International Journal of Data Science and Analytics (2026) 22:249
https://doi.org/10.1007/s41060-026-01191-w

REVIEW

Streamlining anomaly detection for rheumatology: a systematic literature review

Abtin Ijadi Maghsoodi[1,2] · Vicki Quincey[1,3] · Hamidreza Rasouli Panah[1,4] · Matthew Parsons[1,2] · Lincoln C. Wood[5] · Cameron Walker[6] · Michael O'Sullivan[6] · Samaneh Madanian[4]

Received: 23 September 2025 / Accepted: 11 June 2026

Abstract

This study investigated the effectiveness of Artificial Intelligence (AI) and Machine Learning (ML) approaches for anomaly detection in rheumatology, focusing on their potential to improve clinical insights and identify deviations in disease patterns and diagnostics. This systematic review adhered to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. A comprehensive search was conducted across PubMed, Cochrane Library, Web of Science, Scopus, and EBSCO databases to identify cohort studies that developed and/or validated AI and ML models for anomaly detection in rheumatology. Data were extracted and studies were critically evaluated using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Prediction Model Risk of Bias Assessment Tool (PROBAST), with the search last updated on January 28, 2025. The systematic search yielded 2,089 unique citations, from which 40 studies met the inclusion criteria. AI and ML models included demonstrated high performance measures. Most studies focused on rheumatoid arthritis, utilising imaging data, biomarkers, and electronic health records. The quality assessment revealed that 72.5% of the studies had a low risk of bias; however, 25% exhibited high risk due to methodological limitations such as insufficient validation and unclear outcome definitions. Reporting standards were generally well adhered to, although critical gaps such as blinding and risk stratification were frequently overlooked. This study identified a predominance of unclear and high risks of bias across many models, highlighting the need for transparency improvements and addressing such limitations. Key areas for enhancement include incorporating robust calibration measures to improve models and performance measures.

✉ Hamidreza Rasouli Panah
hamid.rasoulipanah@autuni.ac.nz

[1] Health New Zealand Te Whatu Ora, Wellington, New Zealand

[2] Division of Health, Engineering, Computing and Science, School of Health, University of Waikato, Hamilton, New Zealand

[3] School of Medicine, University of Auckland, Auckland, New Zealand

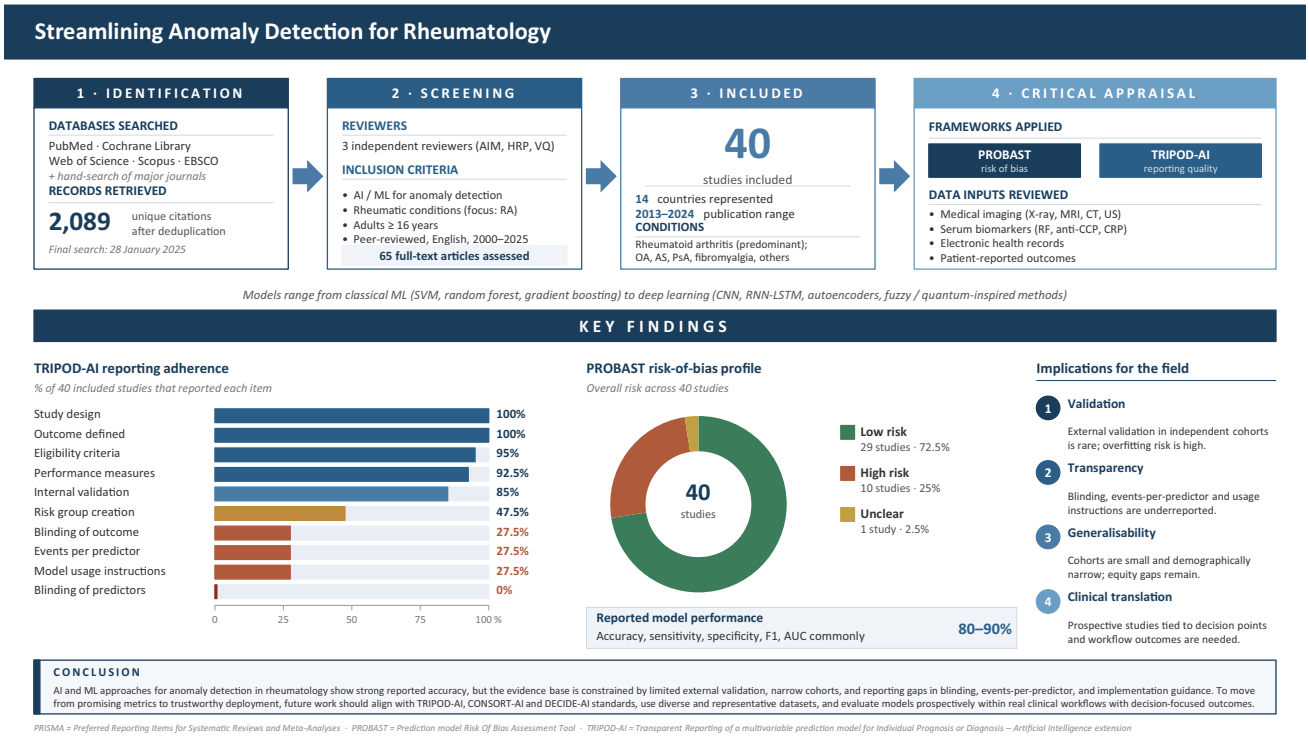
[4] Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland, New Zealand

[5] Department of Business and Management, University of Otago, Dunedin, New Zealand

[6] Department of Engineering Science, Faculty of Engineering, University of Auckland, Auckland, New Zealand

Published online: 31 July 2026

Graphical abstract



Keywords Rheumatology · Anomaly detection · Rheumatoid arthritis · Artificial intelligence · Machine learning · Systematic review

1 Introduction

Rheumatology encompasses a diverse range of disorders affecting joints, muscles, and connective tissues, with rheumatoid arthritis (RA) among the most prevalent and debilitating conditions [1]. RA affects millions globally and remains a leading cause of disability, characterised by chronic inflammation, joint damage, and systemic complications [1, 2], despite advances in targeted therapies [3, 4]. Similarly, for other rheumatic conditions, including osteoarthritis, lupus, and Sjögren's syndrome, outcomes of clinical trials for novel treatments have limited [5]. Early and accurate detection of anomalies in disease patterns, such as deviations in symptoms or diagnostics, is crucial for timely interventions and improving patient outcomes [6, 7]. Anomaly detection focuses on deviations that may not align with typical cases, helping to uncover misdiagnoses and tailor treatments [8]. Despite its promise, implementing effective systems is challenged by heterogeneity in presentation, progression, and treatment response [9].

In clinical settings, anomaly detection is most useful when it is explicitly tied to a decision point and a predefined action. In rheumatology, this commonly aligns with earlier recognition and referral for suspected inflammatory arthritis, earlier identification of flare or subclinical disease activity during treat to target monitoring, detection of treatment non response that prompts timely therapy adjustment, and safety surveillance for medication related adverse events. Framing anomaly detection in this workflow oriented way clarifies how model outputs could translate into measurable clinical and patient relevant outcomes, such as reduced time to specialist assessment, earlier initiation or escalation of disease modifying therapy, fewer severe flares, and prevention of avoidable adverse events.

Among the AI and ML approaches applied to anomaly detection in clinical settings, method families range from traditional machine learning algorithms, including support vector machines, random forests, decision trees, gradient boosting, and logistic regression, to deep learning architectures such as convolutional neural networks for imaging-based tasks, recurrent networks for sequential clinical data, and autoencoders for unsupervised pattern discovery. These approaches can be further distinguished by learning paradigm: supervised methods require labelled case and control groups, unsupervised methods seek to discover anomalous patterns without pre-specified labels, and semi-supervised

methods combine limited labels with larger volumes of unlabelled data. The choice of method family and paradigm has practical implications for the types of clinical data that can be leveraged, the volume and quality of labelled training data required, and the interpretability of model outputs in clinical decision-making.

Technological interventions have become increasingly essential in addressing these challenges and enhancing healthcare systems' efficacy. Among various technologies, Machine Learning (ML), a subset of Artificial Intelligence (AI), has emerged as a transformative tool in healthcare, showcasing its potential in predictive analytics and personalised medicine. ML methods excel in analysing complex datasets, uncovering patterns that may surpass the conventional statistical approaches [10]. In rheumatology, ML models have shown promise in predicting treatment response, disease progression, and identifying atypical cases [3, 4]. Specifically for RA, ML-based anomaly detection may identify subtle deviations before clinical onset [7]. Recent studies [5, 11–13] highlight both promise and persistent gaps; many models suffer from bias and lack external validation [9], and standardised evaluation remains rare [3]. The underrepresentation of diverse patient populations in model development further limits the generalisability of these tools [8]. Studies also frequently omit autoantibody profiles, disease severity measures, and patient-reported outcomes, undermining diagnostic precision [12]. Hence, this heterogeneity demands robust and transparent ML models capable of integrating varied data sources, to improve anomaly detection [6, 7].

In this review, anomaly detection in rheumatology is defined as the identification of atypical or deviant patterns within patient data that can signal early or otherwise hard to recognise disease states, prior to or at the point of formal diagnosis. Practically, this includes detecting abnormal imaging signatures, laboratory profiles, or record based patterns that differentiate individuals at elevated risk or with early inflammatory disease from typical presentations or comparator populations. This framing distinguishes anomaly detection from post diagnostic monitoring tasks such as flare prediction or treatment optimisation.

This review addresses three research questions. First, what AI and ML methods have been applied to anomaly detection in rheumatology, and how do they perform across different data modalities and clinical tasks? Second, what is the methodological quality and reporting transparency of these studies as assessed by PROBAST and TRIPOD-AI? Third, what are the key gaps and priorities for clinical translation, validation, and equitable deployment? In addressing these questions, this review contributes: a targeted synthesis of AI and ML anomaly detection applications in rheumatology, a dual quality appraisal using TRIPOD-AI and PROBAST frameworks, and actionable recommendations for future research and clinical practice.

While previous reviews [3, 4, 8, 9, 12, 13], have evaluated the broader applications of AI and ML in rheumatology, this study specifically examines anomaly detection as a tool for addressing the heterogeneity inherent in rheumatic diseases. Notably, this study prioritises interpretability, diverse data integration, rigorous validation protocols, and highlights underrepresented populations, a critical factor for ensuring equitable healthcare delivery. This research goes beyond prior reviews by proposing actionable recommendations, including adopting rigorous validation protocols and developing comprehensive datasets that capture the multifaceted nature of rheumatic diseases. By systematically evaluating transparency, generalisability, and clinical relevance, this study highlights the potential of ML and provides a roadmap for its effective integration into rheumatology practice.

2 Methods

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [14, 15], and adhered to the recommendations outlined in the Cochrane Handbook for Systematic Reviews of Interventions [16]. The protocol was registered in PROSPERO as *Streamlining blood anomaly detection for rheumatology: A Systematic Review* (CRD42024568370).

