

Artificial Intelligence for Rheumatoid Arthritis and Osteoporosis Diagnosis from Radiography, Computed Tomography, and Multimodal Clinical Data: A Systematic Review

Moulay Youssef ICHAHANE^{1*}, Mohamed LACHGAR², Abdelouafi IKIDID³,
Ahmed DOURHRI⁴, Mohamed EL GHAZOUANI⁵, Nouredine ASSAD⁶

Chouaib Doukkali University, LTI/SI / CED Technology and Engineering, ENSA, El Jadida, Morocco^{1,4,6}

Cadi Ayyad University, Higher Normal School (ENS), Computer Science Department, L2IS Laboratory, Marrakech, Morocco²

Cadi Ayyad University, UCA, Faculty of Sciences Semlalia, LCSS, Marrakesh, Morocco³

Chouaib Doukkali University, ESIM, Polydisciplinary Faculty of Sidi Bennour, El Jadida, Morocco⁵

Abstract—Artificial intelligence (AI) is increasingly used to support musculoskeletal diagnosis from routine imaging, yet evidence remains fragmented across diseases, modalities, tasks, and validation designs. This systematic review synthesizes AI-based diagnosis, screening, grading, scoring, and structural assessment of rheumatoid arthritis (RA) and osteoporosis/osteopenia using radiography, computed tomography (CT), ultrasound, DXA-related imaging, and multimodal clinical data. Following PRISMA 2020 principles, Scopus, PubMed, Web of Science Core Collection, IEEE Xplore, and the ACM Digital Library were searched. Across the final exports, 2,239 records were identified; 247 full-text reports were assessed for eligibility, and 233 studies were included, while 25 reports could not be retrieved. CNN-based and transfer-learning approaches remained dominant, with increasing use of attention mechanisms, transformers, ensembles, explainable AI, and multimodal fusion. Osteoporosis research focused mainly on opportunistic screening and BMD-related prediction, whereas RA studies more often addressed diagnosis, joint-level damage, severity grading, and automated scoring. Methodological maturity remained uneven: 141 of 233 studies (60.5%) relied on internal validation only, whereas 44 (18.9%) reported external or multicenter validation. Future progress therefore depends less on marginal gains in headline metrics than on reproducible study design, robust external validation, calibration, and clinically integrated multimodal decision support.

Keywords—Rheumatoid arthritis; osteoporosis; osteopenia; radiography; X-ray; CT; multimodal learning; deep learning; convolutional neural networks; transfer learning; systematic review

I. INTRODUCTION

Artificial intelligence has become an increasingly important component of musculoskeletal and rheumatologic imaging because it can detect subtle image patterns, automate repetitive measurements, and support decision-making in settings where conventional interpretation is time-consuming or operator-dependent. The clinical rationale is especially strong in rheumatoid arthritis (RA), where structural joint damage requires reproducible detection and longitudinal assessment, and in osteoporosis/osteopenia, where underdiagnosis motivates

opportunistic screening from routinely acquired images. Recent reviews confirm substantial growth of AI in both disease domains, but most syntheses remain disease-, modality-, or task-specific [30]–[35].

The technical landscape is correspondingly heterogeneous. Published systems range from binary disease detection to multiclass grading, bone-mineral-density (BMD) or T-score estimation, joint localization, erosion assessment, modified Sharp/van der Heijde scoring, and image-plus-clinical decision support. Conventional convolutional neural networks (CNNs) and transfer learning remain widely used, while attention mechanisms, vision transformers, ensemble learning, explainability, uncertainty estimation, and multimodal fusion are increasingly represented. The EHR-plus-imaging framework reported by Ichahane et al. [8] and the attention-enhanced ensemble for osteoporosis detection by Ichahane and Assad [28] illustrate this movement from isolated image classifiers toward richer patient-level and ensemble reasoning systems.

Despite rapid publication growth, the evidence is difficult to interpret as a coherent basis for clinical translation. Apparent performance depends on dataset scale and provenance, disease prevalence, task definition, reference-standard construction, patient-level data partitioning, and whether evaluation is internal, independently held out, or genuinely external. Existing meta-analyses have demonstrated promising diagnostic accuracy in selected osteoporosis settings [31]–[35], but such pooled estimates do not resolve broader cross-task heterogeneity, while RA-focused reviews cover a wider clinical and data spectrum [30]. Consequently, headline accuracy or AUC values can overstate maturity when validation design, calibration, leakage risk, and transportability are not considered jointly.

The central research gap is therefore not a lack of high-performing AI models, but the absence of a cross-disease synthesis that jointly evaluates clinical task, imaging modality, reference-standard quality, data provenance, risk of bias, and validation maturity in rheumatoid arthritis and osteoporosis/osteopenia. This review addresses that gap through

*Corresponding author.

a common analytical framework that 1) maps clinical objectives and task formulations; 2) compares imaging modalities, datasets, model families, and evaluation metrics; 3) evaluates reference-standard heterogeneity, internal versus external validation, and adapted QUADAS-2 risk-of-bias domains; 4) contrasts the methodological trajectories of the two disease areas; and 5) identifies barriers to clinical translation through an evidence-maturity perspective. By combining five bibliographic databases with article-level eligibility assessment and structured methodological appraisal, the review is designed to distinguish algorithmic performance from the strength and transportability of the supporting evidence.

The remainder of this study is organized as follows. Section II situates the present work within existing systematic reviews and meta-analyses. Section III describes the search strategy, eligibility criteria, PRISMA-based study selection, data extraction, risk-of-bias appraisal, and evidence-synthesis framework. Section IV presents the main results across the predefined research questions. Section V discusses evidence maturity, methodological limitations, and priorities for clinical translation, while Section VI concludes the study.

II. RELATED WORKS

A. Existing Review Landscape and Comparative Positioning

Existing reviews confirm the rapid growth of AI in both rheumatoid arthritis (RA) and osteoporosis, but their scope is usually disease-specific, modality-specific, or focused on a narrower clinical objective. For RA, Momtazmanesh et al. provided a broad state-of-the-art review spanning imaging, omics, EHRs, sensors, treatment response, and prognosis [30], whereas imaging-focused reviews have generally concentrated on selected arthritis tasks rather than a common diagnostic-imaging framework. For osteoporosis, recent systematic reviews and meta-analyses have addressed plain radiographs [31], panoramic dental radiography [32], CT-based deep learning [33], multi-imaging machine-learning diagnosis [34], and opportunistic CT screening [35]. These reviews demonstrate promising diagnostic performance but repeatedly identify substantial heterogeneity, limited external validation, and inconsistent reporting as barriers to translation.

