Battiato S, Rundo F, et al. Early detection of hip periprosthetic joint infections through CNN on computed tomography images. In: Perner P. Lecture notes in computer science. Cham, Switzerland: Springer; 2023. p. 134–43.
43. Guo S, Zhang J, Li H, Zhang J, Cheng C-K. A multi-branch network to detect post-operative complications following hip arthroplasty on X-ray images. *Front Bioeng Biotechnol*. 2023; 11:1239637. <https://doi.org/10.3389/fbioe.2023.1239637>
44. Gurung B, Liu P, Harris PDR, Sagi A, Field RE, Sochart DH, et al. Artificial intelligence for image analysis in total hip and total knee arthroplasty: a scoping review. *Bone Jt J*. 2022; 104-B:929–37.
45. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. *Mach Learn*. 2021;110(3):451–6.
46. Harish H, Patrot A, Bhavan S, Gousiya S, Livitha A. Knee osteoarthritis prediction using deep learning. In: International conference on recent advances in information technology for sustainable development (ICRAIS). Maipal, India: IEEE; 2023. pp. 14–9.
47. Haseeb A, Attique Khan M, Shehzad F, Alhaisoni M, Ali Khan J, Kim T, et al. Knee osteoarthritis classification using X-ray images based on optimal deep neural network. *Comp Syst Sci Eng*. 2023;47(2):2397–415.
48. Hema Rajini N, Anton Smith A. Osteoarthritis detection and classification in knee X-ray images using particle swarm optimization with deep neural network. In: Kose U. editor. Lecture notes in mechanical engineering. Cham, Switzerland: Springer; 2023. p. 91–101.
49. Henson P, Assad N, Cho B, Schwartz JT, Geng E, Arvind V, et al. Development of a machine learning model to identify elbow implants from radiographic images. *Semin Arthropl*. 2023;33(2):255–60.
50. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022; 12(1):5979.
51. Hill BG, Byrum T, Zhou A, Schilling PL. An algorithmic approach to understanding osteoarthritic knee pain. *JB JS Open Access*. 2023;8(4):e23.00039ER.
52. Hu J, Zheng C, Yu Q, Zhong L, Yu K, Chen Y, et al. DeepKOA: a deep-learning model for predicting progression in knee osteoarthritis using multimodal magnetic resonance images from the osteoarthritis initiative. *Quant Imaging Med Surg*. 2023;13(8):4852–66.
53. Hu K, Wu W, Li W, Simic M, Zomaya A, Wang Z. Adversarial evolving neural network for longitudinal knee osteoarthritis prediction. *IEEE Trans Med Imaging*. 2022;41(11):3207–17.
54. Hu Y, Tang J, Zhao S, Li Y. Deep learning-based multimodal 3 T MRI for the diagnosis of knee osteoarthritis. *Comput Math Methods Med*. 2022;2022:7643487. <https://doi.org/10.1155/2022/7643487>
55. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *Lancet*. 2019; 393(10182):1745–59.
56. Jafarzadeh S, Felson DT, Nevitt MC, Torner JC, Lewis CE, Roemer FW, et al. Use of clinical and imaging features of osteoarthritis to predict knee replacement in persons with and without radiographic osteoarthritis: the MOST study. *Osteoarth Cartil*. 2020;28:S308–9.
57. Jarraya M, Roemer FW, Bäuerle T, Kogan F, Guermazi A. PET imaging in osteoarthritis. *PET Clin*. 2023;18(1):21–9.
58. Johnny Samuel S, Bagewadi N, Malathy C, Balasaraswathi VR, Gayathri M, Batta V, et al. Identification of knee prostheses from lateral radiographs using deep learning techniques. In: 9th world congress on electrical engineering and computer systems and science. London, United Kingdom: Avestia; 2023. pp. 1–7.
59. Kang Y-J, Yoo J-I, Cha Y-H, Park CH, Kim J-T. Machine-learning-based identification of hip arthroplasty designs. *J Orthop Transl*. 2020;21:13–7.
