Review

Diagnostic Performance of Machine Learning for Systemic Lupus Erythematosus: Systematic Review and Meta-Analysis

Bingduo Wang[1], MD; Zichao Wang[2], MPH; Yang Liu[3], MD; Fangfang Ge[4], PhD

[1]Department of Anesthesiology and Intensive Care, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2]Department of Epidemiology and Statistics, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China
[3]Department of Proctology, The First Affiliated Hospital of Xinjiang Medical University, Xinjiang, China
[4]Department of Hematology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Corresponding Author:

Fangfang Ge, PhD
Department of Hematology
The First Affiliated Hospital of Zhengzhou University
No. 1, Jianshe East Road, Erqi District
Zhengzhou
China
Phone: 86 15067607819
Email: Gefangfang3@gmail.com

Abstract

Background: Early and accurate diagnosis of systemic lupus erythematosus (SLE) and its organ involvement is essential. Previous reviews of machine learning (ML) in SLE combined heterogeneous tasks and validation strategies and may have overinterpreted model performance.

Objective: This study evaluated the diagnostic performance of ML and deep learning (DL) models for 3 clinically distinct SLE-related tasks: SLE classification or diagnosis, lupus nephritis (LN) diagnosis, and neuropsychiatric systemic lupus erythematosus (NPSLE) discrimination. We also assessed methodological quality and certainty of evidence.

Methods: PubMed, Embase, Cochrane Library, Web of Science, and IEEE Xplore were searched from January 2014 to April 2026. Eligible peer-reviewed diagnostic accuracy studies developed or validated ML or DL models for 1 of the 3 prespecified tasks, used an accepted reference standard, and provided data for a 2×2 contingency table. Bivariate random-effects meta-analyses with the Hartung-Knapp-Sidik-Jonkman adjustment were used to pool sensitivity and specificity. We reported 95% prediction intervals (PIs), assessed risk of bias using the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence tool (QUADAS-AI; Viknesh Sounderajah [Imperial College London]), and evaluated certainty of evidence using the Grading of Recommendations Assessment, Development, and Evaluation framework for diagnostic test accuracy.

Results: Twenty-nine studies were included: 17 for SLE classification, 5 for LN diagnosis, and 7 for NPSLE discrimination. In the primary task-stratified analysis, pooled sensitivity was 0.91 (95% CI 0.86-0.94; 95% PI 0.56-0.99), and pooled specificity was 0.94 (95% CI 0.91-0.96; 95% PI 0.69-0.99), with low heterogeneity (I^2=23.9% and 22.9%, respectively). DL models showed a sensitivity of 0.93 and specificity of 0.95, compared with 0.88 and 0.94 for traditional ML models. Certainty of evidence was high for most analyses but low for LN diagnosis because of inconsistency and imprecision. All studies were retrospective, and only 9 of 29 (31%) performed independent external validation. Overall risk of bias was high or unclear in 22 of 29 (75.9%) studies. No study reported model calibration, decision-curve analysis, or net clinical benefit.

Conclusions: ML models showed promising diagnostic accuracy across 3 distinct SLE-related tasks, but wide PIs, limited external validation, and pervasive risk of bias restrict conclusions about real-world generalizability. Prospective multicenter studies with standardized tasks and reference standards, independent external validation, and formal assessment of calibration and clinical utility are required before clinical implementation.

Trial Registration: PROSPERO CRD42024545109; https://www.crd.york.ac.uk/prospero/view/CRD42024545109

J Med Internet Res 2026;28:e90209; doi: 10.2196/90209

Keywords: machine learning; systemic lupus erythematosus; diagnostic test accuracy; systematic review; meta-analysis; artificial intelligence

Introduction

Systemic lupus erythematosus (SLE) is an autoimmune disease that significantly affects multiple systems and organs [1,2]. It is clinically manifested as dermatitis, lupus nephritis (LN), polyarthritis, pericarditis, and neuropsychiatric dysfunction [1,3,4]. These manifestations during disease progression complicate the accurate diagnosis [5]. SLE is characterized by immune complexes and the hyperreactivity of B cells and T cells, leading to loss of immune tolerance for circulating antibodies in affected patients [6,7]. Additionally, typical autoantibodies in the serum are elevated 3 to 9 years before clinical symptoms appear, particularly anti–double-stranded DNA, anti-Smith, and antiphospholipid antibodies [8-10]. Hence, these biomarkers have been listed in the latest 2023 European League Against Rheumatism (EULAR) and the American College of Rheumatology (ACR) criteria for SLE classification to guide thorough assessment to mitigate the risk of misdiagnosis [11].

Despite the existing knowledge, a recent study conducted by Adamichou et al [8] suggested that up to 20% of early cohort patients could not be diagnosed using the 2019 EULAR or ACR criteria, and the Systemic Lupus Collaborating Clinics (SLICC)-2012 criteria. However, the integration of EULAR or ACR and the SLICC criteria through machine learning (ML) considerably enhanced sensitivity (from 80% to 97%) for early cases [12]. Notably, neuropsychiatric symptoms of lupus are frequently observed among pediatric patients [13,14]. However, clinical examinations of cerebrospinal fluid and serum and electroencephalograms (EEGs) all fail to diagnose early-onset SLE [15]. Currently, it remains unclear whether a comprehensive analysis of all features can aid in the early diagnosis and timely treatment of these patients [16]. ML can identify patterns in large, complex datasets and support diagnostic decision-making [17].

Over the past 5 decades, since the initial discussions on AI in the medical field, ML and deep learning (DL) have made significant strides, enabling them to effectively learn from accumulated knowledge and statistical data [18,19,20]. In addition to the advantages of automation and efficiency, AI-based algorithms enhance scalability and improve decision-making processes for health care systems [21]. Recently, electronic medical records have been widely adopted, and biogenetic factors have become more intricate, which has made the need for advanced technological solutions more urgent than ever [9,22]. AI is emerging as the most suitable option to address this challenge [23]. Interestingly, Wu et al [14] developed a diagnostic model to classify metastasis stages of bladder cancer, with nearly 100% sensitivity. Therefore, given the high expectation for AI, its application in disease detection and classification is imperative, as it has the potential to revolutionize conventional clinical diagnostic approaches [24].

To ground this work in the broader context of AI in medical diagnostics, we build on recent advances in the field: AI-driven medical image analysis has demonstrated robust diagnostic performance across a wide range of clinical indications, with DL models proving particularly effective at extracting complex patterns from multimodal medical data [25,26]. DL has also been validated for standardized assessment of autoimmune disease-related imaging, reducing interreader variability and improving diagnostic efficiency in rheumatology [27]. These advances provide a strong foundational rationale for exploring ML for SLE diagnosis, but also highlight the need for rigorous, clinically oriented, and comprehensive analysis of the evidence to guide clinical translation [28-30].