B. Limitations of Existing Reviews

The prior literature nevertheless leaves three gaps relevant to the present review. First, RA and osteoporosis/osteopenia are rarely assessed within a single methodological framework, despite their shared reliance on image-based classification, quantitative scoring, transfer learning, and clinically meaningful reference standards. Second, prior syntheses often narrow eligibility to one modality or one task, which improves statistical comparability but limits cross-modality and cross-disease assessment of evidence maturity. Third, although several osteoporosis meta-analyses appropriately use QUADAS-2 [31]–[35], the broader diagnostic-AI literature remains inconsistent in reporting patient-level partitioning, leakage control, calibration, dataset provenance, and genuinely external validation. A recent IJACSA systematic review also illustrates the value of explicit search, eligibility, and comparative-synthesis structures for reproducible review reporting [9].

C. Added Value of the Present Review

The present review therefore extends the existing review landscape rather than duplicating a disease- or modality-specific meta-analysis. It jointly evaluates RA and osteoporosis/osteopenia, covers five bibliographic databases, and applies a common PRISMA-based workflow to 233 included primary journal studies (Table IV). The synthesis combines adapted QUADAS-2 domains with AI-specific appraisal of patient-level splitting, train-test leakage, validation design, data provenance, calibration, and interpretability. Because the included tasks range from binary screening to BMD regression, joint-level damage analysis, severity grading, and structured RA scoring, the review prioritizes structured narrative synthesis and evidence-maturity comparison rather than a single pooled diagnostic estimate.

This cross-disease perspective is particularly useful for separating architectural performance from evidence strength: an internally validated high-AUC classifier is not treated as equivalent to a multicenter or external-validation study. The review thus focuses on transportability, reference-standard quality, and clinical task complexity in addition to headline performance.

Comparative positioning against prior reviews. In osteoporosis, Yen et al. synthesized six plain-radiograph studies with QUADAS-2 and diagnostic meta-analysis [31]; Tarighatnia et al. focused on seven panoramic-radiography studies and additionally evaluated publication bias [32]; Wang et al. synthesized 24 CT-based deep-learning studies [33]; Zhao et al. expanded the scope to 60 X-ray, CT, and MRI studies [34]; and Ghareh Mohammadi and Sebro examined 31 opportunistic CT studies covering osteoporosis and spine-fracture screening [35]. These reviews provide stronger within-modality statistical comparability, but they do not address RA within the same framework.

For RA, the broad review by Momtazmanesh et al. spans imaging, omics, EHRs, sensors, treatment response, and prognosis [30], but was not designed as a diagnostic-imaging systematic review with QUADAS-style appraisal. The present review therefore complements rather than replaces these prior syntheses: it sacrifices a single pooled effect in favor of cross-disease comparison of clinical task, imaging modality, reference standard, data provenance, validation rigor, and AI-specific leakage risk across a substantially larger body of included primary studies.

III. METHODS

This systematic review was conducted and reported according to PRISMA 2020 principles [10]. The review workflow comprised seven predefined components: review questions, information sources and search strategy, eligibility criteria, deduplication and staged study selection, structured data extraction, adapted risk-of-bias appraisal, and cross-dimensional evidence synthesis. Scopus was used as the primary discovery source, complemented by PubMed, Web of Science Core Collection, IEEE Xplore, and the ACM Digital Library. At each stage, study-selection decisions and reasons for exclusion were documented so that screening, retrieval, eligibility, and inclusion counts could be reconciled reproducibly.

A. Research Questions and Review Framework

Clinical AI for rheumatoid arthritis and osteoporosis/osteopenia can be organized along seven complementary dimensions: clinical objective, model family, imaging and data source, evaluation strategy, workflow role, task formulation, and methodological constraints on translation. These dimensions provide the analytical structure used to compare heterogeneous diagnostic-AI studies; their operational formulation is presented in Table I.

TABLE I. RESEARCH QUESTIONS AND ANALYTICAL FOCUS

RQ	Research question / analytical focus
RQ1	Clinical objectives targeted by AI systems.
RQ2	AI model families and methodological evolution.
RQ3	Imaging modalities, datasets, and data provenance.
RQ4	Evaluation metrics, reference standards, and validation rigor.
RQ5	Role of AI in diagnostic and reader-assistance workflows.
RQ6	Task formulations and label granularity.
RQ7	Methodological limitations and priorities for clinical translation.

These questions follow a diagnostic-review logic linking the target condition, AI index test, imaging input, reference standard, and clinically relevant outcome.

B. Search Strategy

The literature search combined four concept blocks: target condition (rheumatoid arthritis, osteoporosis, osteopenia), artificial-intelligence methods, medical-imaging inputs, and diagnostic or assessment tasks. The original Scopus search was generated on April 9, 2026, and the search was subsequently extended to PubMed, Web of Science Core Collection, IEEE Xplore, and the ACM Digital Library using database-specific adaptations of the same conceptual structure. The Scopus formulation is reported below to make the search logic immediately visible; the complete database-specific strategies are provided in Supplementary Table S2 and are publicly available through the Zenodo repository at <https://doi.org/10.5281/zenodo.22751177>. Exact execution dates for the supplementary database searches were not separately preserved in the frozen export metadata and are therefore not assigned retrospectively.

Research query:

TITLE-ABS-KEY (("rheumatoid arthritis" OR osteoporosis OR osteopenia) AND ("deep learning" OR "convolutional neural network*" OR CNN OR "ensemble learning" OR "attention mechanism*" OR "case-based reasoning" OR multimodal) AND ("x-ray" OR radiograph* OR "medical imag*" OR EHR OR "electronic health record*") AND (diagnos* OR detect* OR classif* OR predict* OR screening))

Applying the conceptually equivalent searches across the five sources yielded 1,053 Scopus records, 298 PubMed records, 551 Web of Science Core Collection records, 197 IEEE Xplore records, and 140 ACM Digital Library records (n = 2,239). Source overlap was handled by cross-database deduplication; therefore, database-specific retrieval counts are reported for

transparency and are not interpreted as independent study contributions.

The resulting records were merged into a single screening dataset before DOI- and title-based deduplication.