60. Karnuta JM, Murphy MP, Luu BC, Ryan MJ, Haeberle HS, Brown NM, et al. Artificial intelligence for automated implant identification in total hip arthroplasty: a multicenter external validation study exceeding two million plain radiographs. *J Arthroplasty*. 2023;38(10):1998–2003.e1.

61. Karnuta JM, Shaikh HJF, Murphy MP, Brown NM, Pearle AD, Nawabi DH, et al. Artificial intelligence for automated implant identification in knee arthroplasty: a multicenter external validation study exceeding 3.5 million plain radiographs. *J Arthroplasty*. 2023;38(10):2004–8.
62. Khosravi B, Rouzrokh P, Maradit Kremers H, Larson DR, Johnson QJ, Faghani S, et al. Patient-specific hip arthroplasty dislocation risk calculator: an explainable multimodal machine learning-based approach. *Radiol Artif Intel*. 2022;4(6):e220067. <https://doi.org/10.1148/ryai.220067>
63. Kijowski R, Fritz J, Deniz CM. Deep learning applications in osteoarthritis imaging. *Skelet Radiol*. 2023;52(11):2225–38.
64. Kim B, Lee DW, Lee S, Ko S, Jo C, Park J, et al. Automated detection of surgical implants on plain knee radiographs using a deep learning algorithm. *Medicina*. 2022;58(11):1677.
65. Kim M-S, Cho R-K, Yang S-C, Hur J-H, In Y. Machine learning for detecting total knee arthroplasty implant loosening on plain radiographs. *Bioengineering*. 2023;10(6):632.
66. Klemm C, Uzosike AC, Cohen-Levy WB, Harvey MJ, Subih MA, Kwon Y-M. The ability of deep learning models to identify total hip and knee arthroplasty implant design from plain radiographs. *J Am Acad Orthop Surg*. 2022;30(9):409–15.
67. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097–105.
68. Kunze KN, Jang SJ, Li TY, Pareek A, Finocchiaro A, Fu MC, et al. Artificial intelligence for automated identification of total shoulder arthroplasty implants. *J Shoulder Elbow Surg*. 2023;32(10):2115–22.
69. Lalehzarian SP, Gowd AK, Liu JN. Machine learning in orthopaedic surgery. *World J Orthop*. 2021;12(9):685–99.
70. Lau LCM, Chui ECS, Man GCW, Xin Y, Ho KKW, Mak KKK, et al. A novel image-based machine learning model with superior accuracy and predictability for knee arthroplasty loosening detection and clinical decision making. *J Orthop Transl*. 2022;36:177–83.
71. Lee JJ, Liu F, Majumdar S, Pedoia V. Can AI predict pain progression in knee osteoarthritis subjects from structural MRI. *Osteoarth Cartil*. 2019;27:S24.
72. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. *Radiology*. 2020;296(3):584–93.
73. Li N, Swiecicki A, Said N, O'Donnell J, Jiranek WA, Mazurowski MA. Automatic Kellgren–Lawrence grade estimation driven deep learning algorithms. In: *Computer aided diagnosis*. Houston, United States: SPIE; 2020. p. 11314.
74. Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS. Artificial intelligence in orthopaedic surgery. *Bone Jt Res*. 2023;12(7):447–54.
75. Liu B, Luo J, Huang H. Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. *Int J Comp Assist Radiol Surg*. 2020;15(3):457–66.
76. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*. 2018;289(1):160–9.
77. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–74.
78. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271–97.
79. Mahum R, Rehman SU, Meraj T, Rauf HT, Irtaza A, El-Sherbeeny AM, et al. A novel hybrid approach based on deep CNN features to detect knee osteoarthritis. *Sensors*. 2021;21(18):6189.
80. Masood H, Hassan E, Salam AA, Liaquat M. Osteo-Doc: KL-grading of osteoarthritis using deep-learning. In: *2nd international conference on digital futures and transformative technologies*. Rawalpindi, Pakistan: IEEE; 2022. pp. 1–6.