3 Eligibility criteria

Eligible studies included those evaluating AI-based tools and techniques for the early anomaly detection of rheumatologic conditions, focusing on RA. Rheumatologic conditions represent a diverse group of disorders primarily characterised by inflammation and degeneration of joints, connective tissues, and musculoskeletal structures. These conditions are frequently characterised by autoimmune mechanisms, wherein the immune system aberrantly targets self-tissues, resulting in persistent inflammation, pain, joint stiffness, and swelling. This pathological process may extend beyond the musculoskeletal system in advanced stages, leading to significant systemic manifestations and multi-organ involvement. Among these conditions, RA stands out as a prototypical systemic inflammatory disease targeting synovial joints, which leads to progressive joint damage if untreated. Clinically, RA manifests as symmetrical joint pain and swelling, prolonged stiffness, and systemic symptoms such as fatigue and low-grade fever. Diagnostic findings often include elevated inflammatory markers, such as C-reactive protein and erythrocyte sedimentation rate, and autoantibodies like rheumatoid factor and anti-citrullinated protein antibodies. Imaging studies, including X-rays, ultrasound,

or MRI, further help identify characteristic features such as joint space narrowing and erosions, which are hallmarks of the disease.

Studies focusing on genes, genomics, and phenotyping were excluded, as these approaches, while essential for understanding the molecular underpinnings of rheumatologic conditions, do not directly address early clinical detection workflows using routinely available clinical data. This exclusion does not apply to routine clinical biomarkers, including serology and autoantibodies such as rheumatoid factor and anti citrullinated protein antibodies, when these were used as predictors within AI based early detection models. Genomic and phenotypic studies often aim to identify genetic predispositions or specific phenotypic markers associated with the disease, which falls outside the scope of this review that focuses on predictive tools and techniques for early clinical detection. Additionally, such studies typically require different methodologies and datasets, making their integration into this review's framework infeasible. By focusing on early prediction, the review explores AI-based tools capable of identifying individuals at risk before clinical symptoms fully develop, aligning to enable timely intervention and potentially mitigating disease impact. While such studies provide valuable insights into disease management and progression, they do not align with the primary aim of this review, which is to explore AI applications for early anomaly detection and disease prediction before diagnosis. Similarly, studies predicting disease flare-ups or evaluating the effects of drugs and medications were excluded, as these are concerned with disease monitoring and treatment optimisation rather than the pre-diagnostic identification of at-risk individuals.

Accordingly, studies employing AI methodologies were considered, including but not limited to ML, Neural Networks (NN), Deep Learning (DL), and other computational approaches. These studies needed to have leveraged AI for analysing clinical or non-clinical data to detect anomalies indicative of rheumatologic conditions, such as imaging abnormalities, laboratory findings, or patterns in clinical records. For synthesis, approaches were additionally classified by learning paradigm as supervised, unsupervised, or semi supervised, based on how labels were used during model development. The search was limited to publications in the English language from 2000 to 2025, as it was believed that older results would no longer apply due to recent developments in medicine and healthcare. The review excluded studies focusing solely on non-AI-based methods, genomic or molecular approaches unrelated to disease diagnosis and prognosis with AI, or treatment-centric research. Non-peer-reviewed publications, articles not in English, and those addressing conditions outside the scope of this review were omitted. The population of interest comprised patients aged 16 years and older, ensuring a focus on adult and adolescent-onset rheumatologic conditions while excluding juvenile conditions such as Juvenile Idiopathic Arthritis. Eligible studies were required to include control groups without AI interventions to allow for direct comparisons between AI-based and traditional diagnostic methods, if any. Only studies conducted in clinical settings or using clinical or non-clinical data relevant to real-world applications were included. This criterion was established to ensure the practicality and clinical relevance of the findings.

The search window of 2000 to 2025 was set intentionally broad to ensure comprehensive coverage and avoid prematurely excluding early applications of computational methods in rheumatology. The absence of eligible studies before 2014 in the final set reflects the practical reality that AI and ML applications meeting the inclusion criteria did not appear in the peer-reviewed literature until the mid-2010s, consistent with the broader timeline of modern machine learning adoption in clinical research.

Importantly, rule-based and fuzzy-logic expert systems were not excluded from the review when they met the AI and ML inclusion criteria. Three included studies rely on fuzzy or hybrid fuzzy approaches: the eRheumatologist mobile expert system [17], a quantum-inspired fuzzy cognitive map decision support system for early-stage RA [18], and a fuzzy expert system combined with machine learning for RA diagnosis [19]. The limited number of pre-2014 eligible records therefore reflects the historical pace of AI and ML adoption in clinical rheumatology rather than methodological exclusion of earlier rule-based work, and the post-2014 trend toward deep learning observed in the included literature represents a genuine shift in the underlying field.

Primary outcomes of interest included diagnostic performance metrics of the AI method, such as sensitivity, specificity, positive predictive value, and early detection rates. Studies focusing on the early detection of disease-specific features, such as synovitis, bone erosion, or elevated autoantibody levels (e.g., rheumatoid factor or anti-citrullinated protein antibodies), were prioritised. Additional primary outcomes included reductions in diagnostic delay, identifying subclinical disease activity, and differentiating RA from other rheumatologic conditions. Secondary outcomes investigated included patient-related outcomes such as disease progression, the incidence of complications (e.g., cardiovascular comorbidities or systemic inflammatory sequelae), and quality of life improvements following earlier diagnosis. Studies assessing the feasibility and cost-effectiveness of AI-driven approaches in clinical practice were also considered, along with their impact on healthcare delivery metrics such as time to diagnosis, interdisciplinary care coordination, and patient satisfaction with diagnostic processes.

4 Search strategy and study selection

The literature search was led by A.I.M and H.R.P assisted by a subject librarian to pinpoint medical subject headings and keywords reflective of the inclusion criteria. Electronic databases searched were EMBASE, PubMed, CINAHL, OECD health policies and data, NHSEED, Web of Science, SCOPUS, SSRN, Cochrane Library, and ProQuest. Hand-searching was conducted in publisher databases, including ScienceDirect, SpringerLink, Sage, and Wiley. A manual hand search and text word search was also conducted within the abstracts and titles of the mentioned databases and in high-impact journals in the field of rheumatology, including the Frontiers in Immunology, Arthritis Research & Therapy, Frontiers in Medicine, Arthritis & Rheumatology, Arthritis Care & Research, Annals of the Rheumatic Diseases, Nature Reviews Rheumatology, Rheumatology Journal, Lancet Rheumatology, RMD Open, and bibliography information of comprised studies and previous reviews aligned with the inclusion criteria. The search strategy using the medical subject headings and other extracted keywords is presented in Appendix S1. The last update for the review was performed on January 28, 2025. Titles and abstracts were reviewed in the first screening stage, and articles not meeting the inclusion criteria were excluded. Full-text papers were retrieved for further review, and in cases of insufficient information in the title and abstract, the complete article was reviewed. Three reviewers (A.I.M, H.R.P, V.Q) were involved with the eligibility assessment of the studies. Two reviewers (A.I.M and H.R.P) independently assessed the eligibility of studies retrieved from the literature search for potential inclusion. In cases where consensus could not be achieved, a third team member (V.Q) reviewed the studies to reach a concluding consensus.

5 Data extraction and analysis

The quality assessment of eligible papers was independently conducted by two reviewers (A.I.M and H.R.P), with discrepancies resolved through discussion with a third reviewer (V.Q). Extracted information included the country, data source (if available), type of AI/ML model used, the specific rheumatic disease studied with a primary focus on rheumatoid arthritis, study objectives, dataset characteristics (e.g., sample size, data type, and preprocessing), and performance metrics (e.g., accuracy, sensitivity, specificity, and AUC). We also extracted whether studies reported interpretability approaches, including intrinsically interpretable model choices or post hoc explanation techniques. Furthermore, the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD-AI) [20, 21] was used for further data extraction. This study employed the PROBAST (Prediction model Risk Of Bias Assessment Tool) [22] to evaluate the quality of the included studies. The signalling questions were rated using a structured scale (yes, probably yes, probably no, no, or no information), and these responses were used to determine the risk of bias for each domain as “low,” “high,” or “unclear.” Additionally, the relevance of three domains, participants, predictors, and outcomes, was assessed for applicability, with ratings of “low,” “high,” or “unclear.” For each prediction model, including those from studies with multiple models, the overall risk of bias was determined across four domains based on specific criteria. A "low" risk of bias required all domains to be rated as low, and for the prediction development model to include internal validation or be based on a very large dataset without external validation. A "high" risk of bias was assigned if at least one domain was rated as high, or if a prediction development model lacked internal and external validation despite other domains being rated as low. An "unclear" risk of bias was given when one or more domains were rated as unclear while the others were rated as low. The risk of bias was assessed independently by at least three authors. Any discrepancies in their evaluations were resolved through collaborative discussion to achieve consensus. Given heterogeneity in clinical tasks and reporting, performance metrics were synthesised narratively and interpreted alongside dataset size, outcome definition, and validation design rather than being combined into pooled estimates.

6 Comparison with previous reviews

The current study identified 11 relevant reviews, encompassing narrative reviews [23], scoping reviews [24], and systematic literature reviews [25]. The identified reviews demonstrated varying scopes and methodological approaches. For instance, Kedra et al. [26] examined big data and AI applications in rheumatic and musculoskeletal diseases, finding that clinical data (47%) dominated, followed by imaging (29%) and biological data (15%), with ML techniques accounting for 97% of AI methods. Their work highlights the diverse data landscape in RMDs and lays a robust foundation for future guidelines. Similarly, Mendoza-Pinto et al. [4] reviewed ML approaches for predicting treatment response in rheumatoid arthritis, emphasising the need for better external model validation, increased transparency in reporting, and improved data integration to boost clinical relevance. Danieli et al. [11] expanded the focus to autoimmune diseases, including systemic lupus erythematosus and

rheumatoid arthritis, demonstrating the potential of ML for early diagnosis and prognostic modelling. In another study, Imtiaz et al. [6] surveyed deep learning and ML techniques for diagnosing osteoarthritis and rheumatoid arthritis, identifying promising methods and challenges such as limited datasets, poor model interpretability, and integration issues in clinical practice. Momtazmanesh et al. [3] provided a state-of-the-art review on AI in rheumatoid arthritis, outlining both advancements and the technical and ethical challenges ahead, while Bragazzi et al. [8] explored the use of big data, digital technologies, and AI in psoriatic arthritis, noting their promise for early detection and management alongside research and integration challenges. Stafford et al. [9] offered a broader perspective on autoimmune diseases, revealing that only 7.7% of studies combined multiple data types. This gap underscores the potential benefits of integrating diverse data sources for more robust research. Avramidis et al. [12] compared diagnostic methods for rheumatoid arthritis and demonstrated that deep learning outperforms traditional approaches in accuracy, thereby underscoring the potential of AI to enhance clinical decision-making. Altıkardeş et al. [27] examined ML and deep learning applications in rheumatological screening, focusing on image processing techniques, to assess their effectiveness in enhancing diagnostic accuracy and workflow efficiency. Kedra et al. [28] reviewed ML strategies for enhancing disease management in rheumatoid arthritis, outlining current practices and future directions to optimise patient monitoring and treatment personalisation. Afrazeh and Shomalzadeh [5] provided a comprehensive review of AI applications in arthritis care, detailing advancements in diagnostics, prognostics, and treatment innovations and discussing emerging trends and challenges for future clinical integration.