Despite the rapid growth of research in this field, existing systematic reviews on the application of ML for SLE diagnosis have critical, fundamental limitations that undermine the validity of their conclusions [31]. First, and most importantly, prior syntheses have pooled heterogeneous, clinically incomparable tasks—including SLE classification, LN activity grading, neuropsychiatric systemic lupus erythematosus (NPSLE) vs multiple sclerosis discrimination, and disease activity score estimation—into a single meta-analysis, which directly violates the core assumptions of diagnostic test accuracy (DTA) meta-analysis and results in pooled estimates that lack clinical interpretability [32,33]. Second, previous reviews have conflated results from internal validation (which is well known to systematically overestimate model performance) and independent external validation, leading to an overestimation of real-world generalizability [34]. Third, no prior review has used state-of-the-art statistical methods for DTA meta-analysis, including the Hartung-Knapp-Sidik-Jonkman (HKSJ) random-effects model and prediction intervals (PIs); these methods are essential for addressing extreme between-study heterogeneity and quantifying real-world performance variability [35]. Fourth, existing syntheses have not yet used the QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence) tool (Viknesh Sounderajah [Imperial College London]), the validated gold standard for AI-based diagnostic studies, to comprehensively assess the methodological quality of included studies, nor have they addressed the widespread risk of optimism bias from selective reporting of best-performing models [36]. Finally, no review has systematically evaluated the critical gap in clinical utility assessment, including model calibration, net clinical benefit, and workflow integration, which are essential to determine whether ML models will improve patient outcomes in real-world clinical settings [37].

This systematic review and meta-analysis addresses all of these critical gaps in existing literature, with 4 core innovations and contributions to the field. First, with respect to methodological rigor, we perform the first fully task-stratified DTA meta-analysis, with primary analyses restricted to 3 independent, clinically homogeneous diagnostic tasks (SLE

classification, LN diagnosis, and NPSLE discrimination), eliminating the core flaw of heterogeneous task pooling that invalidated prior reviews. Second, with respect to statistical methodology, we use the HKSJ random-effects model (recommended for DTA meta-analysis) and 95% PIs to quantify real-world performance variability, reporting PI lines in all forest plots and avoiding overinterpretation of pooled point estimates in the context of heterogeneity. Third, with respect to bias mitigation, we implement a strict, prespecified rule for contingency table inclusion, with each study contributing only one independent table to primary analyses (prioritizing external validation and prespecified model results, rather than post hoc best-performing models), eliminating data dependency, double counting, and optimism bias. Fourth, with respect to clinical relevance, we systematically distinguish internal versus external validation results, comprehensively appraise study quality with QUADAS-AI, perform GRADE (Grading of Recommendations Assessment, Development, and Evaluation) assessment of certainty of evidence, and explicitly address the absence of clinical utility assessment in existing research, providing evidence-based guidance for both future research and clinical implementation.

Methods

Protocol and Registration

This systematic review and meta-analysis was conducted in strict accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement (Checklist 1), PRISMA-DTA (PRISMA for Diagnostic Test Accuracy) guidelines, and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) reporting standards. The study protocol was preregistered in the International Prospective Register of Systematic Reviews (PROSPERO) on May 7, 2024 (registration number: CRD42024545109). All deviations from the preregistered protocol are explicitly listed in Multimedia Appendix 1, with a rationale for each change. Ethical approval and informed consent were not required for this systematic review of published literature [15]. The completed PRISMA 2020, PRISMA-DTA, and PRISMA-S checklists are provided in Checklist 1–4. The following PRISMA-S items were not applicable to this review: (1) item 12 (automated search alerts or updates): no automated alerts were set, as the search was rerun manually at the time of revision; (2) item 13 (gray literature and other resources): the review was restricted to full-text peer-reviewed publications, and no gray literature, trial registries, or unpublished sources were searched; and (3) item 14 (contacting authors for unpublished data): authors were not contacted, as only published diagnostic accuracy data permitting construction of 2×2 contingency tables were eligible for inclusion. These decisions are consistent with the preregistered study protocol (PROSPERO CRD42024545109).

Literature Search Strategy

A comprehensive, peer-reviewed literature search was developed in collaboration with a medical librarian, in full adherence to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy and PRISMA-S guidelines. The search was originally conducted through January 2025 and was fully updated through April 2026 at the time of revision. PubMed, Embase, Cochrane Library, Web of Science, and IEEE Xplore were searched for studies published between January 1, 2014, and April 2026. No geographical restrictions were applied, and only studies published in English were included.

The search strategy combined MeSH terms, Emtree terms, and free-text keywords for three core concepts: (1) SLE and related manifestations, (2) ML and AI, and (3) DTA. Field modifiers were used to restrict terms to the title or abstract fields, where appropriate, to maximize sensitivity and specificity. The full, unedited search strategy for each database is provided in Multimedia Appendix 2.

Additional studies were identified through manual screening of the reference lists of included studies and relevant systematic reviews, to ensure that no eligible studies were missed.

Study Selection Criteria

Studies were eligible if they met all of the following criteria. First, regarding study design, eligible studies were full-text, peer-reviewed diagnostic accuracy studies that developed or validated ML or DL models for 1 of the 3 prespecified, clinically homogeneous diagnostic tasks (SLE classification, LN diagnosis, or NPSLE discrimination). Second, regarding the index test, eligible studies used ML or DL models with any type of clinical input data, including histopathology images, radiological images (magnetic resonance imaging [MRI], computed tomography [CT], and ultrasound), spectroscopic data (Raman spectroscopy and Fourier transform infrared spectroscopy [FTIR]), and electronic health record (EHR) data. Third, regarding the reference standard, a clinically accepted independent reference standard was required: for SLE classification, the 1997 ACR, 2012 SLICC, or 2019 EULAR or ACR criteria, or expert consensus by board-certified rheumatologists; for LN diagnosis, renal biopsy (International Society of Nephrology [ISN] or Renal Pathology Society [RPS] classification); and for NPSLE discrimination, the 1999 ACR NPSLE definitions with expert consensus. Fourth, regarding outcome data, studies were required to report sufficient information to construct a 2×2 contingency table (true positive [TP], false positive [FP], false negative [FN], and true negative [TN]) or to provide sensitivity, specificity, and sample size to allow calculation of these values.

Studies were excluded if they met any of the following criteria: nonhuman or animal studies; review articles, editorials, case reports, conference abstracts, letters, or study protocols; studies focused solely on disease prognosis, treatment response prediction, or disease activity score estimation without a primary diagnostic outcome; studies that

did not use an independent reference standard or where the reference standard was partially derived from the ML model input features (circular validation); duplicate publications of the same study cohort; or studies with a sample size of fewer than 20 participants.