81. MHRA. Good machine learning practice for medical device development: guiding principles. 2021. Available from: https://assets.publishing.service.gov.uk/media/617921168fa8f52978e14c0b/GMLP_Guiding_Principles_FINAL.pdf. Accessed 2 Feb 2025.
82. Mikhaylichenko A, Demyanenko Y. Automatic grading of knee osteoarthritis from plain radiographs using densely connected convolutional networks. In: *Perner P. editor. Lecture notes in computer science*. Cham, Switzerland: Springer; 2021. p. 149–61.
83. Mishra A, Ramanathan A, Batta V, Malathy C, Kundu SS, Gayathri M, et al. Harnessing the potential of deep learning for total shoulder implant classification: a comparative study. In: *Perner, P. editor. Lecture notes in computer science*. Cham, Switzerland: Springer; 2024. p. 119–32.
84. Mohammadi S, Salehi MA, Jahanshahi A. Artificial intelligence in osteoarthritis detection: a systematic review and meta analysis. *Osteoarthr Cartil*. 2024; 32(3):241–53.
85. Mohammed AS, Hasanaath AA, Latif G, Bashar A. Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed X-ray images. *Diagnostics*. 2023;13(8):1380.
86. Morales Martinez AG, Flament I, Liu F, Lee J, Cao P, Majumdar S, et al. Discovering knee osteoarthritis bone shape features using deep learning. *Osteoarth Cartil*. 2019;27:S386–7.
87. Müskens J, Kool RB, van Dulmen SA, Westert GP. Overuse of diagnostic testing in healthcare: a systematic review. *BMJ Qual Saf*. 2022;31(1):54–63.
88. Nasser Y, El Hassouni M, Hans D, Jennane R. A discriminative shape-texture convolutional neural network for early diagnosis of knee osteoarthritis from X-ray images. *Phys Eng Sci Med*. 2023;46(2):827–37.
89. Nguyen HH, Saarakkala S, Tiulpin A. Automatic knee osteoarthritis severity assessment from plain radiographs with limited data. *Osteoarth Cartil*. 2020;28:S314–5.
90. Nguyen HH, Saarakkala S, Tiulpin A. Deep semi-supervised learning for knee osteoarthritis severity assessment from plain radiographs. *Osteoarth Cartil*. 2020;28:S311–2.
91. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging*. 2019;32(3):471–7.
92. Norman BD, Pedoia V, Noworolski A, Link TM, Majumdar S. Automatic knee Kellgren–Lawrence grading with artificial intelligence. *Osteoarth Cartil*. 2018;26:S436–7.
93. Olsson S, Akbarian E, Lind A, Razavian AS, Gordon M. Automating classification of osteoarthritis according to Kellgren–Lawrence in the knee using deep learning in an unfiltered adult population. *BMC Musculoskelet Disord*. 2021;22(1):844.
94. Panfilov E, Tiulpin A, Saarakkala S, Nieminen MT, Casula V. Automatic assessment of morphological changes in knee cartilage using deep learning. *Osteoarth Cartil*. 2020;28:S309–10.
95. Patel I, Nham F, Zalikha AK, El-Othmani MM. Epidemiology of total hip arthroplasty: demographics, comorbidities and outcomes. *Arthroplasty*. 2023;5(1):2.
96. Pedoia V, Lee J, Norman B, Link TM, Majumdar S. Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire osteoarthritis initiative baseline cohort. *Osteoarth Cartil*. 2019;27(7):1002–10.

97. Phan Trung H, Nguyen Thiet S, Nguyen Trung T, Le Tan L, Tran Minh T, Quan Thanh T. OsteoGA: an explainable AI framework for knee osteoarthritis severity assessment. Conference proceedings. 2023.
98. Pi S-W, Lee B-D, Lee MS, Lee HJ. Ensemble deep-learning networks for automated osteoarthritis grading in knee X-ray images. *Sci Rep*. 2023;13(1):22887.