The current systematic review distinguishes itself from these previous works in several crucial aspects. First, it is the only review specifically focused on anomaly detection in rheumatology, addressing a critical gap in the literature. Second, the current study employed a rigorous methodology following PRISMA guidelines and utilised both TRIPOD and PROBAST tools for quality assessment, providing a more comprehensive evaluation of the included studies. Unlike previous reviews that broadly examined AI applications, the current study specifically evaluates the effectiveness of AI and ML approaches in identifying deviations in disease patterns and diagnostics. Furthermore, while earlier reviews primarily focused on diagnostic applications or specific disease types, the current review extends beyond diagnosis to examine anomaly detection across the entire spectrum of rheumatological care. The systematic approach to identifying and evaluating studies, combined with the specific focus on anomaly detection, provides a more targeted and comprehensive understanding of how AI and ML can improve clinical insights and identify deviations in disease patterns in rheumatology. A structured comparison of previous reviews and the present systematic review is provided in Table 1.

Table 1 Summary comparison of previous reviews and the current systematic review

Review	Year	Type	Scope	Strengths	Limitations
Kedra et al. [26]	2019	SLR	Big data and AI in RMDs	Broad scope; EULAR aligned	Not focused on anomaly detection
Mendoza-Pinto et al. [4]	2024	SLR	ML for RA treatment response	Clinically focused	Limited to treatment response
Danieli et al. [11]	2024	SLR	ML in autoimmune diseases	Multi-disease scope	No quality appraisal framework
Imtiaz et al. [6]	2022	Review	DL/ML for OA and RA diagnosis	Identifies key methods	No PROBAST or TRIPOD assessment
Momtazmanesh et al. [3]	2022	State-of-art	AI in RA	Comprehensive RA coverage	Not anomaly detection specific
Bragazzi et al. [8]	2022	SLR	Big data/AI in PsA	Digital health perspective	Single disease focus
Stafford et al. [9]	2020	SLR	AI/ML in autoimmune diseases	Broad autoimmune scope	Limited multi-modal analysis
Avramidis et al. [12]	2021	Review	DL vs human diagnosis in RA	DL-human comparison	Narrow diagnostic focus
Afrazeh and Shomalzadeh [5]	2024	Review	AI in arthritis care	Broad clinical coverage	No systematic quality appraisal
Current review	2025	SLR	Anomaly detection in rheumatology	PRISMA + TRIPOD-AI + PROBAST; anomaly detection focus; equity lens	Limited to English; no meta-analysis

7 Results

7.1 Description and characteristics of studies

The literature search of all sources yielded 2089 unique citations, of which 65 were selected for critical and comprehensive reading after reviewing the titles and abstracts. A total of 40 studies reporting on applications and development of AI and ML methods for anomaly detection met the inclusion criteria [17–19, 29–56, 56–58, 58–61] (Fig. 1).

Three studies were conducted in the USA [36, 42, 45], four in China [30, 33, 38, 48], and eight in India [31, 32, 35, 43, 49, 56, 60, 61]. The remaining studies originated from Germany [37, 39, 41, 44, 54], France [29, 57], Japan [53, 62], the Philippines [17], Taiwan [34, 58], Canada [40, 46], Saudi Arabia [31], Egypt [47], Turkey [19, 50, 56, 58], Malaysia [55], the Netherlands [44], New Zealand [7], Cuba [51], Australia [52], Iran [18], and the UK [59]. The included studies encompassed retrospective analyses [34, 36, 42, 44, 58, 59], cross-sectional assessments based on single-time-point observations [19, 29, 30, 32, 33, 35, 37, 38, 40, 43, 47, 48, 50, 51, 53–56, 56–58, 60–62]. Additionally, a few observational surveys specifically compared AI-driven tools with clinical judgment [39, 57]. The settings were predominantly hospital-based (e.g., rheumatology clinics, imaging centres) or involved broader datasets [34, 42, 44, 59], while the time of assessment ranged from short-term feasibility evaluations [18, 45] to multi-year retrospective data collection [33, 36, 58].

Participant sample sizes varied, from fewer than 20 [18] to upwards of thousands or millions of records [34, 44, 59]. Although most studies centred on Rheumatoid Arthritis (RA) detection, classification, or severity assessment [7, 18, 19, 31–44, 46–56, 56–58, 58, 60–62], several also addressed Osteoarthritis (OA) [29, 40, 54, 58], Ankylosing Spondylitis (AS) [30], Psoriatic Arthritis (PsA) [37], Fibromyalgia (FM) [57], and other autoimmune or inflammatory conditions [17, 46]. The primary outcomes often involved diagnostic accuracy or anomaly detection performance measures [7, 31, 43, 47, 48], whereas secondary outcomes included disease activity measurement, severity grading, or comparison to standard clinical indices [36, 42, 53, 58].

Many studies utilised imaging methods, such as X-ray, MRI, thermal scans, high-resolution peripheral quantitative CT (HR-pQCT), or fluorescence optical imaging, to capture structural and inflammatory changes [29, 32, 48, 50, 51, 53, 54, 56, 58, 58], while others leveraged electronic health records (EHR) for diagnosis and prediction purposes

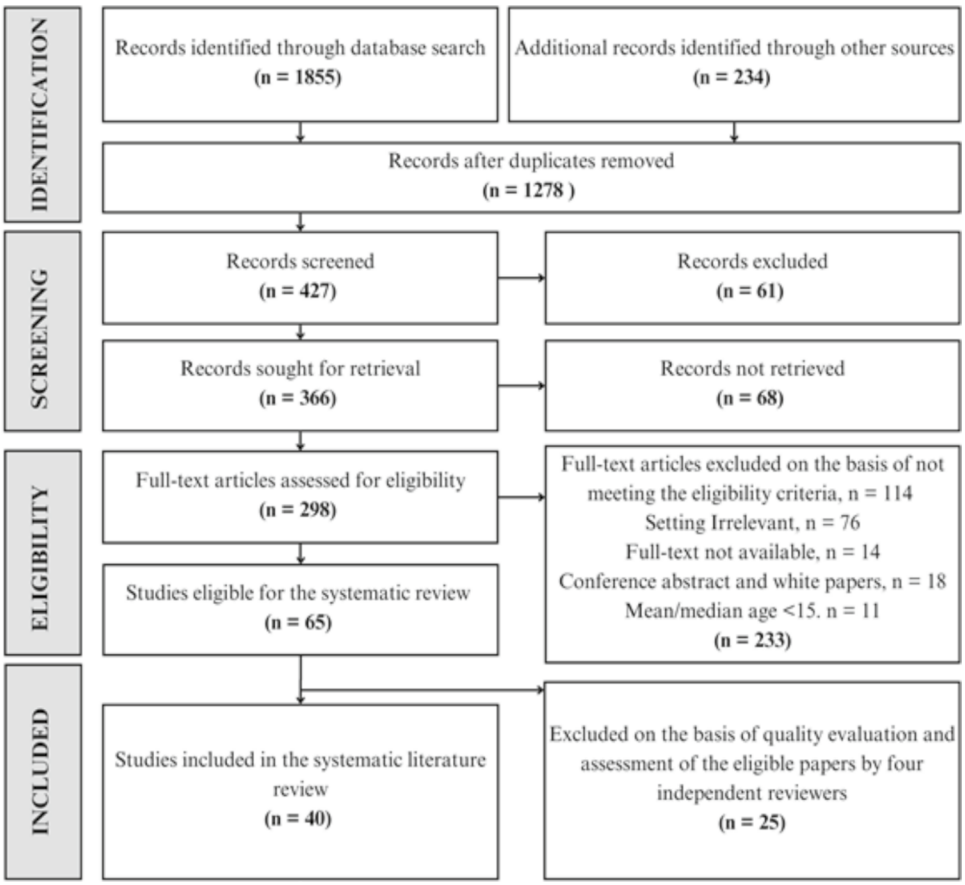


Fig. 1 PRISMA flow diagram for selecting studies

[34, 42, 44, 56, 59, 61]. Additional research relied on serum biomarkers or lab profiles [19, 30, 33, 38, 40] or used synthetic or specialised cohorts [17, 18, 46]. Typical dataset characteristics included patient demographics, clinical markers, such as rheumatoid factor (RF), anti-cyclic citrullinated peptide (anti-CCP), erythrocyte sedimentation rate (ESR), and C-reactive protein (CRP), imaging-derived parameters (e.g., joint space narrowing, erosions), and textual information (clinical notes, symptom checkers). These elements were frequently noted in studies leveraging large databases [34, 36, 59], lab and serum biomarker analyses [19, 40, 46], and imaging-focused investigations [48, 53, 58]. In addition, some projects integrated patient-reported data or questionnaire-based inputs [41, 45].

Deployed AI/ML methods ranged from classical approaches (support vector machines, decision trees, random forests, gradient boosting) to advanced deep learning architectures (CNNs, RNN-LSTM, cost-sensitive neural networks, autoencoders, quantum fuzzy cognitive maps) [18, 31, 38, 41, 47, 60, 62]. Only a minority of studies explicitly described interpretability approaches, and post hoc explainability methods were infrequently reported, despite frequent reliance on complex models Performance metrics, including accuracy, sensitivity, specificity, precision, recall, F1 scores, and AUC, commonly exceeded 80–90% [7, 30, 32–34, 40, 43, 46, 47]. Reported performance was therefore considered in relation to study scale and validation design, as smaller development cohorts with limited validation were more susceptible to optimistic estimates. Meanwhile, comparison groups generally consisted of healthy controls, alternative rheumatic diagnoses, or different ML models [38, 40, 51, 54, 55]. A comprehensive descriptive summary of the 40 studies is provided in Appendix S2. Conceptually, anomaly detection was most often implemented as supervised discrimination between RA and comparators, with fewer studies using unsupervised clustering or representation based discovery, and semi supervised strategies being rarely described.