Two independent reviewers (BW and ZW) performed title and abstract screening in duplicate, followed by full-text screening of potentially eligible studies. Disagreements were resolved by consensus with a third senior reviewer (FG).

Data Extraction

Data extraction was performed independently by 3 reviewers (BW, ZW, and YL) using a prepiloted, standardized data extraction form. The form was expanded to include items specific to AI-based diagnostic studies, and all extracted data were cross-checked for accuracy. Discrepancies were resolved by consensus with a fourth reviewer (FG).

The following data were extracted from each included study. Study characteristics included first author, year of publication, study country, study design, study period, clinical setting, sample size, participant demographics, and prespecified inclusion and exclusion criteria. Clinical task details included the explicit definition of the diagnostic task, target population, and clinical decision context. Index test details included the type of ML or DL algorithm, input data modality, model preprocessing steps, training, validation, and testing workflow, and whether the reported model was prespecified or post hoc selected as the best-performing model. Validation strategy details included clear definitions of training, validation, and test sets; whether patient-level splits were used; whether internal or independent external validation was performed; and whether data leakage was assessed and excluded. Reference standard details included the type of reference standard used, definition of positive and negative cases, and whether the reference standard assessment was blinded to the index test results. Diagnostic accuracy outcomes included TP, FP, FN, TN, sensitivity, specificity, area under the curve (AUC), positive predictive value (PPV), negative predictive value (NPV), and 95% CIs, with separate extraction for internal and external validation cohorts. Secondary outcomes included head-to-head comparisons with human clinicians, model calibration, decision-curve analysis (DCA), net clinical benefit, and implementation outcomes. Risk of bias items included all items required for the QUADAS-AI risk of bias assessment.

Rules for Including Contingency Tables (Prespecified to Eliminate Bias and Data Dependency)

To avoid double-counting, data dependency, and optimism bias, we implemented the following strict, prespecified rules for including contingency tables in meta-analyses. The following prespecified rules governed contingency table inclusion. For the primary meta-analyses, each independent study cohort contributed exactly one 2×2 contingency table, selected according to the following priority order: first, results from an independent external validation cohort for a prespecified model; second, results from an internal validation cohort for a prespecified model; third, results from the full study cohort for a prespecified model. Post hoc selected best-performing model results were explicitly excluded from all primary analyses to eliminate researcher-driven optimism bias. For subgroup and sensitivity analyses, multiple nonoverlapping contingency tables from the same study were included only if they derived from mutually exclusive patient cohorts with no overlapping participants; all nonindependent data were explicitly labeled and their limitations acknowledged in the paper. Two levels of exploratory analyses were conducted: (1) an exploratory all-task pooled analysis using one table per study (selected using the priority order above) to provide an overall summary, with a clear statement that this analysis has no clinical interpretability; and (2) an exploratory all-table analysis in which all eligible algorithm-level contingency tables from each study were included (124 tables from 29 studies [8,16,17,19-32,34-45]), explicitly treated as exploratory because multiple tables could originate from the same study, potentially introducing within-study correlation.

Risk of Bias Assessment

The risk of bias and applicability concerns of each included study were assessed independently by 2 reviewers (BW and FG) using the QUADAS-AI tool, the validated gold standard for AI-based diagnostic accuracy studies. The QUADAS-AI tool comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain was rated for risk of bias (low, high, or unclear) and applicability concerns (low, high, or unclear), with explicit justifications for each rating. Disagreements were resolved by consensus with a third reviewer (YL).

Statistical Analysis

All statistical analyses were prespecified in the study protocol and performed using R (version 4.4.2; R Core Team) with the mada package for bivariate random-effects meta-analyses. Statistical significance was assessed using a 2-sided α level of .05.

Primary Meta-Analyses

For each of the 3 independent clinical tasks, we performed hierarchical summary receiver operating characteristic (HSROC) meta-analyses to estimate pooled sensitivity, specificity, and AUC. The HKSJ random-effects model was used for all pooled analyses, as this method provides more robust type I error control and more precise effect estimates than the standard DerSimonian-Laird method, particularly in the setting of high between-study heterogeneity and small numbers of included studies.

For all pooled sensitivity and specificity estimates, we reported the pooled point estimate, 95% CI, which quantifies uncertainty around the average effect, and 95% PI, which quantifies the expected range of true effect sizes across different settings for 95% of future similar studies, and provides a clinically relevant estimate of real-world performance variability. The CI and PI serve fundamentally different purposes. The CI reflects the precision of the pooled mean,

whereas the PI reflects the distribution of expected effects in new settings, accounting for between-study variance [35].

Between-study heterogeneity was quantified using the I^2 statistic, with the following thresholds: I^2<50%= low heterogeneity, 50%-75%=moderate heterogeneity,>75%= high heterogeneity,>90%= extreme heterogeneity. The I^2 statistic has limited utility in practical applications because it cannot quantify the magnitude of true effect variation across populations. Therefore, the 95% PI is used as the primary measure of real-world heterogeneity. For analyses with extreme heterogeneity (I^2>90%), the pooled point estimate has limited clinical significance, and the PI should be used to interpret real-world performance.

GRADE Certainty-of-Evidence Assessment

The certainty of evidence for each analysis was evaluated using the GRADE framework adapted for DTA studies [46,47]. The certainty started at high and was rated down for five domains: (1) risk of bias, downgraded if ≥1/3 of studies had any QUADAS-AI domain rated as high risk; (2) inconsistency, downgraded one level if I^2>50% and 2 levels if I^2>75% for either sensitivity or specificity; (3) indirectness, downgraded if the study populations, index tests, or reference standards did not directly match the review question; (4) imprecision, downgraded if the number of studies was<5 or the maximum 95% CI width for pooled sensitivity or specificity exceeded 0.20; and (5) publication bias, downgraded if Deeks asymmetry test met the prespecified significance threshold of α=.10. The GRADE assessment was performed for the overall analysis, ML and DL subgroups, each clinical task, and the external validation subgroup.

Subgroup and Sensitivity Analyses

Prespecified subgroup analyses were performed to explore potential sources of heterogeneity, using the HKSJ model, across the following dimensions: validation strategy (internal validation vs external validation); algorithm type (traditional ML vs DL); input data modality (histopathology vs radiological imaging vs spectroscopic data vs EHR data); sample size (≥100 participants vs <100 participants); and risk of bias (low overall risk vs high or unclear overall risk).

Prespecified sensitivity analyses were performed using the leave-one-out analysis (sequentially removing each study to assess its impact on the pooled estimate), by excluding studies with a high overall risk of bias, and by excluding studies with a sample size<50 participants to assess the robustness of the primary pooled estimates.