C. Inclusion and Exclusion Criteria

Eligibility was defined by three core requirements: primary peer-reviewed evidence, medical imaging as an input to the AI system, and quantitative evaluation of a direct rheumatoid arthritis or osteoporosis/osteopenia diagnostic or assessment task. Exclusion rules removed non-primary publications, non-target conditions or outcomes, and image-processing studies without a diagnostic or scoring endpoint. The operational inclusion and exclusion criteria are given in Table II.

TABLE II. ELIGIBILITY CRITERIA

Code	Operational criterion
IC1	Primary journal study evaluating AI for direct RA or osteoporosis/osteopenia diagnosis, screening, grading, scoring, or structural assessment.
IC2	Medical imaging is an AI input, alone or with clinical data; no imaging input fails IC2.
IC3	At least one quantitative evaluation metric is reported.
EC1	Non-primary publication (conference/proceedings, review, editorial, letter/note, erratum, or book chapter).
EC2	Prognosis, epidemiology, fracture-only, treatment-response, or risk-factor study without direct target-condition assessment.
EC3	Image processing only (segmentation, preprocessing, enhancement, or compression) without a diagnostic/scoring endpoint.
EC4	Non-target disease (e.g., osteoarthritis); generic "arthritis" is not treated as RA.

D. Screening Process and PRISMA Flow

Records from all five sources were merged into a common screening dataset. Deduplication was performed first by normalized DOI and subsequently by normalized-title comparison when DOI information was missing or inconsistent. From 2,239 records, 892 duplicates were removed, leaving 1,347 unique records for screening. After pre-retrieval exclusions, 272 reports were sought for retrieval. Twenty-five reports could not be retrieved; 247 reports were assessed for eligibility in full text, 14 were excluded after full-text assessment, and 233 studies were included. The complete selection pathway is shown in Fig. 1.

Study selection followed conservative disease and evidence rules. Generic "arthritis" was not treated as rheumatoid arthritis unless an explicit RA target was demonstrated, and fracture-only studies were not considered osteoporosis evidence unless osteoporosis/osteopenia itself was directly evaluated. A report was considered retrieved only when the complete article had actually been obtained and assessed. Abstracts, previews, and metadata could support exclusion only when an exclusion criterion was already explicit; otherwise, potentially eligible reports remained unresolved until full-text retrieval was attempted. After targeted retrieval, 25 reports could not be retrieved and therefore were not assessed for eligibility. Of the 247 reports assessed in full text, 14 were excluded, and 233 studies were included in the review. The complete study-selection pathway and reconciled counts are shown in Fig. 1.

The 14 reports excluded after full-text assessment and the corresponding reasons for exclusion are detailed in Supplementary Table S3, available at <https://doi.org/10.5281/zenodo.22751177>.

Methodological safeguards. Four rules were enforced throughout the review: 1) a report was considered retrieved only after the complete article had been obtained and assessed; 2) no

report was excluded solely because it was difficult to access; 3) generic arthritis was not interpreted as RA, and fracture-only outcomes were not interpreted as osteoporosis unless the target condition itself was directly assessed; and 4) raw accuracy or AUC values were not ranked across studies without considering task definition, reference standard, data provenance, and validation design.

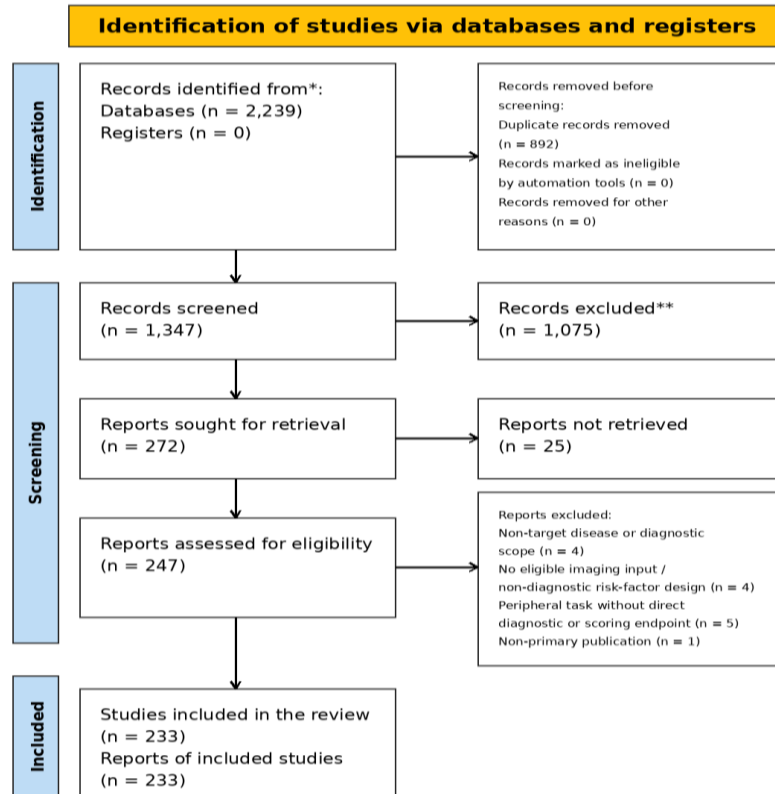


Fig. 1. PRISMA 2020 flow diagram of study identification, screening, retrieval, eligibility assessment, and inclusion.

E. Data Extraction and Synthesis Strategy

After study selection, full-text data extraction was performed using a standardized framework so that every included study could be compared on clinically and methodologically relevant dimensions. Extracted fields included publication year, disease focus, clinical objective, anatomical site, imaging modality, dataset size and provenance, AI architecture, multimodal inputs, task formulation, reference standard, validation design, external-validation status, and reported metrics (e.g., accuracy, sensitivity, specificity, F1-score, AUC, ICC, RMSE, and MAE). Patient-level splitting, potential leakage, class imbalance, calibration, interpretability, and data provenance were also recorded when reported. Metrics were preserved in their original task context rather than normalized into a single performance score.

F. Risk of Bias and AI-Specific Quality Appraisal

Risk of bias was appraised using an adapted QUADAS-2 structure covering patient selection, index test, reference standard, and flow/timing [29]. AI-specific extensions captured patient-level versus image-level splitting, train-test leakage risk, class imbalance, transparency of cohort construction,

calibration, explainability, and independent or multicenter external validation. Domain judgments were retained separately; they were not collapsed into a single numerical “quality score,” because such averaging can obscure distinct sources of diagnostic bias. The appraisal was used to contextualize performance and evidence maturity rather than as an automatic exclusion rule.