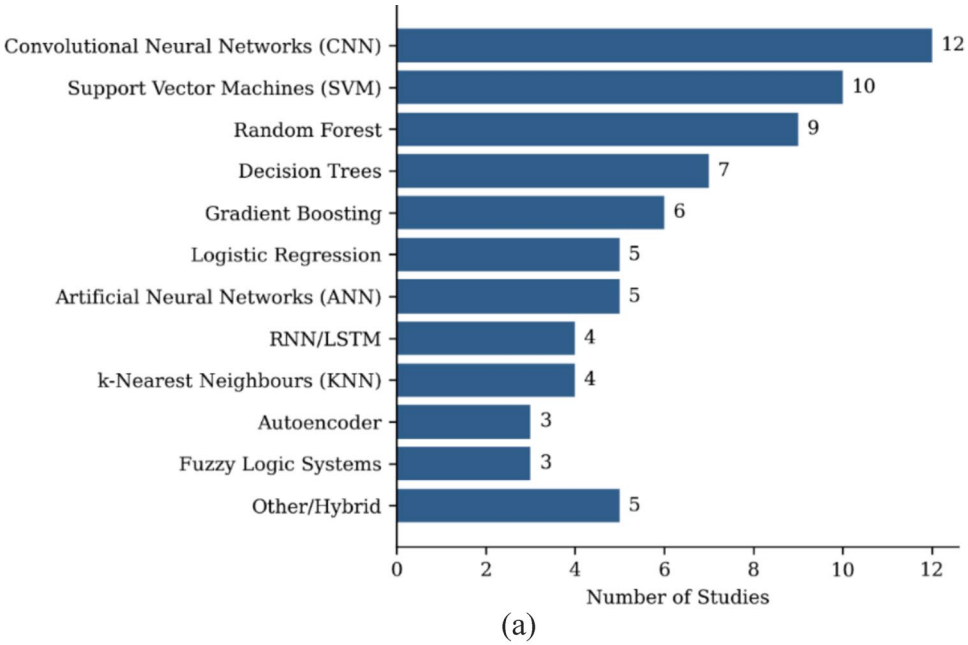
To support new researchers seeking to identify suitable datasets for AI and ML anomaly detection studies in rheumatology, a consolidated dataset reference table is provided in the new Appendix S5. This table summarises the principal datasets used across the 40 included studies, with columns for dataset name or institution of origin, country, data modality (imaging, biomarker, electronic health records, patient-reported outcomes, synthetic), the rheumatic conditions covered, approximate sample size, the studies that used the dataset, and whether the data are publicly accessible (for example, the Osteoarthritis Initiative, Kaggle biomechanical features) or restricted to institutional access. The table is intended as a quick-reference guide for researchers designing new anomaly detection studies in this domain.

Across the included studies, the intended clinical use cases clustered into a small number of decision points, most commonly early case identification or triage for inflammatory arthritis, monitoring for flare or changes in disease activity, imaging based identification of early structural or inflammatory changes, prediction of treatment response or non response to support therapy selection, and safety surveillance using laboratory or record based signals. This categorisation helps interpret performance metrics in context, because the evidentiary requirements and potential harms differ by use case, particularly where false positives could drive unnecessary testing or escalation, or where false negatives could delay treatment.

8 AI and ML methods

The 40 included studies employed a range of AI and ML approaches that can be organised into two broad families. Traditional machine learning methods, used in the majority of studies, included support vector machines, random forests, decision trees, gradient boosting, logistic regression, k-nearest neighbours, and fuzzy logic systems. These methods were typically applied to structured clinical data, laboratory profiles, and biomarker panels. Deep learning architectures were used in studies involving high-dimensional or image-based data and included convolutional neural networks for radiograph and thermal image classification, recurrent neural networks and LSTM models for sequential clinical records, autoencoders for representation learning, cost-sensitive neural networks for imbalanced class problems, and hybrid architectures combining multiple paradigms. The distribution of AI/ML method families across the included studies is shown in Fig. 2a. By learning paradigm, the majority of studies implemented supervised classification, training models on labelled case–control data to discriminate between disease-positive and comparator groups. A smaller subset employed unsupervised clustering or representation-based discovery to identify atypical patterns without pre-specified labels, while semi-supervised strategies were rarely described. The breakdown of learning paradigms is illustrated in Fig. 2b. Among traditional ML methods, ensemble approaches such as random forests and gradient boosting were frequently reported as top-performing classifiers for structured data tasks, while CNNs dominated imaging-based detection. However, direct comparison across studies is limited by heterogeneity in datasets, outcome definitions, and validation procedures.

Datasets across the 40 studies were categorised by primary data modality. Imaging data, including X-ray, MRI, thermal imaging, fluorescence optical imaging, and CT, were used in 15 studies. Electronic health records provided the primary data source in six studies, while laboratory and biomarker data, including serology, autoantibodies, and inflammatory markers, were used in five studies. Patient-reported



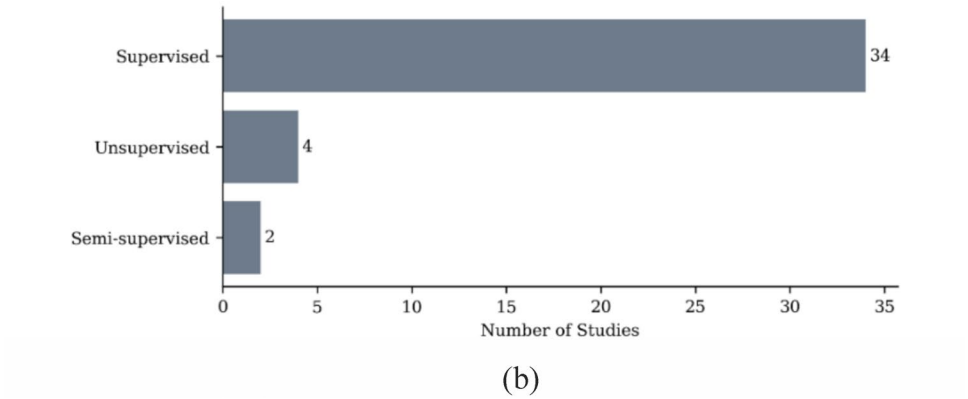


Fig. 2 Distribution of AI/ML method families (**a**) and learning paradigms (**b**)

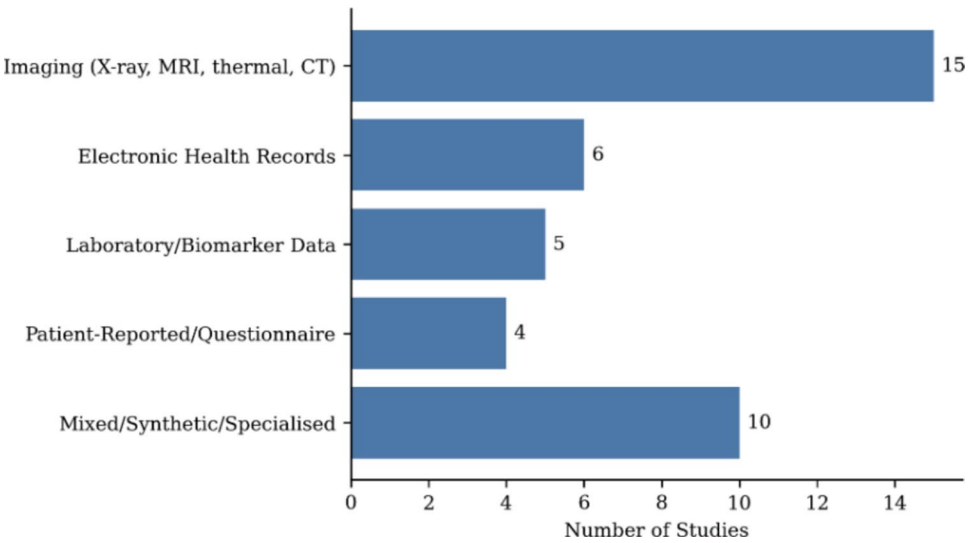


Fig. 3 Distribution of primary data modalities

data, questionnaire-based inputs, or symptom checker platforms were used in four studies. The remaining studies employed mixed, synthetic, or specialised cohort data. Figure 3 presents the distribution of primary data modalities. Evaluation metrics varied but were predominantly discrimination-based: accuracy, sensitivity, specificity, and AUC were the most frequently reported, followed by precision, recall, F1 score, and positive or negative predictive value. Calibration metrics and measures of clinical utility, such as decision curve analysis or net reclassification improvement, were rarely reported, representing a notable gap in the evidence base.

9 Risk of bias in studies

The risk of bias assessment utilising the PROBAST framework was conducted on 40 studies. The total risk of bias was low in 29 studies [7, 19, 29–34, 38–40, 42–44, 48–50, 52–54, 56, 56–58, 58–62]. These studies had all evaluation factors (Participants, Predictors, Outcome, and Analysis) rated as low risk. Figure 4 visualises the distribution of risk categories across studies based on the PROBAST evaluation.

Ten studies had a high risk of bias due to one or more evaluation factors rated as high risk. Corpuz et al. [17] showed high risk in the outcome and analysis domains due to inadequate outcome definitions and inappropriate analysis techniques. Singh et al. [35] also had high Outcome and Analysis risk due to a lack of validation and suboptimal statistical methods. Feldman et al. [36] presented high risk in the participant and outcome domains due to using non-representative samples and unclear definitions of outcomes. The study by Knitza et al. [41] had high risk in 'outcome' and 'analysis' due to inconsistent outcome definitions and analytical limitations. Mickley et al. [45] demonstrated unclear definitions across multiple domains, with high risk in the analysis due to incomplete data handling. Mills et al. [46] and Obayya et al. [47] were rated highly in the participant and analysis domains due to non-representative participant selection and a lack of proper validation methods. Rahimi et al. [18] and Betancourt-Hernández et al. [51] had high risk in participant and analysis for similar reasons. Sharon et al. [55] exhibited high risk in the participant and analysis domains, linked to non-randomised sampling and insufficient statistical rigour. The study of Folle et al. [37] was rated as unclear, due to inadequate information in the outcome domain, while all other factors were low risk. None of the 40 studies appeared to have the risk of other sources of bias not addressed in the prior assessment areas. Across the set, several small sample studies reporting very high discrimination metrics also lacked sufficient internal validation or events per predictor reporting, which increased concern for optimism bias within the PROBAST analysis domain. A summary of the potential sources of bias assessment is presented in Appendix S3.

10 TRIPOD-AI evaluations

The TRIPOD-AI assessment evaluated the selected studies' adherence to established reporting guidelines. This analysis examined 15 methodological and seven results-based criteria to determine the extent to which studies reported key aspects of predictive modelling. In contrast, a comprehensive analysis of the studies based on the TRIPOD-AI framework is provided in Appendix 4, and a summary is presented in Table 2. The results provide insight into the strengths and weaknesses in current reporting practices, highlighting areas where improvement is needed to enhance transparency, reproducibility, and applicability of predictive models. A visual summary of TRIPOD-AI reporting adherence across methodological and results domains is presented in Fig. 5.