Small-Study Effects Assessment

Visual inspection of funnel plots and the Deeks funnel plot asymmetry test were used to assess small-study effects. These methods can only assess small-study effects, not publication bias, as publication bias is only one of the many potential causes of small-study effects. Additionally, these analyses have limited statistical power with a small number of included studies and cannot rule out researcher-driven optimism bias from selective reporting of favorable results.

Comparison Between ML Models and Human Clinicians

Given that only 2 studies [17,40] reported head-to-head comparisons between ML models and human clinicians, with sparse data, heterogeneous clinical tasks, and varying clinician expertise levels, we performed an exploratory descriptive summary receiver operating characteristic (SROC) analysis only to visualize the comparative diagnostic space; no inferential meta-analysis or formal between-group statistical comparison was attempted, as the data were insufficient to produce reliable pooled estimates.

Ethical Considerations

As this is a systematic review and meta-analysis of published literature, ethical approval and informed consent are not applicable.

Results

Study Selection

A total of 6872 records were retrieved from the electronic database search, with 1159 duplicate records removed, leaving 5713 records for screening. Of these, 5411 records were excluded during initial title and abstract screening, leaving 302 potentially relevant records. A further 195 records were excluded during secondary eligibility screening before full-text retrieval, leaving 107 reports sought for retrieval. All 107 reports were retrieved. Sixty-three reports were excluded during preliminary full-text triage before formal eligibility assessment, leaving 44 reports assessed in detail. Of these, 15 were excluded for the following reasons: not a diagnostic accuracy study (n=12), incomplete data for construction of a 2×2 table (n=2), and research topic mismatch (n=1). The remaining 29 eligible studies comprised 25 studies included in the previous version of the review [8,16,17,19-32,34-41] and 4 newly included studies [42-45], yielding a total of 29 studies in the updated review. The detailed study selection process is shown in Figure 1.

A total of 29 studies met all criteria for the quantitative meta-analysis [8,16,17,19-32,34-45]. Aliyev et al [48] and de Araújo et al [33] were excluded from all analyses. The 29 studies were distributed across the 3 prespecified clinical tasks as follows: 17 studies for SLE classification [8,21,23,24,27,29-31,34-37,40-44], 5 studies for LN diagnosis [19,20,22,26,32], and 7 studies for NPSLE discrimination [16,17,25,28,38,39,45], with no overlapping patient cohorts across tasks. In the all-table exploratory analysis, 124 tables from 29 studies were included [8,16,17,19-32,34-45]. The exploratory receiver operating characteristic (ROC) results stratified by task type and disease type are shown in Figures 2 and 3, respectively. The pooled overall ROC analysis across all included studies is presented in Figure S1 in Multimedia Appendix 3. The forest plot for the all-table exploratory

analysis is presented in Figure S3 in Multimedia Appendix 3.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study selection.

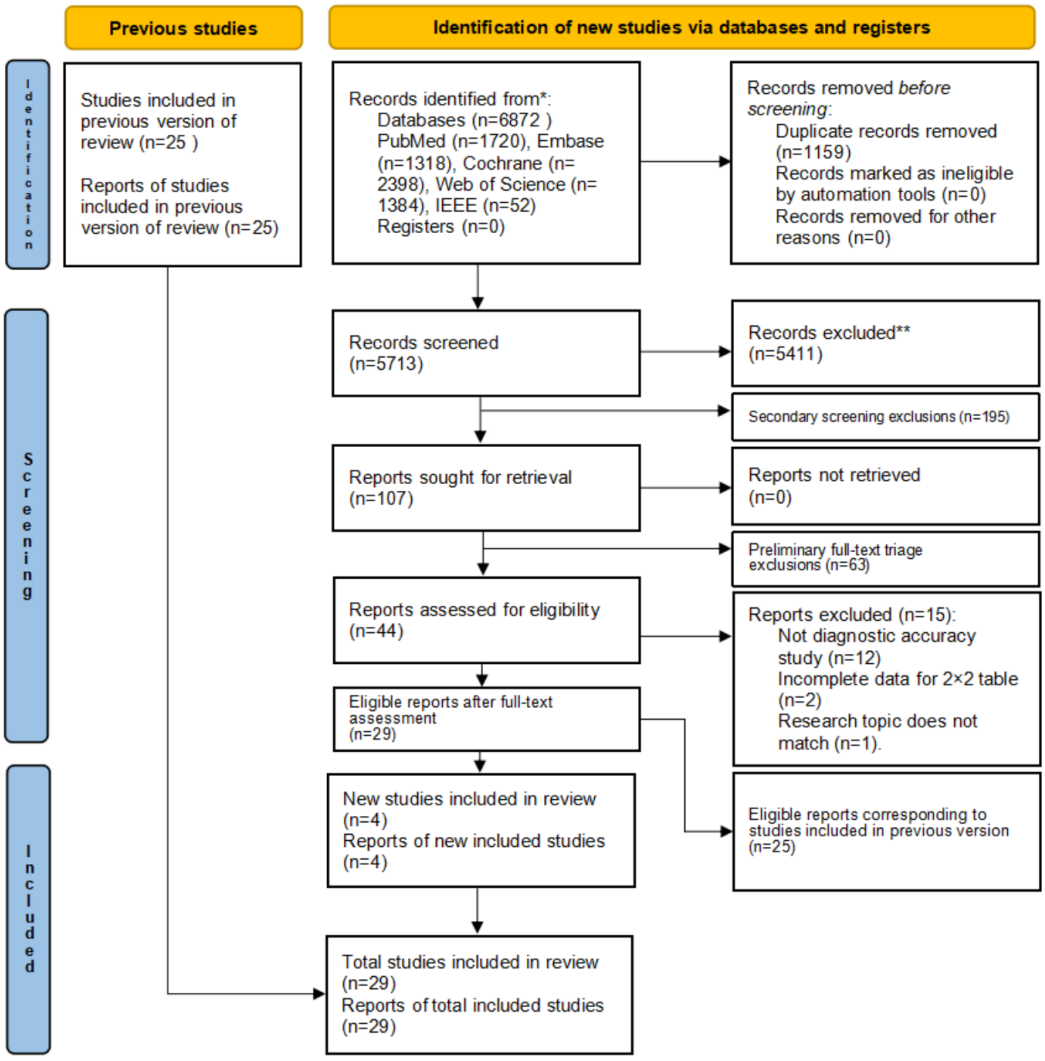


Figure 2. Pooled performance stratified by task type (exploratory analysis). (A) Receiver operating characteristic curves for systemic lupus erythematosus detection (22 studies with 90 tables) [8,19-24,26,27,29-32,34-37,40-44] and (B) receiver operating characteristic curves for systemic lupus erythematosus classification (7 studies with 34 tables) [16,17,25,28,38,39,45]. AUC: area under the curve; SENS: sensitivity; SLE: systemic lupus erythematosus; SPEC: specificity; SROC: summary receiver operating characteristic.