G. Evidence Synthesis and Evidence-Maturity Logic

A structured narrative synthesis was used because disease targets, anatomical sites, imaging modalities, class definitions, reference standards, dataset construction, model objectives, validation designs, and outcome metrics were too heterogeneous for a clinically defensible meta-analysis. Studies were synthesized by disease, clinical task, modality, model family, reference standard, and validation strategy. To avoid pseudo-precision, accuracy or AUC values from non-equivalent tasks were not pooled or averaged. A three-axis Evidence Maturity Framework was used to interpret the body of evidence across 1) clinical task complexity, 2) data richness/provenance, and 3) validation rigor. This framework is interpretive rather than a scoring system and is intended to distinguish high internal performance from stronger evidence of transportability and

clinical readiness. A single pooled effect estimate was therefore not pursued, because pooling non-equivalent targets, reference standards, imaging modalities, and validation designs would violate the clinical and statistical comparability required for a defensible diagnostic meta-analysis.

1) *Evidence-maturity framework — clinical task.* The continuum ranges from binary recognition to multiclass grading or BMD regression, and then to structured scoring, reader assistance, or clinically integrated decision support.

2) *Evidence-maturity framework — data.* The continuum ranges from small private single-center image sets to larger institutional/public cohorts and, at the highest level, multicenter or external cohorts combined with clinically meaningful structured data.

3) *Evidence-maturity framework — validation.* The continuum ranges from internal split or cross-validation to independent testing and, at the highest level, external/multicenter validation, calibration, reader studies, or prospective evaluation.

IV. RESULTS

A. Characteristics of the Included Studies

Fig. 2 shows the publication-year distribution of the 233 included studies and highlights the rapid expansion of the evidence base after 2019. Fig. 3 presents the distribution of imaging modalities, showing the predominance of radiography and computed tomography. Fig. 4 summarizes validation maturity by distinguishing internal validation, independent test sets, and external or multicenter evaluation.

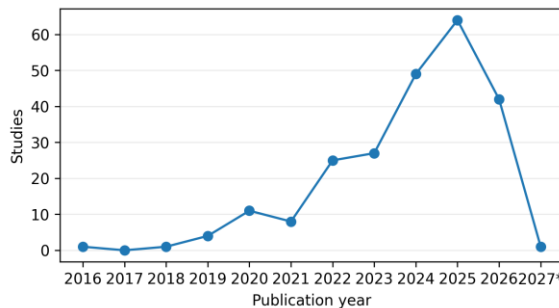


Fig. 2. Publication year of the 233 included studies. *The single study assigned to a 2027 issue had a 2026 DOI/online-publication record.

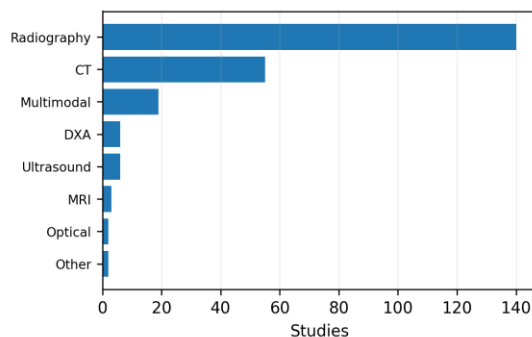


Fig. 3. Distribution of imaging modalities among the 233 included studies. Multimodal denotes studies using more than one imaging family.

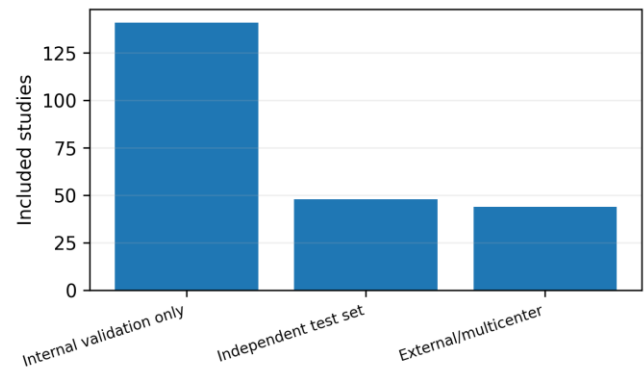


Fig. 4. Validation characteristics of the included studies: 141 internal-validation-only studies, 48 studies with an independent test set, and 44 studies with external or multicenter validation.

B. Overview of the Included Studies

Following deduplication, screening, report retrieval, and full-text eligibility assessment, 233 primary research studies published as journal articles were included: 198 addressing osteoporosis/osteopenia and 35 rheumatoid arthritis (Table V). Their target-condition distribution is shown in Fig. 5. The included evidence is methodologically diverse, but most studies still rely on internal validation; the corresponding risk-of-bias profile across adapted QUADAS-2 domains and the AI-specific split/leakage domain is shown in Fig. 6.

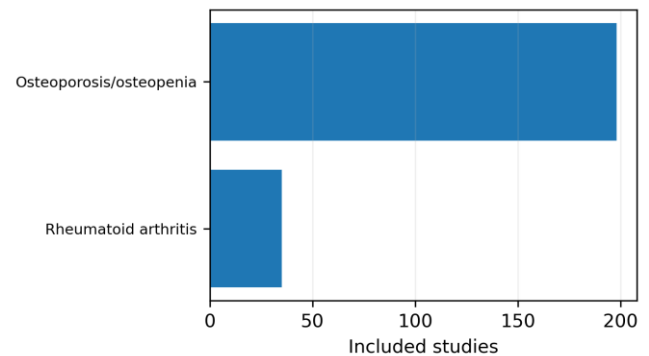


Fig. 5. Target conditions represented among the 233 included studies: 198 osteoporosis/osteopenia studies and 35 rheumatoid arthritis studies.

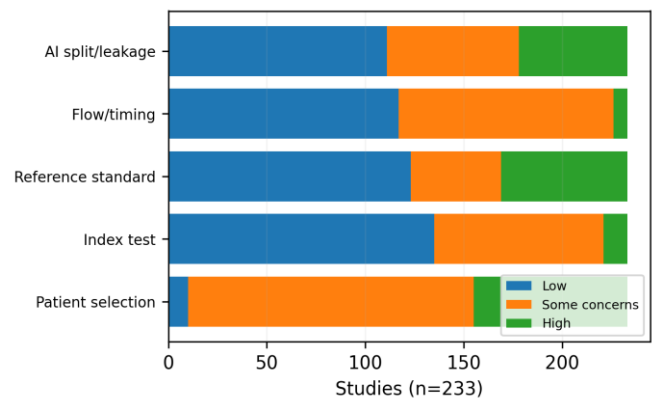


Fig. 6. Risk-of-bias assessment of the 233 included studies using adapted QUADAS-2 domains and an AI-specific split/leakage domain.