99. Pongsakonpruttikul N, Anghong C, Kittichai V, Chuwongin S, Puengpipattrakul P, Thongpat P, et al. Artificial intelligence assistance in radiographic detection and classification of knee osteoarthritis and its severity: a cross-sectional diagnostic study. *Eur Rev Med Pharmacol Sci*. 2022;26(5):1549–58.
100. Rahman T, Khandakar A, Islam KR, Soliman MM, Islam MT, Elsayed A, et al. HipXNet: deep learning approaches to detect aseptic loosening of hip implants using X-ray images. *IEEE Access*. 2022;10:53359–73.
101. Rehman A, Raza A, Alamri FS, Alghofaily B, Saba T. Transfer learning-based smart features engineering for osteoarthritis diagnosis from knee X-ray images. *IEEE Access*. 2023;11:71326–38.
102. Ribas LC, Riad R, Jennane R, Bruno OM. A complex network based approach for knee osteoarthritis detection: data from the osteoarthritis initiative. *Biomed Signal Process Control*. 2022;71:103133.
103. Rouzrokh P, Mickley JP, Khosravi B, Faghani S, Moassefi M, Schulz WR, et al. THA-AID: deep learning tool for total hip arthroplasty automatic implant detection with uncertainty and outlier quantification. *J Arthroplasty*. 2024;39(4):966–73.e17.
104. Rouzrokh P, Ramazanian T, Wyles CC, Philbrick KA, Cai JC, Taunton MJ, et al. Deep learning artificial intelligence model for assessment of hip dislocation risk following primary total hip arthroplasty from postoperative radiographs. *J Arthroplasty*. 2021;36(6):2197–203.e3.
105. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865–84.
106. Saini D, Khosla A, Chand T, Chouhan DK, Prakash M. A deep learning approach for development of web application for automated knee OA classification. Conference proceedings. 2023.
107. Schwartz AJ, Clarke HD, Spangehl MJ, Bingham JS, Etzioni DA, Neville MR. Can a convolutional neural network classify knee osteoarthritis on plain radiographs as accurately as fellowship-trained knee arthroplasty surgeons? *J Arthroplasty*. 2020;35(9):2423–8.
108. Sekhri A, Kerkouri MA, Chetouani A, Tliba M, Nasser Y, Jennane R, et al. Automatic diagnosis of knee osteoarthritis severity using Swin transformer. In: 20th International Conference on Content Based Multimedia Indexing. Orleans, France: ACM DL Digital Library; 2023. pp. 1–5.
109. Shah RF, Bini SA, Martinez AM, Pedroia V, Vail TP. Incremental inputs improve the automated detection of implant loosening using machine-learning algorithms. *Bone Jt J*. 2020;102–B(6A):101–6.
110. Sharma G, Anand V, Kumar V. Classification of osteo-arthritis with the help of deep learning and transfer learning. In: 5th international conference on inventive research in computing applications. Coimbatore, India IEEE; 2023. pp. 446–52.
111. Sharma S, Batta V, Chidambaramathan M, Mathialagan P, Mani G, Kiruthika M, et al. Knee implant identification by fine-tuning deep learning models. *Indian J Orthop*. 2021;55(5):1295–305.
112. Shichman I, Roof M, Askew N, Nherera L, Rozell JC, Seyler TM, et al. Projections and epidemiology of primary hip and knee arthroplasty in Medicare patients to 2040–2060. *JB JS Open Access*. 2023;8(1):e22.00112.
113. Sindhu N, Mishra S, Gowrishankar S, Anushka S, Ha S, Veena A. Prediction of knee-replacement using deep-learning approach. In: International conference on edge computing and applications. Tamilnadu, India: IEEE; 2022. pp. 1213–8.
114. Sivakumari T, Vani R. Implementation of AlexNet for classification of knee osteoarthritis. In: 7th International conference on communication and electronics systems. Coimbatore, India: IEEE; 2022. pp. 1405–9.