The study design and data source reporting were well-documented across all studies, with all providing clear descriptions of their study design and specifying the data sources used [7, 17–19, 29–56, 56–58, 58–61]. Similarly, the study setting was reported in all cases, ensuring a strong foundation for understanding the context of each study. Eligibility criteria were also reported, with 95% of studies specifying the criteria used to include or exclude participants, allowing for a clear understanding of the target population. A small gap remained, as 5% of studies did not report their eligibility criteria, making it unclear how participants were selected [17, 31]. However, one of the most

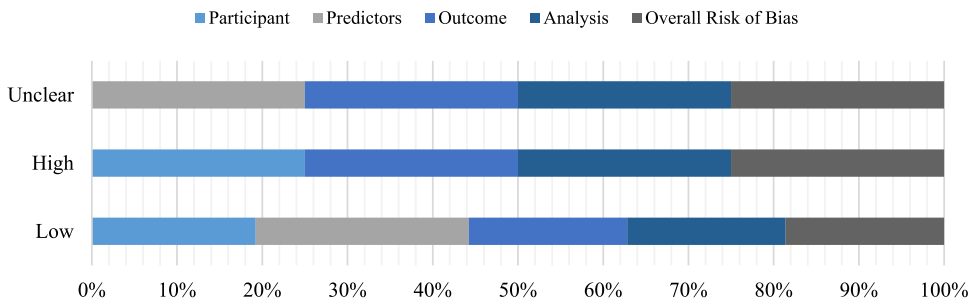


Fig. 4 The distribution of risk categories based on PROBAST framework

Table 2 TRIPOD-AI summary

Methods	Yes	No	Yes (%)	No (%)
Study design described	40	0	100	0
Data source specified	40	0	100	0
Study setting described	40	0	100	0
Eligibility criteria provided	38	2	95	5
Treatments described	0	40	0	100
Outcome clearly defined	40	0	100	0
Blinding of outcome assessment	11	29	27.5	72.5
Predictors clearly defined	40	0	100	0
Blinding of predictors	0	40	0	100
Sample size described	38	2	95	5
Missing data handling described	37	3	92.5	7.5
Methods described	40	0	100	0
Pre-processing described	40	0	100	0
Internal validation mentioned	34	6	85	15
Risk group creation described	19	21	47.5	52.5
Results				
Flow of participants described	38	2	95	5
Participant characteristics reported	32	8	80	20
Events per predictor reported	11	29	27.5	72.5
Unadjusted associations between predictors reported	32	8	80	20
Full model specification provided	32	8	80	20
Instructions on model usage included	11	29	27.5	72.5
Performance measures reported	37	3	92.5	7.5

significant shortcomings observed was the reporting of treatments, which was absent in all studies. This is likely due to the nature of predictive modelling studies, where treatment details may not be the primary focus.

Outcome definitions and predictor variables were consistently reported across studies, with all of them clearly defining their outcomes and specifying the predictor variables used in their models. However, the blinding of outcome assessment was only reported in 27.5% of the studies and was missing in 72.5% [17, 29–31, 33–40, 46–51, 53–56, 56–58, 60, 62]. Also, the blinding of predictor variables was completely missing from all the studies, raising potential concerns about the risk of bias in model development. Sample size descriptions were consistently provided in the studies, but were missing in two research [17, 31], but missing data handling methods were reported in only 75% of cases [7, 17–19, 30, 32–40, 42–51, 53–56, 56–58, 58–63], indicating that a significant portion of studies did not adequately address how they handled incomplete data [29, 31]. Reporting on statistical analysis methods was strong, with all studies detailing their methodological approaches and data preprocessing steps. Internal validation was performed in 85% of studies, reflecting a good level of model performance evaluation. However, 15% did not report any validation, raising concerns about the reliability and generalisability of their findings [17, 18, 29, 31, 35, 56]. Risk group creation was only reported in 70% of the studies, suggesting that many studies did not explicitly stratify risk levels, which could limit the applicability of the models in different clinical settings [18, 19, 32, 35, 44, 45, 47–50, 53–56, 56–58, 58, 59]. The flow of participants was well-described in 95% of studies, ensuring transparency in how data were processed and analysed, while 5% did not report this information [17, 31]. Model performance reporting was also strong, with 92.5% of the studies providing performance metrics and only 7.5% failing to report an evaluation for the models [35, 46, 47]. However, the reporting of participant characteristics, though relatively high, was still incomplete in 20% of studies [17, 29–31, 33–35, 60], which could limit the ability to assess the representativeness of the study populations.

One of the significant weaknesses observed was the underreporting of events per predictor, which was detailed in only 27.5% of studies [19, 32, 44, 50, 54, 56, 56–59, 61]. This is a critical limitation, as understanding the number of outcome events per predictor variable is essential for assessing the robustness and reliability of predictive models. Similarly, while 80% of studies provided a full model specification, a significant number still lacked comprehensive details

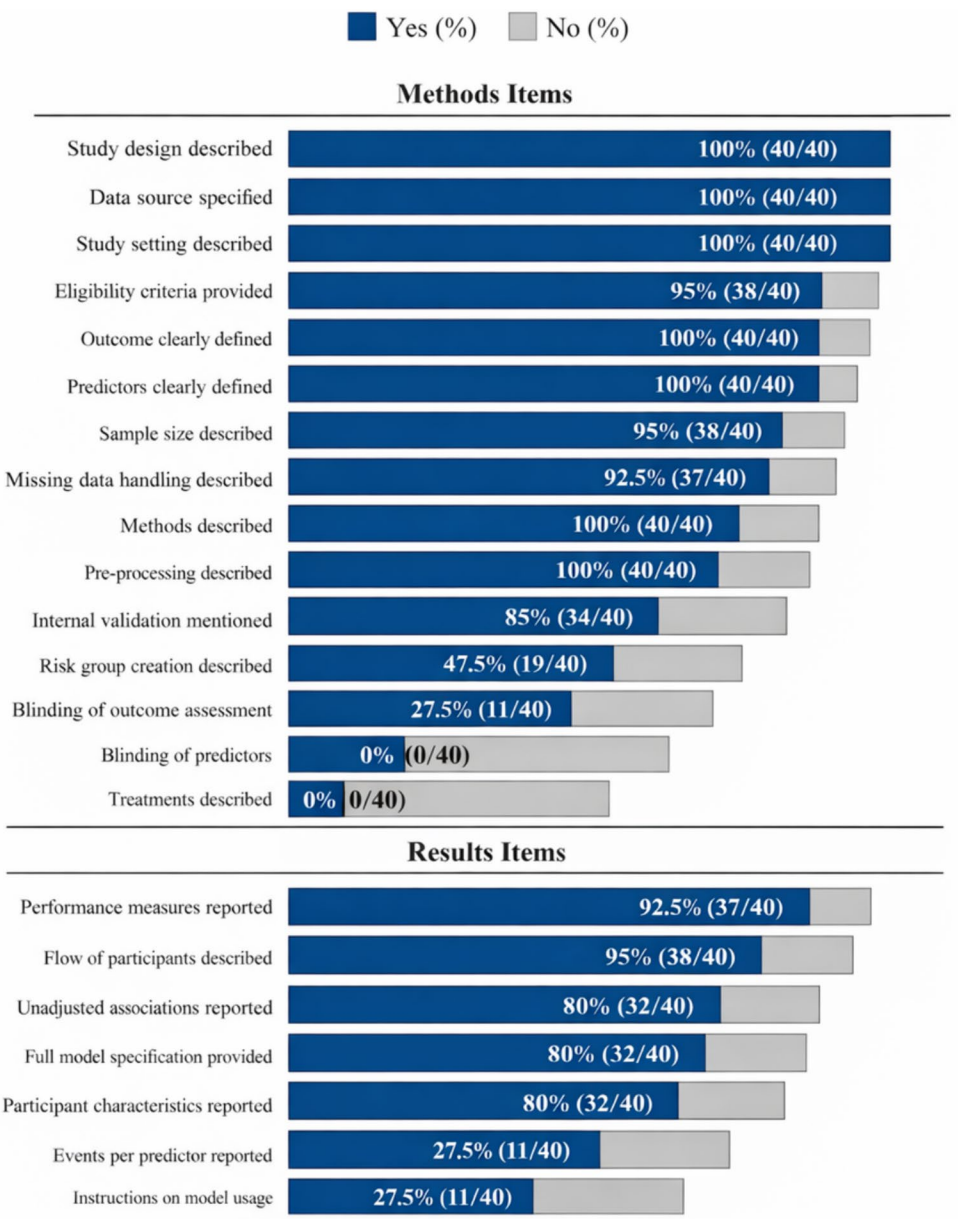


Fig. 5 TRIPOD-AI reporting adherence rates across methodological and results domains

[17, 29–35], which could hinder replication and external validation. Instructions on how to use the developed models were included in 27.5% of studies [18, 19, 44, 45, 47, 49, 50, 56, 56, 59, 61], and the rest failed to guide model implementation, limiting their applicability in real-world settings.

11 Discussion and implications for policy and research

This systematic review highlights both the promise and the present shortcomings of AI-driven anomaly detection in rheumatology. On one hand, these ML tools could augment clinical care by flagging atypical disease patterns or outliers that clinicians might otherwise overlook, enabling earlier diagnosis and intervention. In practical terms, a well-implemented anomaly detection system could flag patients with atypical patterns suggestive of early inflammatory arthritis or subclinical disease, enabling risk stratified triage for expedited rheumatology assessment and more targeted imaging or serologic work-up. By making risk status explicit, such tools could influence decisions about who is prioritised for assessment and investigation, and could help reduce diagnostic delay and avoidable progression when linked to predefined clinical actions. On the other hand, our findings clarify that most current models remain in early translation stages. This evidentiary immaturity continues to limit immediate clinical deployment. However, limited explainability remains a barrier to clinician trust and adoption [64]. Thus, a key practical implication is the need for

explainability and user-cantered design: clinicians should be able to interrogate AI outputs, and training programmes will be needed so that healthcare staff can interpret and act on AI findings appropriately [65]. Moreover, clear protocols must be in place to handle false positives or negatives from anomaly alerts to avoid alarm fatigue or missed diagnoses.

Clinical translation depends on placing anomaly detection within existing care pathways, clarifying who receives the alert, what the alert means, and what action follows. For example, in primary care and early arthritis services, an anomaly flag derived from symptom patterns, inflammatory markers, and serology could function as a triage signal that prompts expedited rheumatology review or early imaging, with the intended impact being shorter time to diagnosis and earlier initiation of disease modifying therapy. In established rheumatoid arthritis under treat to target monitoring, anomalies in longitudinal trends such as patient reported outcomes, inflammatory markers, and visit history could trigger earlier follow up, medication adjustment, or additional imaging, with the intended impact being fewer severe flares and prevention of cumulative joint damage.

An additional high value pathway is safety surveillance, where anomalies in routine laboratory monitoring or coded infection patterns could prompt medication review, repeat testing, or temporary treatment pause, with the intended impact being reduced serious adverse events. Importantly, these examples also highlight the evidence gap identified in this review: most studies report discrimination metrics, but few evaluate calibration, clinical usefulness, or real world impact. Future evaluations should therefore align validation and reporting with the intended pathway and include decision focused outcomes, such as change in time to specialist assessment, flare related urgent care, treatment escalation appropriateness, and adverse event detection, alongside careful management of alert burden and false alarm risk.