C. RQ1. Main Clinical Objectives Targeted by AI Systems

In osteoporosis and osteopenia, the dominant objective is opportunistic screening for low bone mineral density or disease using images acquired for other routine purposes [4]–[7], [12]–[14]. In RA, the objective space is more clinically granular and includes diagnosis, differential diagnosis, severity grading, joint-space or erosion analysis, and automated estimation of structured radiographic scores [2], [16], [17]. Overall, osteoporosis studies remain driven mainly by case finding and opportunistic screening, whereas RA studies are more deeply embedded in complex radiographic interpretation workflows.

D. RQ2. AI Model Families used Across the Included Studies

The dominant methodological pattern is the strong prevalence of deep convolutional image analysis. CNN-based and transfer-learning pipelines form the backbone of most osteoporosis studies and a large portion of RA studies. More recent work increasingly combines complementary architectures through attention or ensemble mechanisms; for example, the attention-enhanced E-DWE CNN ensemble proposed by Ichahane and Assad [28] integrates multiple pretrained backbones with adaptive weighting for knee-radiograph osteoporosis detection. In RA, multistage systems combining localization and classification are particularly prominent, while transformer-enhanced, radiomics-assisted, and explicitly interpretable systems are increasingly represented across both disease families. This methodological diversity is also illustrated by fine-tuned transfer learning for knee-radiograph osteoporosis classification [11], multimodal hand-radiograph/thermal RA modeling [21], interpretable hand-radiograph RA classification [22], and contextual deep-learning assessment of RA bone destruction [24].

E. RQ3. Data Types, Imaging Modalities, and Datasets

Plain radiography is the dominant study-level primary modality: 140/233 studies (60.1%). CT accounts for 55 (23.6%), multimodal imaging for 19 (8.2%), DXA-only and ultrasound for 6 each (2.6% each), MRI for 3 (1.3%), optical imaging for 2 (0.9%), and other imaging for 2 (0.9%). Osteoporosis research spans multiple anatomical sites and opportunistic imaging settings, whereas hand radiography remains central to RA. True image-plus-structured-data fusion remains comparatively uncommon. Additional included examples span cervical-radiography osteoporosis screening [12], spectral-CT osteoporosis assessment [23], few-shot knee-radiograph diagnosis [26], and custom deep-learning approaches for knee osteoporosis/osteopenia detection [27].

F. RQ4. Evaluation Metrics and Reported Performance

The included studies report a heterogeneous set of evaluation metrics. Accuracy, sensitivity, specificity, F1-score, and AUC dominate binary diagnostic tasks, whereas BMD estimation and RA scoring additionally use correlation, RMSE, MAE, and ICC. These metrics cannot be interpreted independently of validation design. A representative example is the hip-radiograph study by Jang et al. [4], in which AUC decreased from 0.867 on internal evaluation to 0.700 on an external cohort, illustrating the importance of transportability assessment.

Among the included studies, 141/233 (60.5%) used internal validation only, 48/233 (20.6%) reported an independent test set, and 44/233 (18.9%) reached external or multicenter validation. Consequently, the synthesis emphasizes validation rigor and reference-standard quality rather than ranking studies by headline accuracy.

G. RQ5. Impact of AI on Conventional Diagnostic and Reader-Assistance Workflows

In osteoporosis research, AI frequently turns routine images into opportunistic screening opportunities, allowing patients to be flagged for confirmatory DXA without a separate imaging pathway [5], [6], [13], [14]. In RA, AI is more often positioned as reader assistance, supporting joint localization, damage scoring, structured interpretation, and consistency [2], [17]. The evidence supports AI as a complementary decision-support layer rather than as a replacement for expert judgment.

H. RQ6. Dominant Task Formulations and Label Granularities

Osteoporosis studies are dominated by binary or near-binary screening tasks and BMD-related prediction [6], [7], [18], whereas RA studies are more heterogeneous and include binary recognition, RA-versus-OA discrimination, severity grading, joint-wise scoring, and end-to-end prediction of aggregate radiographic scores [2], [15]–[17]. Direct comparison of headline metrics is therefore misleading unless task complexity is considered.

I. RQ7. Limitations of the Current Literature and Opportunities for Future Work

Recurrent limitations include small private datasets, retrospective single-center designs, heterogeneous reference standards, insufficient reporting of patient-level splitting, limited calibration, and scarce external validation. The adapted risk-of-bias profile quantifies these concerns: 78/233 studies (33.5%) received a high-risk judgment for patient selection, 64/233 (27.5%) for the reference standard, and 55/233 (23.6%) for AI split/leakage. Only 44/233 studies (18.9%) reached external or multicenter validation. These weaknesses create risks of overfitting, spectrum bias, hidden data leakage, and site-specific performance inflation.

J. Cross-Dimensional Evidence Maturity

When the included studies are interpreted across task complexity, data richness, and validation rigor, a clear imbalance emerges. The field is strongest in algorithmic experimentation and internal discrimination performance, but substantially less mature in transportability and workflow-level evidence. Many studies occupy an intermediate position: they use clinically relevant imaging and modern architectures, yet remain dependent on retrospective single-center data and internal validation. The highest-maturity studies are those that combine clinically meaningful endpoints with independent or multicenter testing, transparent reference standards, and explicit attention to calibration, interpretability, or reader interaction. This pattern indicates that the next phase of progress is more likely to depend on study design and evidence quality than on marginal architectural novelty.

K. Summary of Key Findings

The consolidated evidence therefore reveals four consistent patterns. First, conventional radiography remains the most broadly deployed modality, while CT plays an important role in opportunistic osteoporosis screening and ultrasound remains particularly relevant to RA. Second, CNNs and transfer learning continue to dominate, although attention, transformers, ensemble strategies, and explainable AI are increasingly represented. Third, osteoporosis research is more frequently framed as opportunistic screening or BMD-related prediction, whereas RA research more often involves complex grading, localization, and scoring tasks. Fourth, validation quality remains the principal methodological bottleneck across both disease families, with external and prospective evaluation substantially less common than internal testing.