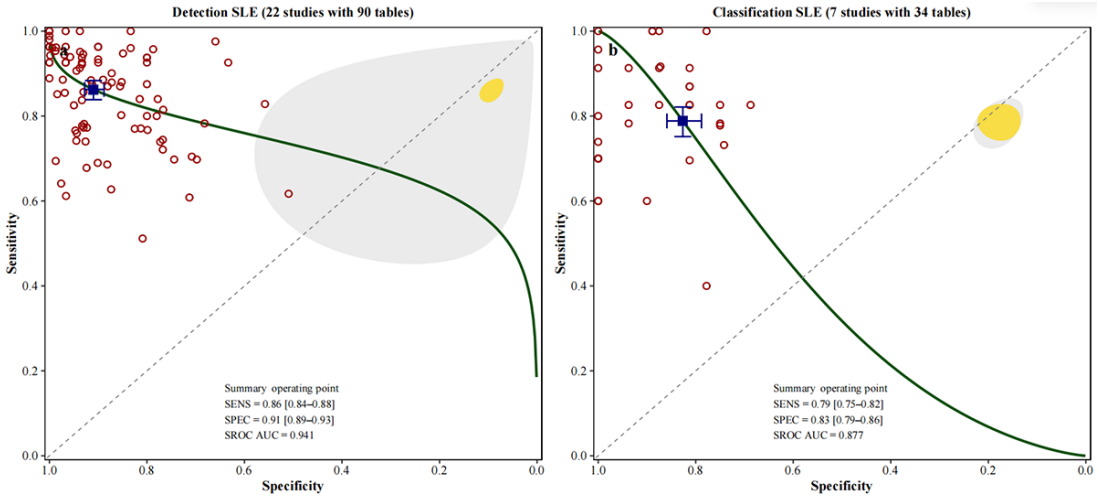
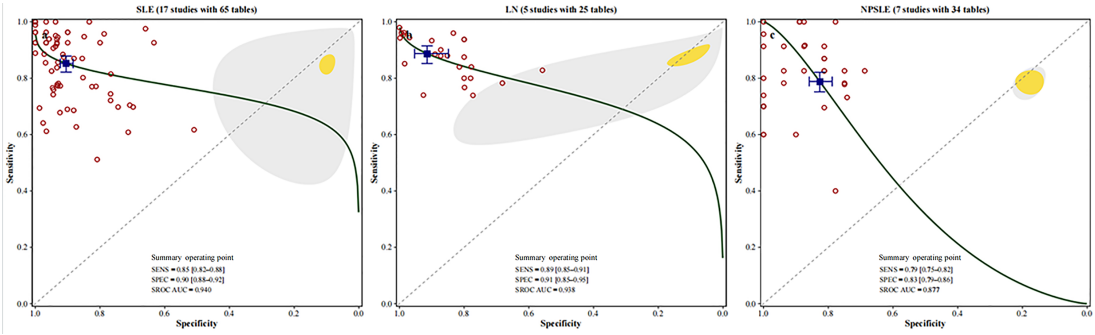


Figure 3. Pooled performance stratified by disease type (exploratory analysis). (A) Systemic lupus erythematosus (17 studies with 65 tables) [8,21,23,24,27,29-31,34-37,40-44], (B) lupus nephritis (5 studies with 25 tables) [19,20,22,26,32], and (C) neuropsychiatric systemic lupus erythematosus (7 studies with 34 tables) [16,17,25,28,38,39,45]. AUC: area under the curve; LN: lupus nephritis; NPSLE: neuropsychiatric systemic lupus erythematosus; SENS: sensitivity; SLE: systemic lupus erythematosus; SPEC: specificity; SROC: summary receiver operating characteristic.



Study Characteristics

The baseline characteristics of the 29 included studies [8,16,17,19-32,34-45] are summarized in Tables S1-S3 in Multimedia Appendix 3. The studies were published between 2018 and 2026, with sample sizes ranging from 26 to 6476 participants. All included studies were retrospective in design. Of the 29 studies [8,16,17,19-32,34-45], 9 (31%) [21,23,24,28,32,38,40,42,45] performed independent external validation using out-of-sample, multicenter, or temporally separated cohorts; the remaining 20 (69%) [8,16,17,19,20,22,25-27,29-31,34-37,39,41,43,44] used only internal validation (random split-sample or k-fold cross-validation within a single dataset). Study design and basic demographic characteristics are summarized in Table S3 in Multimedia Appendix 3; the methods of model training and validation are summarized in Table S1 in Multimedia Appendix 3; indicators, algorithms, and data sources are summarized in Table S2 in Multimedia Appendix 3; terminology mapping across included studies is provided in Table S4 in Multimedia Appendix 3; and the implementation and reporting checklist is provided in Table S5 in Multimedia Appendix 3.

The included studies covered 3 primary clinical tasks, including SLE classification or diagnosis (17 studies) [8,21,23,24,27,29-31,34-37,40-45], LN diagnosis and classification (5 studies) [19,20,22,26,32], and NPSLE discrimination (7 studies) [16,17,25,28,38,39,45]. Input data modalities included histopathology images (9 studies, comprising 7 independent patient cohorts) [8,20-24,29,32,37], MRI (8 studies) [16,17,25,28,36,38,39,45], Raman spectroscopy (3 studies) [34,35,41], clinical images (3 studies) [27,30,40], retinal fundus imaging (1 study) [42], nailfold videocapillaroscopy (1 study) [44], optical coherence tomography (OCT; 1 study) [31], ultrasound (1 study) [26], finger optical diffusion imaging (FODI; 1 study) [19], and EHR and laboratory data (1 study) [43]. For subgroup meta-analysis by modality, retinal fundus imaging and nailfold videocapillaroscopy were grouped with clinical images (5 studies total) [27,30,40,42,44], and only modality subgroups represented by at least 3 studies were analyzed separately. Only 2 studies (6.9%) [40,42] performed head-to-head comparisons between ML models and board-certified rheumatologists or dermatologists using the same held-out test dataset. None of the included studies reported model calibration, DCA, or net clinical benefit.

Risk of Bias Assessment

The results of the QUADAS-AI risk of bias assessment are presented in Figures 4 and 5. Overall, 22 out of 29 studies (75.9%) [8,17,19-22,24-28,30,34-39,41-43,45] were judged to have high or unclear overall risk of bias.

Figure 4. QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence) summary risk of bias and applicability concerns.