From a policy and public health perspective, the rise of AI in rheumatology raises essential considerations for regulators and health systems. Regulatory oversight is becoming increasingly pertinent as these algorithms move from the lab to the bedside. International efforts are underway to ensure AI tools meet safety and efficacy standards like traditional medical devices [66]. For example, the European Union's forthcoming AI Act (set to be enforced in 2026) will likely classify clinical AI systems as high-risk technologies, mandating strict requirements for transparency, risk management, and ongoing monitoring [67]. In the United States, the FDA has also signalled the need for oversight tailored to AI/ML-based medical software, emphasising the importance of a Total Product Life Cycle approach that includes post-market surveillance for algorithm performance drifts [68]. These policy moves reflect a broader recognition that continuous validation and governance of AI is essential; an algorithm is not a one-time product but may evolve, and its real-world behaviour must be tracked to ensure patient safety. Public health relevance is likewise significant. If anomaly detection algorithms can reliably identify early disease manifestations or predict flare-ups across large patient populations, they could enable earlier interventions on a broad scale, potentially reducing cumulative joint damage and disability in diseases like rheumatoid arthritis. Such preventative potential speaks to improved outcomes and possibly reduced healthcare costs by avoiding late-stage complications. However, realising these benefits for the public requires ensuring the technology is equitable and widely accessible.

Bias and equity issues in AI have been well documented. A model trained predominantly on data from one demographic or healthcare setting may perform poorly in others, risking exacerbation of health disparities [69]. In rheumatology, this could mean minority or underserved patient groups receiving less accurate predictions from AI, which is unacceptable from an equity standpoint. Therefore, policymakers and developers must mandate fairness checks and use diverse, representative data in developing anomaly detectors. Techniques like bias audits and dataset curation for inclusivity should become standard, alongside the adoption of FAIR data principles (Findable, Accessible, Interoperable, Reusable) to improve data sharing and quality [70]. From a policy angle, Federated may address previously noted data integration and representativeness gaps. Lastly, health authorities should consider guidelines for accountability, for instance, requiring that an AI's recommendation in patient care be traceable and that responsibility is assigned among clinicians, developers, and institutions. In summary, the translation of AI anomaly detection into public health impact will depend on regulatory frameworks that enforce transparency and efficacy, and on proactive measures to ensure all patient groups benefit equally from these innovations.

Despite the appeal of federated approaches, none of the 40 included studies used a federated or distributed learning architecture, and several concrete barriers explain this gap in rheumatology. First, the absence of standardised electronic health record schemas and rheumatology-specific common data models across centres makes harmonisation of predictors, outcomes and assessment instruments difficult, even before any federated training can begin. Second, regulatory and governance constraints for multi-centre, often multi-jurisdictional collaborations, including HIPAA, GDPR, the forthcoming European Union AI Act and national health data legislation, impose strict controls on the movement and processing of sensitive serology, imaging and free-text clinical notes. Third, the relatively small per-centre cohort sizes typical in rheumatology reduce the statistical benefit of federated aggregation relative to its engineering and governance overhead. Fourth, there is a lack of widely adopted, clinically validated federated learning infrastructure and trust frameworks within specialty practice; most current

federated learning toolkits are designed for radiology or genomics rather than for the heterogeneous data types found in rheumatology. Finally, the mix of modalities encountered in rheumatology, including structured biomarkers, imaging, free text and patient-reported outcomes, complicates the design of a single federated pipeline. Future work should therefore frame federated learning as one possible future direction among several, alongside well-governed centralised research consortia, synthetic-data approaches with rigorous fidelity testing, and prospective harmonisation initiatives at the rheumatology specialty level.

It is also notable that the majority of included studies were published after 2020, coinciding with the post-COVID-19 period during which interest in digital health and AI-based analytical tools accelerated markedly across medicine, including in rheumatology and related rehabilitation fields. This temporal clustering may partially reflect the pandemic-driven impetus to develop remote monitoring, automated triage, and data-driven diagnostic tools. Broader systematic work has documented the emerging role of artificial intelligence in immune-mediated rheumatic diseases and in understanding long COVID sequelae, further situating the present review within a rapidly evolving research landscape.

Relatedly, many datasets used to train the reviewed AI and ML models were collected during or shortly after the COVID-19 pandemic, raising the possibility of pandemic-related biases. Disrupted care pathways, delayed diagnoses, changes in healthcare utilisation patterns, and altered biomarker distributions during and after the pandemic may have influenced the composition and characteristics of training data in ways that do not reflect typical pre-pandemic clinical practice. Models trained on such data may therefore capture pandemic-era artefacts rather than stable disease patterns. To mitigate these risks, developers should document data collection periods, assess temporal stability of model performance, and where feasible validate models on both pre-pandemic and post-pandemic cohorts.

Specifically, none of the 40 included studies explicitly tested for temporal stability across pre-COVID and COVID/post-COVID strata, for example by training on pre-2020 data and evaluating on 2020 to 2022 data, nor did any perform formal dataset-shift diagnostics on biomarker distributions, diagnostic timing or referral patterns. Future developers should therefore (i) perform chronologically split internal validation with pre-2020, 2020 to 2022 and post-2022 strata; (ii) apply dataset-shift detection on key predictors and outcome incidence; (iii) plan periodic model recalibration and prospective monitoring of calibration drift after deployment; and (iv) report data collection windows transparently alongside performance metrics, in line with TRIPOD-AI guidance. These steps are particularly important for distinguishing genuine post-acute rheumatic disease phenomena from transient pandemic-era artefacts in the training data.

Our review underscores the pressing need for greater transparency, rigorous evaluation standards, and demonstration of generalisability moving forward. The clinical and methodological complexities contribute to fragmented datasets and constrained external validation. In rheumatology, external validation is additionally constrained by pragmatic barriers. Early detection often requires longitudinal pre diagnostic data that are not routinely curated across institutions, while outcome definitions can vary between settings due to differences in classification criteria, coding practices, and clinician documentation. Imaging based studies face additional challenges linked to acquisition protocols and annotation availability. These factors contribute to dataset silos and limit transportability testing, reinforcing the need for multicentre collaborations and, where data sharing is restricted, federated or privacy preserving validation approaches. Consistent with findings in other areas of medical AI, many published models likely suffer from optimistic bias or overfitting, highlighting that the field as a whole must improve its methodological rigour [71]. To operationalise the identified reporting gaps, adherence to TRIPOD-AI for prediction model studies [72], CONSORT-AI for clinical trials of AI interventions [73], and SPIRIT-AI for trial protocols [74] is essential and encouraged. These guidelines promote transparent reporting of key details, facilitating reproducibility and critical appraisal. Likewise, stage-specific guidance like DECIDE-AI [75] provides a roadmap for early-phase prospective studies of AI decision support tools, bridging the gap between retrospective model development and large-scale trials. Embracing such standards will improve the quality of evidence generated and help the field move beyond proof-of-concept towards clinically robust applications. Moreover, future research should focus on prospective evaluations, such as integrating an AI tool into a rheumatology clinic on a trial basis to observe how it performs and influences care in real time. Such deployment studies, possibly in cluster randomised trials or adaptive trials, will shed light on not just accuracy, but also workflow impact, clinician behaviour changes, and patient outcomes. They will also help uncover unanticipated failure modes or safety issues not evident in retrospective analyses.

Clinical implications of AI-based anomaly detection also extend to complex diagnostic scenarios that are increasingly common in current practice. For example, distinguishing genuine disease flares from post-infectious or post-COVID inflammatory manifestations presents a significant challenge, as overlapping clinical features, laboratory profiles, and imaging findings may confound automated classification. None of the reviewed models were specifically trained or validated to distinguish between these presentations, representing a gap that future development should address through clinical phenotyping that accounts for overlapping

aetiologies and temporal disease trajectories in the post-pandemic era.

The generalisability of AI and ML models across healthcare systems and geographic regions also warrants more explicit attention than is currently provided in the reviewed studies. Differences in diagnostic pathways, classification criteria, data availability, coding practices, imaging acquisition protocols, and patient demographics between health systems may substantially limit external validity. Models validated in a single centre or country cannot be assumed to transfer reliably to other settings without local recalibration and prospective validation in the target population. Addressing this will require multicentre and multinational collaboration, standardised data definitions, and, where data sharing is restricted, federated or privacy-preserving approaches to model validation.

From a technical perspective, the evidence suggests that model performance varies meaningfully by method family and application context. For imaging-based anomaly detection, convolutional neural networks and their variants consistently reported the highest discrimination metrics, reflecting their established capacity for visual pattern recognition. For structured clinical and laboratory data, ensemble methods such as random forests and gradient boosting frequently outperformed single classifiers, benefiting from their robustness to feature noise and ability to handle mixed data types. Dataset size and diversity also influenced reported performance: studies using large electronic health record databases generally reported more conservative but potentially more generalisable estimates, while studies on small, curated cohorts often reported very high accuracy that may not replicate in broader populations. These patterns reinforce the importance of interpreting reported performance in the context of study scale, validation design, and intended clinical deployment.

Data scarcity and fragmentation remain practical barriers to robust development and external validation in rheumatology. Methods from adjacent biomedical AI domains that explicitly target low data learning may therefore be valuable, including meta learning and few shot approaches, graph based representation learning, and large scale text mining resources. While these approaches have been developed largely in drug discovery and molecular modelling contexts, their underlying strategies could be adapted to clinical anomaly detection to improve sample efficiency, portability, and reuse across sites, provided they are evaluated within appropriate clinical workflows and under transparent reporting and governance requirements [76–79].

The need for methodological standards and stewardship in AI evaluation cannot be overstated. Just as evidence-based medicine in therapeutics relies on consensus guidelines and regulatory standards, AI in healthcare should be evaluated against accepted benchmarks before widespread adoption. Efforts to develop specialised risk-of-bias assessment tools and standardised outcome reporting frameworks will be crucial to objectively compare studies and avoid hyped conclusions. In rheumatology, where datasets can be smaller and diseases heterogeneous, following standards will help ensure that novel AI models are accurate in a research setting and reliable in diverse clinical scenarios. Future evaluations should extend beyond technical metrics to consider clinical and cost-effectiveness endpoints. Finally, our analysis suggests that AI-based anomaly detection in rheumatology is on the cusp of translating into meaningful clinical advances. Still, it will require sustained collaborative efforts to navigate the translational gaps. The intersection of rheumatology and AI is a microcosm of the broader digital transformation in medicine; it holds great promise, but realising that promise responsibly will demand high standards of evidence, thoughtful integration into practice, and continuous oversight. Embracing these principles will help clinicians and policymakers harness such algorithms as trustworthy allies in improving care, rather than unvetted gadgets, moving the field toward a future where AI-driven insights support better health for all.

12 Limitations

This systematic review has several limitations. Many of the included studies exhibited methodological weaknesses. Outcome definitions were often ambiguous or inconsistent across studies, and blinding of outcome assessments was limited or not reported, introducing potential bias [80]. Inadequate external validation was a common issue; most prediction models were evaluated only on training or internal data, with few being tested on independent cohorts, leading to a high risk of overfitting and optimistic performance estimates. Additionally, more than half of the developed models included in this study were derived with fewer than the recommended outcome events per predictor variable (or failed to report this metric), raising concerns about model stability, and contributing to a high risk of bias in the published models. Furthermore, there were significant deficiencies in how the studies were reported, as highlighted by our TRIPOD and evaluations. Overall, transparency was suboptimal; crucial details needed for appraisal and replication were often missing. For instance, few papers fully adhered to reporting guidelines for key items such as outcome definitions, participant flow diagrams, handling of missing data, or model performance measures. In line with findings from other reviews [4, 81], median adherence to TRIPOD items in similar ML studies has been around only 40–50%, and a majority of models were judged to be at high risk of bias. Moreover, only a minority provided access to their algorithms or detailed their input variables, output formats,

and model coefficients, hindering independent validation and clinical application of the proposed models. This lack of implementation guidance and poor reporting of model details undermine the reproducibility and usability of the research findings.