V. DISCUSSION

A. From Algorithmic Performance to Evidence Maturity

The revised evidence base indicates that the central question is no longer whether AI can achieve high performance on selected RA or osteoporosis datasets, but whether that performance is supported by a sufficiently mature evidence design. Across the 233 included studies, 141 (60.5%) relied on internal validation only, whereas 44 (18.9%) provided external or multicenter evidence. This imbalance explains why high reported discrimination should not automatically be interpreted as clinical transportability.

B. Task Complexity, Data Richness, and Validation Trade-offs

The three-axis synthesis highlights an important trade-off. Simpler binary screening tasks can achieve high discrimination with relatively modest datasets, whereas structured RA scoring, multiclass grading, and multimodal decision support demand richer labels, stronger reference standards, and more robust validation. Similarly, increasing data richness does not automatically improve clinical credibility if data provenance, patient-level partitioning, or external transportability remain weak. This is why performance metrics must be interpreted jointly with task complexity and validation design rather than ranked in isolation. For osteoporosis, opportunistic screening remains the clearest near-term translational pathway. For RA, the strongest role is likely to be structured reader assistance, where AI supports localization, scoring consistency, and repetitive interpretation while preserving clinician oversight.

C. Persistent Validation and Reporting Gaps

The strongest recurring weakness is limited external validation. Only 44 of the 233 included studies (18.9%) reached external or multicenter evaluation, whereas 141 (60.5%) relied on internal validation only. High-risk judgments were assigned to 78 studies for patient selection, 64 for the reference standard, and 55 for AI split/leakage. These findings reinforce the need for multicenter evaluation, prespecified patient-level partitioning, transparent provenance, calibration, and reporting of clinically meaningful uncertainty.

D. Limitations of this Review

At the review level, restriction to primary journal articles improves comparability but excludes potentially informative conference research. Search syntax required database-specific adaptation, and the evidence base remains too heterogeneous for a valid pooled diagnostic-effect estimate. Twenty-five potentially eligible reports could not be retrieved despite targeted searches and therefore could not be assessed for eligibility; they are reported separately in the PRISMA flow as reports not retrieved. In addition, reporting quality in the source literature was uneven for patient-level splitting, calibration, reference-standard construction, and external validation, limiting the precision of cross-study risk-of-bias judgments. These constraints should be considered when interpreting completeness and generalizability. Because the review did not estimate a single common pooled diagnostic effect, formal funnel-plot or regression-based tests of publication bias were not methodologically appropriate across the full evidence base. Nevertheless, publication and selective-reporting bias remain plausible: AI studies with favorable discrimination may be more likely to reach journal publication, while negative, null, or poorly performing models may remain unpublished or incompletely reported. Restricting the synthesis to peer-reviewed journal articles may further accentuate this dissemination bias. These considerations reinforce the need to interpret high reported performance cautiously, particularly when external validation, calibration, or prespecified analysis protocols are absent.

E. Future Research and Translational Priorities

Future research should prioritize methodological and translational maturity over incremental architectural novelty. High-value priorities include multicenter external validation using geographically and technically diverse cohorts; prospective reader-assistance studies; standardized patient-level data partitioning and leakage controls; calibration and uncertainty reporting; explicit comparison against clinically meaningful baselines; transparent reference standards; and multimodal fusion that combines imaging with laboratory, demographic, and EHR information. Shared benchmarks and reporting checklists tailored to diagnostic AI would improve comparability, while reader studies and prospective workflow evaluation would clarify whether algorithmic gains translate into better clinical efficiency, consistency, and patient-level decision support.

F. Methodological Priorities for Clinical Translation

Methodological maturity in diagnostic AI depends on validation strategy, dataset scale and provenance, reference-standard consistency, patient-level partitioning and leakage control, calibration and uncertainty assessment, clinically grounded explainability, prospective workflow evaluation, and multimodal integration. Each recurring limitation therefore implies a corresponding study-design or reporting priority; these gap-to-action relationships are synthesized in Table III.

TABLE III. METHODOLOGICAL GAP-TO-ACTION MATRIX

Observed gap	Recommended reporting / study-design response
Limited external validation	Use prespecified independent or multicenter test cohorts; report transportability and subgroup performance.
Small or private datasets	Use larger shared or multicenter cohorts with transparent provenance and patient-level partitioning.
Heterogeneous reference standards	Report task-specific reference standards, label generation, adjudication, and image-reference timing.
Unclear splitting/leakage control	Partition at patient level; isolate preprocessing and augmentation; document leakage controls.
Sparse calibration/uncertainty	Report calibration, uncertainty, and decision analysis when clinically relevant.
Explainability without reader evidence	Link XAI to error analysis and reader-centered evaluation rather than plausibility alone.
Few prospective workflow studies	Use prospective reader or workflow studies with time, consistency, and decision-impact outcomes.
Limited multimodal context	Fuse imaging with clinical variables only when provenance and validation remain transparent.

VI. CONCLUSION

This systematic review synthesizes 233 primary research studies on AI-based diagnosis, screening, grading, scoring, and structural assessment of rheumatoid arthritis and osteoporosis/osteopenia. Across the included evidence, radiography remains the dominant imaging modality, while CNNs and transfer learning continue to underpin most systems and newer work increasingly incorporates attention, transformers, ensembles, explainability, and multimodal fusion.

The principal finding is methodological rather than architectural. Of the 233 included studies, 141 (60.5%) relied on internal validation only, whereas just 44 (18.9%) reported external or multicenter validation. Thus, high discrimination on retrospective or single-center datasets should not be equated with clinical transportability. The adapted risk-of-bias assessment further indicates persistent concerns related to patient selection, reference standards, and AI-specific split/leakage practices.

Future progress should therefore prioritize evidence maturity: multicenter and prospective evaluation, reproducible patient-level partitioning, transparent reference standards, calibration and uncertainty reporting, reader-interaction studies, and genuinely multimodal integration. For osteoporosis, opportunistic screening from routine imaging appears particularly promising; for RA, structured reader assistance and automated scoring are likely to be the most clinically relevant near-term applications.

DECLARATION ON GENERATIVE AI

Generative AI tools were used only for language refinement, document-formatting assistance, and consistency checking. All eligibility decisions, data extraction, methodological judgments, interpretation, and conclusions were made or critically verified by the authors, who retain full responsibility for the manuscript.