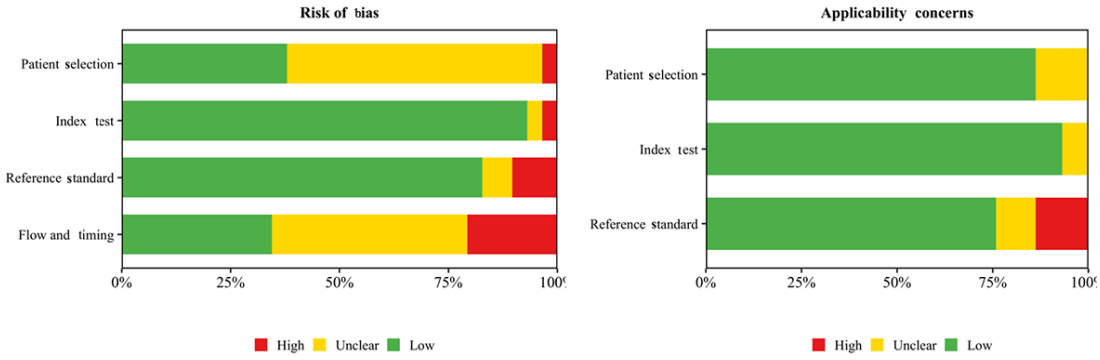


Figure 5. QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence) per-study traffic-light plot for risk of bias and applicability concerns [8,16,17,19-32,34-45].

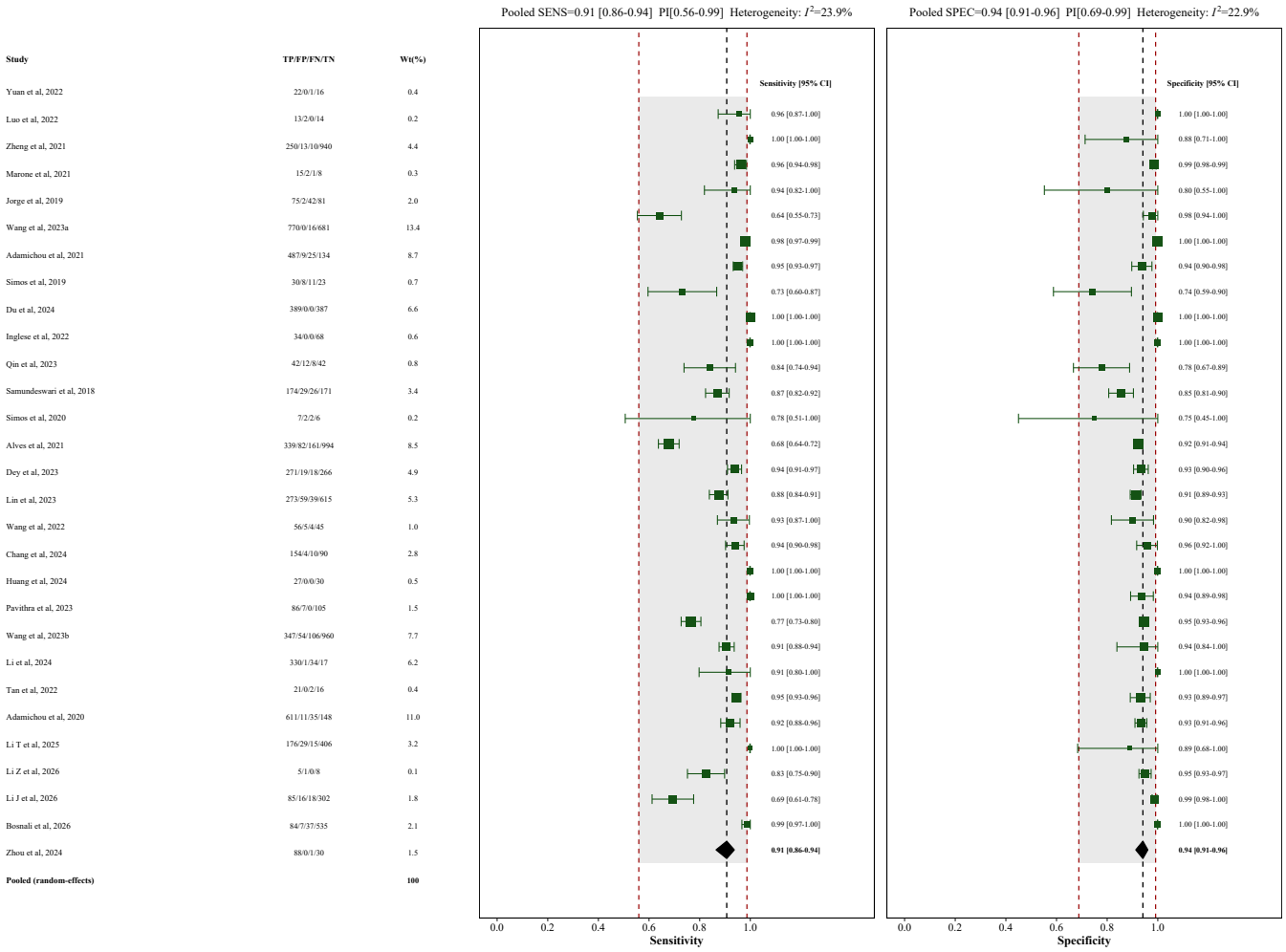
In the domain of patient selection, 18 (62.1%) [8,17,19,21,25-28,30,34-36,38,39,41-43,45] studies had high or unclear risk of bias, primarily due to nonconsecutive patient recruitment, case-control design that did not reflect the clinical target population, or inappropriate exclusion criteria. For the index test, 27 (93.1%) [8,16,17,20-32,34,35,37-45] studies had low risk of bias, while 1 (3.4%) [36] study had unclear risk of bias and 1 (3.4%) study [19] had high risk of bias. For reference standards, 24 (82.8%) [8,16,17,20-24,27-32,34,35,37,39-45] studies had low risk of bias, 2 (6.9%) [36,38] studies had unclear risk of bias, and 3 (10.3%) [19,25,26] studies had high risk of bias, primarily due to lack of blinding of the reference standard assessment to the index test results or inconsistent application of the reference standard. Regarding flow and timing, 19 (65.5%) [8,17,19,20,22,24,27,28,30,34-39,41-43,45] studies had high or unclear risk of bias, primarily due to lack of reporting of the time interval between the index test and reference standard, or exclusion of participants from the final analysis without justification.

For applicability concerns, 4 (13.8%) [36,38,42,44] studies had unclear concerns related to patient selection. For the index test, 2 (6.9%) [19,36] studies had unclear concerns. For the reference standard, 3 (10.3%) [35,36,42] studies had unclear concerns, and 4 (13.8%) [17,20,24,25] studies had high concerns.