Another concern is the limited diversity of study populations and the consequent generalisability issues. Most models were developed and tested on relatively small, homogeneous patient cohorts, underrepresenting certain demographic and geographic groups. This lack of diverse population data can inflate apparent accuracy and limit the applicability of models to broader rheumatology populations [82]. Small sample sizes and narrow development settings were a frequent problem, a point noted in other ML studies wherein performance is often overestimated due to restricted or biased datasets [82, 83]. Consequently, many models may perform poorly when deployed in different clinical contexts or under-represented subpopulations. In our review, guidance on generalisability was seldom provided in the original studies. This raises concerns that the 'streamlined' anomaly detection methods reported in the literature may not be directly generalisable across other study settings, especially given that external validations often revealed attenuated performance outside the development sample.

Another critical research direction is improving the explainability and interpretability of AI models. Black-box algorithms, as noted, pose a barrier to trust; thus, techniques must make anomaly detection results more interpretable to clinicians. Explainable AI methods (such as attention maps for images or feature importance scores for predictions) should be incorporated and evaluated for their efficacy in conveying useful information to end-users. Doing so will improve clinician acceptance and allow better scrutiny of the model's reasoning, which is vital for safety. In tandem, researchers should report performance in human-cantered terms; rather than exclusively reporting statistical metrics, discussing how the AI's outputs could translate to clinical decisions or patient benefits will make the research more clinically meaningful. We also recommend that future work consider implementation science aspects early on; involving rheumatologists, patients, and IT in co-designing AI tools will address usability challenges and ensure that deployment, when it happens, is smooth. Factors like interoperability with electronic health record systems, and minimising disruption to clinical workflows, are often afterthoughts in research studies but are decisive for real-world adoption.

In addition to limitations of the primary studies, there are scope and timeframe limitations to consider. We deliberately narrowed the focus of this review to anomaly detection applications in rheumatology, which led us to exclude certain categories of studies. In particular, we did not include studies primarily examining genomic predictors (e.g., genetic risk scores), pharmacological treatment response modelling, or disease flare prediction algorithms. These are important research areas in rheumatology AI, but they fall outside our predefined scope. Consequently, our findings do not generalise to those domains; for example, our conclusions about model performance and reporting may differ from those in pharmacogenomics studies or flare forecasting. Additionally, we limited the publication timeframe to 2000–2025. We also limited inclusion to English language publications, which may have excluded relevant studies from non English speaking settings and may bias the evidence base toward regions and health systems more represented in English language journals. While this was intended to capture the era of modern ML applications, it inherently omits any earlier pioneering work. The breadth of the time also means that methodological standards evolved during the study period, potentially contributing to heterogeneity in quality. We recognise that these review-level restrictions (in scope and time frame) may have left some gaps in the surveyed literature. Despite these limitations, we believe the insights gained are valuable for understanding the current state of anomaly detection in rheumatology, and this study underscores the need for future research to address these noted weaknesses in both primary studies and review methodologies.

Additionally, the restriction to English-language publications may have excluded relevant studies from non-English-speaking settings and may bias the evidence base toward regions and health systems that are more prominently represented in English-language journals. Future updates of this review should consider incorporating multilingual searching and screening to improve the representativeness and global generalisability of the evidence synthesis.

13 Conclusions

This systematic review highlights the potential of AI and ML to support anomaly detection in rheumatology, particularly by enabling earlier recognition of atypical disease patterns and more timely clinical review. However, the current evidence base remains constrained by limitations in transparency, validation rigour, and generalisability, with recurrent risks of bias linked to unclear outcome definitions, limited blinding, and insufficient external validation. To translate these tools into safe and equitable clinical practice, future work must align development with regulatory expectations for high risk clinical AI, and must address ethical requirements including fairness, accountability, and minimising harm from false alerts. We therefore call for multicentre

collaborations that enable prospective evaluation in real workflows, supported by standardised, well documented datasets that represent diverse populations, and consistent reporting using established frameworks. These steps are essential to move from promising performance metrics to demonstrable patient benefit and trustworthy deployment. Addressing these methodological and reporting gaps is essential for successful translation from model development to clinical deployment. Ultimately, AI and ML hold transformative potential in rheumatology, promising enhanced diagnostic precision, improved clinical decision-making, and more personalised patient management. To achieve these outcomes, a collaborative approach involving clinicians, data scientists, patients, and healthcare stakeholders is essential. By systematically addressing current limitations and promoting transparent, reproducible research practices, the rheumatology community can leverage AI effectively to improve patient outcomes and advance clinical care.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s41060-026-01191-w.

Author contribution AIM, VQ, and HRP contributed to the conception, design, data collection, analysis, and drafting of the manuscript. MP, LCW, CW, MOS, and SM provided supervision, critical feedback, and advisory support throughout the study and manuscript preparation. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work was supported by the New Zealand Health Research Council (Grant Number 23/919).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Smolen, J.S., Aletaha, D., McInnes, I.B.: Rheumatoid arthritis. Lancet **388**(10055), 2023–2038 (2016). https://doi.org/10.1016/S0140-6736(16)30173-8
2. Scott, D.L., Wolfe, F., Huizinga, T.W.J.: Rheumatoid arthritis. Lancet **376**(9746), 1094–1108 (2010). https://doi.org/10.1016/S0140-6736(10)60826-4
3. Momtazmanesh, S., Nowroozi, A., Rezaei, N.: Artificial intelligence in rheumatoid arthritis: current status and future perspectives: a state-of-the-art review. Rheumatol Ther **9**(5), 1249–1304 (2022). https://doi.org/10.1007/s40744-022-00475-4
4. Mendoza-Pinto, C., et al.: Machine learning in the prediction of treatment response in rheumatoid arthritis: A systematic review. Semin. Arthritis Rheum. **68**, 152501 (2024). https://doi.org/10.1016/j.semarthrit.2024.152501
5. F. Afrazeh and M. Shomalzadeh, "Revolutionizing Arthritis Care with Artificial Intelligence: A Comprehensive Review of Diagnostic, Prognostic, and Treatment Innovations," *International Journal of Applied Data Science in Engineering and Health,* 2024.
6. Imtiaz, M., Shah, S.A.A.: A review of arthritis diagnosis techniques in artificial intelligence era: Current trends and research challenges. Neurosci. Inform. **2**(4), 100079 (2022). https://doi.org/10.1016/j.neuri.2022.100079
7. A. I. Maghsoodi, D. Harries, and V. Quincey, "Shallow Learning for Predictive Blood Test Anomaly Detection: Case Study for Rheumatic Diseases," in *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 4–6 Dec. 2023 2023, pp. 1–7, https://doi.org/10.1109/CSDE59766.2023.10487145.
8. Bragazzi, N.L., Bridgewood, C., Watad, A., Damiani, G., Kong, J.D., McGonagle, D.: Harnessing big data, smart and digital technologies and artificial intelligence for preventing, early intercepting, managing, and treating psoriatic arthritis: insights from a systematic review of the literature. Front. Immunol. (2022). https://doi.org/10.3389/fimmu.2022.847312
9. Stafford, I.S., Kellermann, M., Mossotto, E., Beattie, R.M., MacArthur, B.D., Ennis, S.: A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit Med **3**, 30 (2020). https://doi.org/10.1038/s41746-020-0229-3
10. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Computer Sci. **2**(3), 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
11. Danieli, M.G., et al.: Machine learning application in autoimmune diseases: State of art and future prospectives. Autoimmun. Rev. **23**(2), 103496 (2024). https://doi.org/10.1016/j.autrev.2023.103496
12. Avramidis, G.P., Avramidou, M.P., Papakostas, G.A.: Rheumatoid arthritis diagnosis: deep learning vs. humane. Appl. Sci. **12**(1), 10 (2021). https://doi.org/10.3390/app12010010
13. AltikardeŞ, Z.A., Canayaz, E., ÜNsal, A.: Machine learningdeep learning in rheumatological screening a systematic review. Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi **16**(3), 940–969 (2023). https://doi.org/10.18185/erzifbed.1211547
14. Liberati, A., et al.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ **339**, b2700 (2009). https://doi.org/10.1136/bmj.b2700
15. Shamseer, L., et al.: Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ **350**, g7647 (2015). https://doi.org/10.1136/bmj.g7647