115. Sivakumari T, Vani R. Deep learning-based automated knee joint localization in radiographic images using Faster R-CNN. *Curr Med Imaging*. 2024;20:e15734056262464.
116. Suresha S, Kidziński L, Halilaj E, Gold GE, Delp SL. Automated staging of knee osteoarthritis severity using deep neural networks. *Osteoarth Cartil*. 2018;26:S441.
117. Swiecicki A, Li N, O'Donnell J, Said N, Yang J, Mather RC, et al. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. *Comput Biol Med*. 2021;133:104334. <https://doi.org/10.1016/j.compbiomed.2021.104334>
118. Tariq T, Suhail Z, Nawaz Z. Knee osteoarthritis detection and classification using X-rays. *IEEE Access*. 2023;11:48292–303.
119. The Institute for Health Metrics and Evaluation. Osteoarthritis – Level 3 cause. *The Lancet*. 2019. https://www.healthdata.org/results/gbd_summaries/2019/osteoarthritis-level-3-cause. Accessed 3 Jan 2025.
120. Thomas KA, Kidziński Ł, Halilaj E, Fleming SL, Venkataraman GR, Oei EHG, et al. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiol Artif Intel*. 2020;2(2):190065.
121. Tibrewala R, Ozhinsky E, Shah R, Flament I, Crossley K, Srinivasan R, et al. Computer-aided detection AI reduces interreader variability in grading hip abnormalities with MRI. *J Magn Reson Imaging*. 2020;52(4):1163–72.
122. Tibrewala R, Ozhinsky E, Shah R, Foreman SC, Pedroia V, Majumdar S. Detecting hip osteoarthritic degenerative changes in MRI using deep learning. *Osteoarth Cartil*. 2019;27:S387–8.
123. Tiulpin A, Klein S, Bierma-Zeinstra S, Thevenot J, van Meurs J, Oei E, et al. Deep learning predicts knee osteoarthritis progression from plain radiographs. *Osteoarth Cartil*. 2019;27:S397–8.
124. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, van Meurs J, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci Rep*. 2019;9:20038. <https://doi.org/10.1038/s41598-019-56527-3>
125. Tiulpin A, Panfilov E, Vaattovaara E, Niinimäki J, Nevalainen M, Saarakkala S. DeepKnee: towards open source fully-automatic Kellgren–Lawrence grading. *Osteoarth Cartil*. 2019;27:S25.
126. Tiulpin A, Saarakkala S. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *Osteoarthr Cartil*. 2020;28:S308.
127. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci Rep*. 2018;8:1727.
128. Tiwari A, Poduval M, Bagaria V. Evaluation of artificial intelligence models for osteoarthritis of the knee using deep learning algorithms for orthopedic radiographs. *World J Orthop*. 2022;13(6):603–14.
129. Tiwari A, Yadav AK, Bagaria V. Application of deep learning algorithm in automated identification of knee arthroplasty implants from plain radiographs using transfer learning models: are algorithms better than humans? *J Orthop*. 2022;32:139–45.
130. Tolpadi AA, Lee JJ, Pedroia V, Majumdar S. Deep learning predicts total knee replacement from magnetic resonance images. *Sci Rep*. 2020;10(1):6371.

131. Upadhyay A, Sawant O, Choudhary P. Detection of knee osteoarthritis stages using convolutional neural network. *SN Comp Sci.* 2023;4(3):257.
132. Urban G, Porhemmat S, Stark M, Feeley B, Okada K, Baldi P. Classifying shoulder implants in X-ray images using deep learning. *Comput Struct Biotechnol J.* 2020;18:967–72.
133. Üreten K, Arslan T, Gültekin KE, Demir AND, Özer HF, Bilgili Y. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. *Skeletal Radiol.* 2020;49(9):1369–74.