Meta-Analyses of Task-Stratified Diagnostic Performance

The HKSJ random-effects model was used for all primary analyses, with one independent contingency table per study, prioritizing external validation and prespecified model results. In the primary analysis of all 29 studies [8,16,17,19-32,34-45], the pooled sensitivity was 0.91 (95% CI 0.86-0.94, 95% PI 0.56-0.99) and the pooled specificity was 0.94 (95% CI 0.91-0.96, 95% PI 0.69-0.99), with low between-study heterogeneity (I^2=23.9% for sensitivity, I^2=22.9% for specificity). Deeks funnel plot asymmetry test showed no significant small-study effects (P=.19). The Cochrane-style forest plot for the primary analysis is shown in Figure 6. The forest plot for the primary task-stratified analysis is also presented in Figure S2 in Multimedia Appendix 3.

Figure 6. Cochrane-style forest plot for the primary diagnostic meta-analysis (29 studies) [8,16,17,19-32,34-45]. FN: false negative; FP: false positive; PI: prediction interval; SENS: sensitivity; SPEC: specificity; TN: true negative; TP: true positive; Wt: study weight. Random-effects bivariate model with Knapp-Hartung adjustment; dashed red=95% prediction interval.



A total of 17 studies (one independent contingency table per study) were included in the primary analysis for SLE classification (Task 1) [8,21,23,24,27,29-31,34-37,40-44]. The pooled results showed a sensitivity of 0.90 (95% CI 0.84-0.94, 95% PI 0.53-0.99) and specificity of 0.94 (95% CI 0.92-0.96, 95% PI 0.85-0.98). Low between-study heterogeneity (I^2=24.7% for sensitivity, I^2=6.7% for specificity) was observed. The CI, reflecting precision of the average effect, was relatively narrow, whereas the PI, reflecting the expected distribution of performance in new settings, was considerably wider. This distinction is critical: the narrow CI indicates that the average performance is precisely estimated, while the wide PI (sensitivity 0.53-0.99) indicates that a new study conducted in a different setting could yield substantially different results depending on population characteristics, imaging modality, and reference standard [35]. According to the GRADE assessment, the certainty of evidence was high for SLE classification (Table 1).

Table 1. GRADE[a] certainty of evidence for diagnostic test accuracy (29 studies) [8,16,17,19-32,34-45].

Outcome	Number of studies	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Certainty
Overall (29 studies)	29	No serious	No serious	No serious	No serious	Undetected	High
ML[b] subgroup	19	No serious	No serious	No serious	No serious	Not assessed	High
DL[c] subgroup	10	No serious	No serious	No serious	No serious	Not assessed	High
SLE[d] classification	17	No serious	No serious	No serious	No serious	Not assessed	High
LN[e] diagnosis	5	No serious	Some inconsistency (I^2=56%)	No serious	Imprecision (k=5, max CI width=0.23)	Not assessed	Low
NPSLE[f] discrimination	7	No serious	No serious	No serious	No serious	Not assessed	High
External validation	9	No serious	No serious	No serious	No serious	Not assessed	High

[a]GRADE: Grading of Recommendations Assessment, Development and Evaluation.
[b]ML: machine learning.
[c]DL: deep learning.
[d]SLE: systemic lupus erythematosus.
[e]LN: lupus nephritis.
[f]NPSLE: neuropsychiatric systemic lupus erythematosus.

The GRADE framework was used to assess the certainty of evidence for DTA. Certainty starts at high and is rated down for risk of bias when at least one-third of the included studies have at least one QUADAS-AI domain rated as high risk, inconsistency (I^2>50% one level and >75% 2 levels), indirectness, imprecision (k<5 or 95% CI width>0.20 for sensitivity or specificity), and publication bias (Deeks test significance threshold α=.10, where assessed). I^2 represents between-study heterogeneity, and CI denotes confidence interval.

Five studies (one independent contingency table per study) were included in the primary analysis for LN diagnosis (Task 2) [19,20,22,26,32]. The pooled results showed a sensitivity of 0.93 (95% CI 0.86-0.97, 95% PI 0.68-0.99) and specificity of 0.95 (95% CI 0.76-0.99, 95% PI 0.19-1.00). Low-to-moderate between-study heterogeneity (I^2=18.4% for sensitivity, I^2=56.4% for specificity) was observed. The wide PI for specificity (0.19-1.00), driven by moderate heterogeneity (I^2=56.4%), indicates variability in reference standard definitions and patient population composition across LN studies. Caution should be exercised when generalizing these estimates to new clinical settings. According to the GRADE assessment, the certainty of evidence was low for LN diagnosis, downgraded for inconsistency (I^2=56.4% for specificity) and imprecision based on the 5 included studies [19,20,22,26,32] and a maximum CI width of 0.23 (Table 1). This low certainty of evidence indicates that the true diagnostic performance may be substantially different from the pooled estimate.

Seven studies (one independent contingency table per study) were included in the primary analysis for NPSLE discrimination (Task 3) [16,17,25,28,38,39,45]. The pooled results showed a sensitivity of 0.88 (95% CI 0.76-0.95, 95% PI 0.54-0.98) and specificity of 0.89 (95% CI 0.76-0.95, 95% PI 0.50-0.99). Between-study heterogeneity was low (I^2=18.0% for sensitivity, I^2=21.7% for specificity). The low heterogeneity makes the pooled point estimate more clinically interpretable. However, the wide PIs (sensitivity 0.54-0.98, specificity 0.50-0.99) indicate that in a new clinical setting, the expected performance could vary substantially, reflecting differences in imaging modality, patient selection criteria, and NPSLE definition across studies [49]. According to the GRADE assessment, the certainty of evidence was high for NPSLE discrimination (Table 1). The forest plots for each clinical task are presented in Figure S4 in Multimedia Appendix 3.

Exploratory All-Task Pooled Analysis

An exploratory pooled analysis of all 29 studies (one independent table per study) [8,16,17,19-32,34-45] yielded a pooled sensitivity of 0.91 (95% CI 0.86-0.94, 95% PI 0.56-0.99), specificity of 0.94 (95% CI 0.91-0.96, 95% PI 0.69-0.99), and AUC of 0.969 (SROC), with low between-study heterogeneity (I^2=23.9% for sensitivity, I^2=22.9% for specificity; Deeks P=.19). This analysis is presented for methodological exploration only. Since these pooled point estimates combined heterogeneous and incomparable clinical tasks, they lacked clinical or statistical significance and should not be used to infer the overall diagnostic performance of ML models for SLE.