16. M. Cumpston *et al.*, (2019) "Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions, *Cochrane Database Syst Rev,* 10: 000142, https://doi.org/10.1002/14651858.ED000142.
17. M. J. S. Corpuz, J. Avestro, F. B. Calanda, R. C. Morco, J. M. Angeles, and C. M. S. Alcala, "eRheumatologist: Mobile-based Expert System for Rheumatology Utilizing Fuzzy Logic Algorithm," in *ACM International Conference Proceeding Series*, 2017: Association for Computing Machinery, pp. 153–157, https://doi.org/10.1145/3168390.3168428.
18. Rahimi, S.A., et al.: Quantum-inspired interpertable AI-empowered decision support system for detection of early-stage rheumatoid arthritis in primary care using scarce dataset. Mathematics **10**(3), 496 (2022). https://doi.org/10.3390/math10030496
19. Zhou, S.-M., et al.: Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. PLoS ONE **11**(5), e0154515 (2016)
20. G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement," *Circulation,* vol. 131, no. 2, pp. 211-219, 2015
21. Moons, K.G., et al.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. **162**(1), W1–W73 (2015)
22. Moons, K.G.M., et al.: PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ **388**, e082505 (2025). https://doi.org/10.1136/bmj-2024-082505
23. Fukae, J., et al.: Convolutional neural network for classification of two-dimensional array images generated from clinical information may support diagnosis of rheumatoid arthritis. Sci. Rep. **10**(1), 5648 (2020). https://doi.org/10.1038/s41598-020-62634-3
24. S. Shanmugam and J. Preethi, "Design of rheumatoid arthritis predictor model using machine learning algorithms," in *SpringerBriefs in Applied Sciences and Technology*: Springer Verlag, 2018, pp. 67–77.
25. Bae, J.-M.: Narrative reviews. Epidemiol. Health **36**, e2014018 (2014)
26. Peters, M.D., et al.: Scoping reviews: reinforcing and advancing the methodology and application. Syst. Rev. **10**, 1–6 (2021)
27. Mengist, W., Soromessa, T., Legese, G.: Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX **7**, 100777 (2020)
28. Kedra, J., et al.: Current status of use of big data and artificial intelligence in RMDs: a systematic literature review informing EULAR recommendations. RMD Open **5**(2), e001004 (2019)
29. Brahim, A., et al.: A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. Comput. Med. Imaging Graph. **73**, 11–18 (2019)
30. Shuai, W., et al.: Rapid diagnosis of rheumatoid arthritis and ankylosing spondylitis based on Fourier transform infrared spectroscopy and deep learning. Photodiagnosis Photodyn. Ther. **45**, 103885 (2024)
31. Priyadarshini, K., Sikkandar, M.Y., AlDuraywish, A., Alqahtani, T.M.: Integrating relational and sequential information for enhanced detection of autoimmune disorders with relational neural networks and long short-term memory networks. Biomed. Signal Process. Control **96**, 106495 (2024). https://doi.org/10.1016/j.bspc.2024.106495
32. Ahalya, R.K., Almutairi, F.M., Snekhalatha, U., Dhanraj, V., Aslam, S.M.: RANet: a custom CNN model and quanvolutional neural network for the automated detection of rheumatoid arthritis in hand thermal images. Sci. Rep. **13**(1), 15638 (2023). https://doi.org/10.1038/s41598-023-42111-3
33. Bai, L., Zhang, Y., Wang, P., Zhu, X., Xiong, J.-W., Cui, L.: Improved diagnosis of rheumatoid arthritis using an artificial neural network. Sci. Rep. **12**(1), 9810 (2022)
34. Chin, C.-Y., Hsieh, S.-Y., Tseng, V.S.: eDRAM: effective early disease risk assessment with matrix factorization on a large-scale medical database: a case study on rheumatoid arthritis. PLoS ONE **13**(11), e0207579 (2018)
35. U. V. Singh, E. Gupta, and T. Choudhury, "Detection of rheumatoid arthritis using machine learning," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019: IEEE, pp. 25–29.
36. Feldman, C.H., et al.: Supplementing claims data with electronic medical records to improve estimation and classification of rheumatoid arthritis disease activity: a machine learning approach. ACR Open Rheumatology **1**(9), 552–559 (2019). https://doi.org/10.1002/acr2.11068
37. Folle, L., et al.: Deep learning-based classification of inflammatory arthritis by identification of joint shape patterns—how neural networks can tell us where to "deep dive" clinically. Front. Med. **9**, 850552 (2022). https://doi.org/10.3389/fmed.2022.850552
38. Geng, L., et al.: Prediction of diagnosis results of rheumatoid arthritis patients based on autoantibodies and cost-sensitive neural network. Clin. Rheumatol. **41**(8), 2329–2339 (2022). https://doi.org/10.1007/s10067-022-06109-y
39. Gräf, M., et al.: Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol. Int. **42**(12), 2167–2176 (2022). https://doi.org/10.1007/s00296-022-05202-4
40. Heard, B.J., Rosvold, J.M., Fritzler, M.J., El-Gabalawy, H., Wiley, J.P., Krawetz, R.J.: A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers. J. Royal Soc. Interface **11**(97), 20140428 (2014)
41. Knitza, J., et al.: Diagnostic accuracy of a mobile ai-based symptom checker and a web-based self-referral tool in rheumatology: multicenter randomized controlled trial. J. Med. Internet Res. **26**, e55542 (2024). https://doi.org/10.2196/55542
42. Lin, C., et al.: Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS ONE **8**(8), e69932 (2013)
43. Mate, G.S., Kureshi, A.K., Singh, B.K.: An efficient cnn for hand X-ray classification of rheumatoid arthritis. J. Healthcare Eng. **2021**, 6712785 (2021). https://doi.org/10.1155/2021/6712785
44. Maarseveen, T.D., et al.: Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. JMIR Med. Inform. **8**(11), e23930 (2020)
45. Mickley, J.P., et al.: A stepwise approach to analyzing musculoskeletal imaging data with artificial intelligence. Arthrit. Care Res. **76**(5), 590–599 (2024). https://doi.org/10.1002/acr.25260
46. Mills, G.A., Dey, D., Kassim, M., Yiwere, A., Broni, K.: Diagnostic tool for early detection of rheumatic disorders using machine learning algorithm and predictive models. BioMedInformatics **4**(2), 1174–1201 (2024). https://doi.org/10.3390/biomedinformatics4020065
47. Obayya, M., et al.: Artificial intelligence driven biomedical image classification for robust rheumatoid arthritis classification. Biomedicines **10**(11), 2714 (2022). https://doi.org/10.3390/biomedicines10112714
48. Peng, Y., Huang, X., Gan, M., Zhang, K., Chen, Y.: Radiograph-based rheumatoid arthritis diagnosis via convolutional neural network. BMC Med Imag. **24**(1), 180 (2024). https://doi.org/10.1186/s12880-024-01362-w

49. Phatak, S., Chakraborty, S., Goel, P.: "Computer vision detects inflammatory arthritis in standardized smartphone photographs in an Indian patient cohort. Front. Med. **10**, 1280462 (2023). https://doi.org/10.3389/fmed.2023.1280462
50. Prabavathy, K., Nalini, M.: Deep learning enabled whale optimization algorithm for accurate prediction of RA disease. Salud Cienc. Tecnol. - Ser. Conf. **3**, 652 (2024). https://doi.org/10.56294/sctconf2024652
51. Betancourt-Hernández, M., Viera-López, G., Serrano-Muñoz, A.: Automatic diagnosis of rheumatoid arthritis from hand radiographs using convolutional neural networks. Rev. Cuba. Fis. **35**(1), 39–43 (2018)
52. Reed, M., Le Souëf, T., Rampono, E.: Pilot study of a machine-learning tool to assist in the diagnosis of hand arthritis. Int. Med. J. **52**(6), 959–967 (2022). https://doi.org/10.1111/imj.15173
53. Hirano, T., et al.: Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. Rheumatol. Adv. Pract. **3**(2), rkz047 (2019)
54. Rothe, F., et al.: Fluorescence optical imaging feature selection with machine learning for differential diagnosis of selected rheumatic diseases,". Front. Med. **10**, 1228833 (2023). https://doi.org/10.3389/fmed.2023.1228833
55. H. Sharon, I. Elamvazuthi, C. K. Lu, S. Parasuraman, and E. Natarajan, "Classification of Rheumatoid Arthritis using Machine Learning Algorithms," in *2019 IEEE Student Conference on Research and Development, SCOReD 2019*, 2019: Institute of Electrical and Electronics Engineers Inc., pp. 345–350, https://doi.org/10.1109/SCORED.2019.8896344.
56. S. Sundaramurthy, C. Saravanabhavan, and P. Kshirsagar, "Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches," in *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, 2020: Institute of Electrical and Electronics Engineers Inc., pp. 17–21, https://doi.org/10.1109/DASA51403.2020.9317253. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100574783&doi=10.1109%2fDASA51403.2020.9317253&partnerID=40&md5=d57f0f36c321f25311181938da8995dd
57. Üreten, K., Erbay, H., Maraş, H.H.: Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. Clin. Rheumatol. **39**(4), 969–974 (2020). https://doi.org/10.1007/s10067-019-04487-4
58. Üreten, K., Maraş, H.H.: Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. J. Digit. Imag. **35**(2), 193–199 (2022). https://doi.org/10.1007/s10278-021-00564-w
59. Vera Cruz, G., et al.: Machine learning reveals the most important psychological and social variables predicting the differential diagnosis of rheumatic and musculoskeletal diseases. Rheumatol. Int. **42**(6), 1053–1062 (2022). https://doi.org/10.1007/s00296-021-04916-1
60. Wang, H.J., 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 **10**(6), 1355 (2022). https://doi.org/10.3390/biomedicines10061355
61. Ahalya, R.K., Snekhalatha, U., Krishnan, P.T.: Hybrid AI model for the detection of rheumatoid arthritis from hand radiographs. Biomed. Eng. – Appl. Bas. Commun. **35**(3), 2350009 (2023). https://doi.org/10.4015/S1016237223500096
62. Tarakci, F., Ozkan, I.A., Yilmaz, S., Tezcan, D.: Diagnosing rheumatoid arthritis disease using fuzzy expert system and machine learning techniques. J. Intell. Fuzzy Syst. **44**(4), 5543–5557 (2023). https://doi.org/10.3233/JIFS-221582
63. Ramasamy, U., Santhoshkumar, S.: Benchmark datasets and real-time autoimmune disease dataset analysis using machine learning algorithms with implementation, analysis and results. J. Intell. Fuzzy Syst. **45**(2), 2449–2463 (2023). https://doi.org/10.3233/JIFS-224115
64. Altıkardeş, Z.A., Canayaz, E., Ünsal, A.: Machine learningdeep learning in rheumatological screening a systematic review. Erzincan Univ. J. Sci. Technol. **16**(3), 940–969 (2023)
65. Kedra, J., Davergne, T., Braithwaite, B., Servy, H., Gossec, L.: Machine learning approaches to improve disease management of patients with rheumatoid arthritis: review and future directions. Expert Rev. Clin. Immunol. **17**(12), 1311–1321 (2021)
66. Sequí-Sabater, J.M., Benavent, D.: Artificial intelligence in rheumatology research: what is it good for? RMD Open **11**(1), e004309 (2025)
67. J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, P. Q. Consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **20**, 1–9 (2020)
68. M. o. Health, "Precision health: exploring opportunities and challenges to predict, prevent, diagnose, and treat health needs more precisely in Aotearoa New Zealand," Wellington, New Zealand, August 2023 2023.
69. E. Commission. "AI Act: Regulatory Framework for Artificial Intelligence." European Commission. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed 14 April 2025).
70. Abràmoff, M.D., et al.: Considerations for addressing bias in artificial intelligence for health equity. NPJ Dig. Med. **6**(1), 170 (2023)
71. Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G.: Potential biases in machine learning algorithms using electronic health record data. JAMA Intern. Med. **178**(11), 1544–1547 (2018)
72. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci Data **3**(1), 1–9 (2016)
73. C. L. A. Navarro *et al.*, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review," *bmj,* vol. 375, 2021.
74. G. S. Collins *et al.*, "TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *bmj,* vol. 385, 2024.
75. Liu, X., et al.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health **2**(10), e537–e548 (2020)
76. Shi, Y., et al.: Advancing precision rheumatology: applications of machine learning for rheumatoid arthritis management. Front. Immunol. **15**, 1409555 (2024)
77. Moon, S.J., Lee, S., Hwang, J., Lee, J., Kang, S., Cha, H.-S.: Performances of machine learning algorithms in discriminating sacroiliitis features on MRI: a systematic review. RMD Open **9**(4), e003783 (2023)
78. Lv, Q., Chen, G., He, H., et al.: TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining. Chem. Sci. **14**(39), 10684–10701 (2023)
79. Lv, Q., Chen, G., Yang, Z., Zhong, W., Chen, C.Y.-C.: Meta learning with graph attention networks for low-data drug discovery. IEEE Trans. Neural Netw. Learn. Syst. **33**(7), 3134–3144 (2022)
80. Rivera, S.C., et al.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. The Lancet Digital Health **2**(10), e549–e560 (2020)
81. B. Vasey *et al.*, "Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *bmj,* vol. 377, 2022.

82. Dhiman, P., et al.: Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn. Progn. Res. **6**(1), 13 (2022)
83. Munguía-Realpozo, P., et al.: Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: A systematic review. Autoimmun. Rev. **22**(5), 103294 (2023)