SUPPLEMENTARY MATERIALS

The supplementary materials associated with this systematic review are publicly available through the Zenodo repository at <https://doi.org/10.5281/zenodo.22751177>. Supplementary Table S1 provides the complete study-level extraction for the 233 included studies; Supplementary Table S2 provides the database-specific search strategies for Scopus, PubMed, Web of Science Core Collection, IEEE Xplore, and ACM Digital Library; and Supplementary Table S3 lists the reports excluded after full-text assessment together with the corresponding exclusion reasons.

REFERENCES

- [1] G. S. Mate, A. K. Kureshi, and B. K. Singh, "An efficient CNN for hand X-ray classification of rheumatoid arthritis," *Journal of Healthcare Engineering*, vol. 2021, 2021.
- [2] Z. Zhou et al., "RATING: Medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning," *Patterns*, vol. 3, no. 10, 2022.
- [3] Y. Peng, X. Huang, M. Gan, K. Zhang, and Y. Chen, "Radiograph-based rheumatoid arthritis diagnosis via convolutional neural network," *BMC Medical Imaging*, vol. 24, no. 1, 2024.
- [4] R. Jang et al., "Prediction of osteoporosis from simple hip radiography using deep learning algorithm," *Scientific Reports*, vol. 11, no. 1, 2021.
- [5] K. Fang, X. Zheng, X. Lin, and Z. Dai, "A comprehensive approach for osteoporosis detection through chest CT analysis and bone turnover markers: harnessing radiomics and deep learning techniques," *Frontiers in Endocrinology*, vol. 15, 2024.
- [6] T.-Y. Yen et al., "Predicting osteoporosis from kidney-ureter-bladder radiographs utilizing deep convolutional neural networks," *Bone*, vol. 184, 2024.
- [7] H. G. Nguyen et al., "Artificial intelligence system for predicting areal bone mineral density from plain X-rays," *Osteoporosis International*, vol. 36, no. 11, pp. 2167–2176, 2025.
- [8] M. Y. Ichahane, N. Assad, and H. Ouahmane, "Enhancing rheumatoid arthritis diagnosis: Combining case-based reasoning on EHR data with deep learning on medical images," *Journal of Universal Computer Science*, vol. 32, no. 3, pp. 373–404, 2026.
- [9] M. Aatila, M. Lachgar, H. Hrimech, and A. Kartit, "Machine learning techniques in keratoconus classification: A systematic review," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 648–656, 2023.
- [10] M. J. Page et al., "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, p. n160, 2021.
- [11] U. B. Abubakar, M. M. Boukar, and S. Adeshina, "Evaluation of parameter fine-tuning with transfer learning for osteoporosis classification in knee radiograph," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 246–252, 2022.
- [12] K. Tamai et al., "Deep learning algorithm for identifying osteopenia/osteoporosis using cervical radiography," *Scientific Reports*, vol. 15, no. 1, 2025.
- [13] F. Gharehmohammadi and R. Sebro, "Deep learning opportunistic screening for osteoporosis and osteopenia using radiographs of the foot or ankle – a pilot study," *European Journal of Radiology*, vol. 184, 2025.
- [14] B. S. S. Dias et al., "Osteoporosis screening: Leveraging EfficientNet with complete and cropped facial panoramic radiography imaging," *Biomedical Signal Processing and Control*, vol. 100, 2025.
- [15] K. Ureten and H. H. Maras, "Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods," *Journal of Digital Imaging*, vol. 35, no. 2, pp. 193–199, 2022.

- [16] H.-J. Wang et al., "Deep learning-based computer-aided diagnosis of rheumatoid arthritis with hand X-ray images conforming to modified total Sharp/van der Heijde score," *Biomedicines*, vol. 10, no. 6, 2022.
- [17] H. Moradmand and L. Ren, "Multistage deep learning methods for automating radiographic Sharp score prediction in rheumatoid arthritis," *Scientific Reports*, vol. 15, no. 1, p. 3391, 2025. doi:10.1038/s41598-025-86073-0.
- [18] S. Lee, E. K. Choe, H. Y. Kang, J. W. Yoon, and H. S. Kim, "The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population," *Skeletal Radiology*, vol. 49, no. 4, pp. 613–618, 2020.
- [19] S. Kumar, P. Goswami, and S. Batra, "Enriched diagnosis of osteoporosis using deep learning models," *International Journal of Performativity Engineering*, vol. 19, no. 12, pp. 824–833, 2023, doi: 10.23940/ijpe.23.12.p7.824833.
- [20] D. Çetintaş, G. Kılıçarslan, T. Tuncer, and D. Çetintaş, "Automated diagnosis of rheumatoid arthritis from hand radiographs using artificial intelligence: A retrospective study," *Clinical Medicine Insights: Arthritis and Musculoskeletal Disorders*, vol. 19, 2026.
- [21] A. R. Kesavapillai, S. M. Aslam, S. Umapathy, and F. Almutairi, "RA-XTNet: A novel CNN model to predict rheumatoid arthritis from hand radiographs and thermal images," *Diagnostics*, vol. 14, no. 17, 2024.
- [22] K. Cai et al., "Deep learning classification of rheumatoid arthritis in hand radiographs: interpretability insights and web application," *ImmunoTargets and Therapy*, vol. 14, pp. 1333–1345, 2025.
- [23] H. Zhang et al., "Screening for osteoporosis based on IQon spectral CT virtual low monoenergetic images: Comparison with conventional 120 kVp images," *Heliyon*, vol. 9, no. 10, 2023.
- [24] K. Miyama et al., "Deep learning-based automatic bone-destruction evaluation system using contextual information from other joints," *Arthritis Research & Therapy*, vol. 24, no. 1, 2022.
- [25] Y.-J. Huang et al., "Radiographic bone texture analysis using deep learning models for early rheumatoid arthritis diagnosis," *Journal of Imaging Informatics in Medicine*, 2025.
- [26] H. Xie et al., "A few-shot learning framework for the diagnosis of osteopenia and osteoporosis using knee X-ray images," *Journal of International Medical Research*, vol. 52, no. 9, 2024.
- [27] S. M. Naguib, M. K. Saleh, H. M. Hamza, K. M. Hosny, and M. A. Kassem, "A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images," *Scientific Reports*, vol. 14, no. 1, 2024.
- [28] M. Y. Ichahane and N. Assad, "Multi-Model Attention-Enhanced CNN Ensemble for Automated Osteoporosis Detection in Radiographic Knee Images," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 3, pp. 177–194, 2025, doi: 10.22266/ijies2025.0430.13.
- [29] P. F. Whiting et al., "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," *Ann. Intern. Med.*, vol. 155, no. 8, pp. 529–536, 2011.
- [30] S. Momtazmanesh, A. Nowroozi, and N. Rezaei, "Artificial Intelligence in Rheumatoid Arthritis: Current Status and Future Perspectives: A State-of-the-Art Review," *Rheumatology and Therapy*, vol. 9, pp. 1249–1304, 2022, doi: 10.1007/s40744-022-00475-4.
- [31] T.-Y. Yen, C.-S. Ho, Y.-P. Chen, and Y.-C. Pei, "Diagnostic Accuracy of Deep Learning for the Prediction of Osteoporosis Using Plain X-rays: A Systematic Review and Meta-Analysis," *Diagnostics*, vol. 14, no. 2, art. 207, 2024, doi: 10.3390/diagnostics14020207.
- [32] A. Tarighatnia, M. Amanzadeh, M. Hamedan, A. Mohammadnia, and N. D. Nader, "Deep learning-based evaluation of panoramic radiographs for osteoporosis screening: a systematic review and meta-analysis," *BMC Medical Imaging*, vol. 25, art. 86, 2025, doi: 10.1186/s12880-025-01626-z.
- [33] A. Wang et al., "Deep Learning-Assisted Automated Diagnosis of Osteoporosis Based on Computed Tomography Scans: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research*, vol. 27, e77155, 2025, doi: 10.2196/77155.
- [34] R. Zhao et al., "The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research*, vol. 28, e75965, 2026, doi: 10.2196/75965.
- [35] F. Ghareh Mohammadi and R. Sebro, "Artificial Intelligence for Opportunistic Screening for Osteoporosis and Spine Fractures Using Computed Tomography: A Systematic Review and Meta-Analysis," *Journal of Computer Assisted Tomography*, online ahead of print, 2026, doi: 10.1097/RCT.0000000000001899.