134. Vijaya Kishore V, Kalpana V, Hemanth Kumar G. Evaluating the efficacy of deep learning models for knee osteoarthritis prediction based on Kellgren–Lawrence grading system. *e-Prime Adv Electr Eng Electron Energy.* 2023;5:100266.
135. Vijaya Kishore V, Batthala S, Chamarthi JV, Achyutasai C, Subrahmanyam B. Knee osteoarthritis prediction driven by deep learning and the Kellgren–Lawrence grading. *Proc Eng Sci.* 2023;5(3):475–84.
136. von Schacky CE, Liu F, Foreman SC, Jungmann PM, Nardo L, Nevitt MC, et al. Automated severity grading of radiographic hip osteoarthritis features with deep learning. *Osteoarth Cartil.* 2019;27:S396–7.
137. von Schacky CE, Liu F, Ozhinsky E, Jungmann PM, Nardo L, Foreman SC, et al. Artificial intelligence to grade hip osteoarthritis features on radiographs. *Semin Musculoskelet Radiol.* 2019;23(Suppl 2).
138. von Schacky CE, Sohn J, Foreman SC, Liu F, Jungmann PM, Nardo L, et al. Development and performance comparison with radiologists of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Osteoarth Cartil.* 2020;28:S306–8.
139. von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, et al. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Radiology.* 2020;295(1):136–45.
140. Wahyuningrum RT, Anifah L, Purnama IKE, Purnomo MH. A new approach to classify knee osteoarthritis severity from radiographic images based on CNN-LSTM method. 2019 IEEE 10th international conference on awareness science and technology (iCAST), IEEE. Morioka, Japan; 2019. p. 528–33.
141. Wang J, Hall TAG, Musbahi O, Jones GG, van Arkel RJ. Predicting hip-knee-ankle and femorotibial angles from knee radiographs with deep learning. *Knee.* 2023;42:281–8.
142. Wang K, Niu X, Dou Y, Xie D, Yang T. A siamese network with adaptive gated feature fusion for individual knee OA features grades prediction. *Sci Rep.* 2021;11(1):16833.
143. Wang K, Niu X, Dou Y, Yang D, Xie D, Yang T. Two-stage convolutional neural network for knee osteoarthritis diagnosis in X-rays. In: Su R, Zhang YD, Liu H, editor. *International Conference on Medical Imaging and Computer-Aided Diagnosis.* Springer; 2022. pp. 205–17.
144. Wang X, Liu S, Zhou CC. Detection algorithm of knee osteoarthritis based on magnetic resonance images. *Intel Automat Soft Comput.* 2023;37(1):221–34.
145. Wang Y, Li S, Zhao B, Zhang J, Yang Y, Li B. A ResNet-based approach for accurate radiographic diagnosis of knee osteoarthritis. *CAAI Trans Intel Technol.* 2022;7(3):512–21.
146. Wang Z, Chetouani A, Hans D, Lespessailles E, Jennane R. Siamese-gap network for early detection of knee osteoarthritis. In: *International conference on medical imaging and computer-aided diagnosis.* Birmingham, United Kingdom: IEEE; 2022. pp. 205–17.
147. Wei J, Li D, Sing DC, Yang J, Beeram I, Puvanesarajah V, et al. Detecting total hip arthroplasty dislocations using deep learning: clinical and Internet validation. *Emerg Radiol.* 2022; 29(5):801–8.
148. Weinstein AM, Rome BN, Reichmann WM, Collins JE, Burbine SA, Thornhill TS, et al. Estimating the burden of total knee replacement in the United States. *J Bone Jt Surg Am.* 2013;95(5):385–92.
149. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–36.
150. Wilhelm NJ, von Schacky CE, Lindner FJ, Feucht MJ, Ehmann Y, Pogorzelski J, et al. Multicentric development and validation of a multi-scale and multi-task deep learning model for comprehensive lower extremity alignment analysis. *Artif Intel Med.* 2024;150:102843. <https://doi.org/10.1016/j.artmed.2024.102843>
151. Xu Y, Xiong H, Liu W, Liu H, Guo J, Wang W, et al. Development and validation of a deep-learning model to predict total hip replacement on radiographs: the total hip replacement prediction (THREP) model. *J Bone Jt Surg Am.* 2023;106(5):389–96.
152. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLoS One.* 2017;12(6):e0178992.
153. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* 2018;9(4):611–29.
154. Yan S, Ramazanian T, Sagheb E, Fu S, Sohn S, Lewallen DG, et al. DeepTKAClassifier: brand classification of total knee arthroplasty implants using explainable deep convolutional neural networks. *Conference proceedings.* 2020.
155. Yang J, Ji Q, Ni M, Zhang G, Wang Y. Automatic assessment of knee osteoarthritis severity in portable devices based on deep learning. *J Orthop Surg.* 2022;17(1):540.
156. Yassine N, El Hassouni M, Jennane R. Discriminative deep neural network for predicting knee osteoarthritis in early stage. In: Perner P, editor. *Lecture notes in computer science.* Cham, Switzerland: Springer; 2022. p. 126–36.
157. Yeoh PSQ, Lai KW, Goh SL, Hasikin K, Wu X, Li P. Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: data from the osteoarthritis initiative. *Front Bioeng Biotechnol.* 2023;11:1164655. <https://doi.org/10.3389/fbioe.2023.1164655>
158. Yi PH, Kim TK, Wei J, Li X, Hager GD, Sair HI, et al. Automated detection and classification of shoulder arthroplasty models using deep learning. *Skeletal Radiol.* 2020;49(10):1623–32.
159. Yi PH, Wei J, Kim TK, Sair HI, Hui FK, Hager GD, et al. Automated detection & classification of knee arthroplasty using deep learning. *Knee.* 2020;27(2):535–42.
160. Yin R, Chen H, Tao T, Zhang K, Yang G, Shi F, et al. Expanding from unilateral to bilateral: a robust deep learning-based approach for predicting radiographic osteoarthritis progression. *Osteoarth Cartil.* 2024;32(3):338–47.
161. Yoon JS, Yon C-J, Lee D, Lee JJ, Kang CH, Kang S-B, et al. Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of radiographic knee osteoarthritis. *BMC Musculoskelet Disord.* 2023;24(1):869.
162. Yue Y, Wang X, Zhao M, Tian H, Cao Z, Gao Q, et al. Pre-operative prediction of prosthetic size in total knee arthroplasty based on multimodal data and deep learning. In: *5th International Conference on Computer and Communications.* Chengdu, China: IEEE; 2019. pp. 2077–81.
163. Yuniarno EM, Purnomo MH. Classification anterior and posterior of knee osteoarthritis X-ray images grade KL-2 using deep learning with random brightness augmentation. In: *International conference on computer engineering, network, and intelligent multimedia.* Surabaya, Indonesia: IEEE; 2022. pp. 128–32.
164. Yunus U, Amin U, Sharif M, Yasmin M, Kadry S, Krishnamoorthy S. Recognition of knee osteoarthritis (KOA) using YOLOv2 and classification based on convolutional neural network. *Life.* 2022;12(8):1126.
165. Zhang B, Tan J, Cho K, Chang G, Deniz CM. Attention-based CNN for KL grade classification: data from the osteoarthritis

initiative. In: International symposium on biomedical imaging. Iowa City, United States: IEEE; 2020. pp. 731–5.

166. Zhou T, Ye X, Lu H, Zheng X, Qiu S, Liu Y. Dense convolutional network and its application in medical image analysis. *BioMed Res Int*. 2022;2022:2384830. <https://doi.org/10.1155/2022/2384830>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Musbahi O, Ahmed A, Hall T, Kutyla K, Sadiq M, Cobb J, et al. Deep learning classification models demonstrate high accuracy and clinical potential in radiograph interpretation in the arthroplasty clinical pathway: a systematic review and meta-analysis. *J Exp Orthop*. 2025;12:e70342. <https://doi.org/10.1002/jeo2.70342>