APPENDIX A

Tables IV and V present representative osteoporosis/osteopenia and rheumatoid arthritis studies, respectively, illustrating the range of imaging modalities, model families, performance measures, and validation designs discussed in Section IV. The complete study-level extraction for all 233 included studies is provided in Supplementary Table S1. The database-specific search strategies are provided in Supplementary Table S2, and the 14 reports excluded after full-text assessment with their exclusion reasons are listed in Supplementary Table S3. All supplementary files are publicly available through Zenodo at <https://doi.org/10.5281/zenodo.22751177>.

TABLE IV. REPRESENTATIVE INCLUDED OSTEOPOROSIS/OSTEOPENIA STUDIES

Study	Dataset / Modality	Model	Metrics / Best values	Validation / comparative insight
Lee et al. (2020) [18]	Spine X-ray; 334 participants	VGG feature extraction + RF	AUC 0.74; Acc. 0.71; Sens. 0.81; Spec. 0.60	Internal only; early proof-of-concept with moderate performance.
Jang et al. (2021) [4]	Hip radiography; 1001 images + external validation	VGG16 + nonlocal network	Internal AUC 0.867; External AUC 0.700	External validation; AUC decreased from 0.867 internally to 0.700 externally.
Kumar et al. (2023) [19]	Knee X-rays; 372 images	Eight ImageNet-pretrained transfer-learning models	Best average Acc. 86.36%	Internal train/test split; broad transfer-learning benchmark without external validation.
Fang et al. (2024) [5]	Chest CT + BMD; 488 patients	Radiomics + 2D/3D DL	Best AUC 0.906	Independent single-center test set; strong CT-based comparative study.
Yen et al. (2024) [6]	KUB radiographs + DXA; 8913 pairs	DeepDXA-KUB CNN	AUROC 0.939 lumbar; 0.947 hip	Internal validation only; large real-world opportunistic screening cohort.
Nguyen et al. (2025) [7]	7060 pelvis/spine radiographs	Ensemble of seven DL models	AUC 0.96–0.97	Independent within-cohort test set; large population-based BMD prediction study.
Gharehmohammadi & Sebro (2025) [13]	Foot/ankle radiographs; 907 patients	Customized DL	Acc. 94.65%; AUC 0.87	Internal patient-level split; opportunistic screening from routine imaging.
Dias et al. (2025) [14]	Panoramic radiography	EfficientNet B5/B6/B7	Development CV Acc. 95–99%; separate DXA-referenced cohort evaluated	Independent DXA-referenced test cohort after development CV; stronger than CV alone.

TABLE V. REPRESENTATIVE INCLUDED RHEUMATOID ARTHRITIS STUDIES

Study	Dataset / Modality	Model	Metrics / Best values	Validation / comparative insight
Üreten & Maraş (2022) [15]	Hand radiographs; held-out test	YOLOv4 + VGG-16	RA vs normal: Acc. 90.7%; Sens. 92.6%; Spec. 88.7%; AUC 0.97	Internal held-out test; single-center differential-diagnosis study.
Mate et al. (2021) [1]	Hand radiographs; 290 images	Efficient CNN	Acc. 94.46%	Internal held-out test; strong binary classification on a compact dataset.
Zhou et al. (2022) [2]	Multimodal ultrasound; prospective + external test sets	RATING knowledge-guided DL	Prospective Acc. 86.1%; External 85.0%	Prospective and external test sets; comparatively mature multimodal RA validation.
Peng et al. (2024) [3]	9714 hand radiograph exams	CNN	AUC 0.955 RA vs not RA	Patient-level internal train/test split; large multicenter-source development dataset but no independent external cohort.
Wang et al. (2022) [16]	240 train / 104 test hand radiographs	Five CNN architectures	Recognition AUC 97.80%	Internal test set; strong recognition performance, with staging/scoring remaining more complex.
Moradmand & Ren (2025) [17]	679 training patients + 291 external test subjects	UNet + YOLOv7 + ViT multistage	Joint ID Acc. 99%; RMSE 9.73; MAE 5.35; ICC 0.702	Independent unseen-patient test set; multistage automated global Sharp-score prediction.
Huang et al. (2025) [25]	891 early RA + 1237 non-RA	Deep-TEN vs ResNet-50	AUC 0.69–0.73	Internal train/validation/test split; highlights the difficulty of early-RA discrimination.
Cetintas et al. (2026) [20]	311 RA + 259 controls	Attention DenseNet121/169	Acc. 88%; Recall 91%	Internal single-center train/validation/test split; attention-based classification with moderate data volume.