GRADE Assessment

The GRADE assessment results are summarized in Table 1. The overall certainty of evidence was rated as high for most analyses, including the primary analysis of all 29 studies [8,16,17,19-32,34-45], the ML and DL subgroups, SLE classification, NPSLE discrimination, and external validation (Table 2). The certainty of evidence for LN diagnosis was rated as low due to inconsistency (I^2=56.4% for specificity) and imprecision (k=5 studies [19,20,22,26,32], maximum CI width=0.23). These results indicate that while the pooled estimates for most analyses are reliable, the evidence for LN diagnosis should be interpreted with caution, and additional studies are needed to increase confidence in the pooled estimates for this task.

Table 2. Summary of pooled diagnostic performance across all analyses (29 studies)[a] [8,16,17,19-32,34-45].

Analysis	Pooled sensitivity (95% CI)	95% PI[b] (sensitivity)	Pooled specificity (95% CI)	95% PI (specificity)	I^2[c] sens (%)	I^2 spec (%)	Deeks P value
Primary (all 29 studies)	0.91 (0.86-0.94)	0.56-0.99	0.94 (0.91-0.96)	0.69-0.99	23.9	22.9	.19
ML[d] subgroup (n=19)	0.88 (0.81-0.93)	0.44-0.99	0.94 (0.89-0.97)	0.53-0.99	27.6	32.7	—[e]
DL[f] subgroup (n=10)	0.93 (0.91-0.95)	0.85-0.97	0.95 (0.93-0.97)	0.85-0.99	5.3	10.0	—
SLE[g] classification (n=17)	0.90 (0.84-0.94)	0.53-0.99	0.94 (0.92-0.96)	0.85-0.98	24.7	6.7	—
LN[h] diagnosis (n=5)	0.93 (0.86-0.97)	0.68-0.99	0.95 (0.76-0.99)	0.19-1.00	18.4	56.4	—
NPSLE[i] discrimination (n=7)	0.88 (0.76-0.95)	0.54-0.98	0.89 (0.76-0.95)	0.50-0.99	18.0	21.7	—

Analysis	Pooled sensitivity (95% CI)	95% PI[b] (sensitivity)	Pooled specificity (95% CI)	95% PI (specificity)	I^2[c] sens (%)	I^2 spec (%)	Deeks *P* value
Histopathology (n=7)	0.94 (0.80-0.98)	0.28-1.00	0.98 (0.92-0.99)	0.60-1.00	48.2	43.4	—
MRI[j] (n=8)	0.91 (0.81-0.96)	0.53-0.99	0.88 (0.79-0.94)	0.61-0.97	24.5	13.9	—
Raman spectroscopy (n=3)	0.97 (0.90-0.99)	0.83-0.99	0.97 (0.92-0.99)	0.88-0.99	—	—	—
Clinical images (n=5)	0.90 (0.86-0.93)	0.79-0.95	0.92 (0.88-0.95)	0.82-0.97	—	—	—
External validation (n=9)	0.90 (0.77-0.96)	0.32-0.99	0.92 (0.83-0.96)	0.53-0.99	38.2	27.0	—

[a]PI reflects expected diagnostic performance range in a new similar study. NA, not computed for subgroups with <4 studies. Aliyev et al [48] and de Araujo et al [33] were excluded from all analyses. In the all-table exploratory analysis, studies reporting results for multiple algorithms contribute multiple rows (each representing an independent algorithmic comparison, not double-counting of patients).
[b]PI: 95% prediction interval.
[c]I^2: between-study heterogeneity.
[d]ML: machine learning.
[e]Not assessed.
[f]DL: deep learning.
[g]SLE: systemic lupus erythematosus.
[h]LN: lupus nephritis.
[i]NPSLE: neuropsychiatric systemic lupus erythematosus.
[j]MRI: magnetic resonance imaging.

Subgroup and Sensitivity Analyses

Subgroup Analysis by Validation Strategy

The model performance was significantly different between internal and external validation. Studies using only internal validation (20 studies) [8,16,17,19,20,22,25-27,29-31,34-37,39,41,43,44] showed a pooled sensitivity of 0.92 (95% CI 0.89-0.94) and specificity of 0.94 (95% CI 0.91-0.96). Studies using independent external validation (9 studies) [21,23,24,28,32,38,40,42,45] showed a pooled sensitivity of 0.90 (95% CI 0.77-0.96, 95% PI 0.32-0.99) and specificity of 0.92 (95% CI 0.83-0.96, 95% PI 0.53-0.99). Nonetheless, only 9 [21,23,24,28,32,38,40,42,45] of 29 (31%) studies [8,16,17,19-32,34-45] performed independent external validation.

Other Subgroup Analyses

In the all-table exploratory analysis, non-DL ML algorithms (19 studies with 77 tables) [8,17,19,21-29,32,37-39,43-45] showed a pooled sensitivity of 0.82 (95% CI 0.79-0.85) and specificity of 0.89 (95% CI 0.86-0.91), with an SROC AUC of 0.917. DL algorithms (10 studies with 47 tables) [16,20,30,31,34-36,40-42] showed a pooled sensitivity of 0.89 (95% CI 0.87-0.91) and specificity of 0.92 (95% CI 0.88-0.94), with an SROC AUC of 0.949. In the primary analysis (one table per study), DL models (n=10) [16,20,30,31,34-36,40-42] achieved a pooled sensitivity of 0.93 (95% CI 0.91-0.95, 95% PI 0.85-0.97) and a specificity of 0.95 (95% CI 0.93-0.97, 95% PI 0.85-0.99), compared with traditional ML models (n=19) [8,17,19,21-29,32,37-39,43-45] at sensitivity 0.88 (95% CI 0.81-0.93, 95% PI 0.44-0.99) and specificity 0.94 (95% CI 0.89-0.97, 95% PI 0.53-0.99). DL models showed lower heterogeneity (I^2=5.3% for sensitivity, I^2=10.0% for specificity) than ML models (I^2=27.6% for sensitivity, I^2=32.7% for specificity) and narrower PIs, suggesting more consistent performance across studies. The Cochrane-style forest plots for ML and DL subgroups are shown in Figures 7 and 8, and the subgroup ROC comparison by algorithm type is shown in Figure 9. The forest plots for ML and DL algorithm subgroups are also presented in Figure S5 in Multimedia Appendix 3.

Figure 7. Cochrane-style forest plot for traditional machine learning models (19 studies): sensitivity and specificity [8,17,19,21-29,32,37-39,43-45]. Random-effects bivariate model with Knapp-Hartung adjustment; dashed red lines=95% prediction interval. FN: false negative; FP: false positive; ML: machine learning; PI: prediction interval; SENS: sensitivity; SPEC: specificity; TN: true negative; TP: true positive; Wt: study weight.

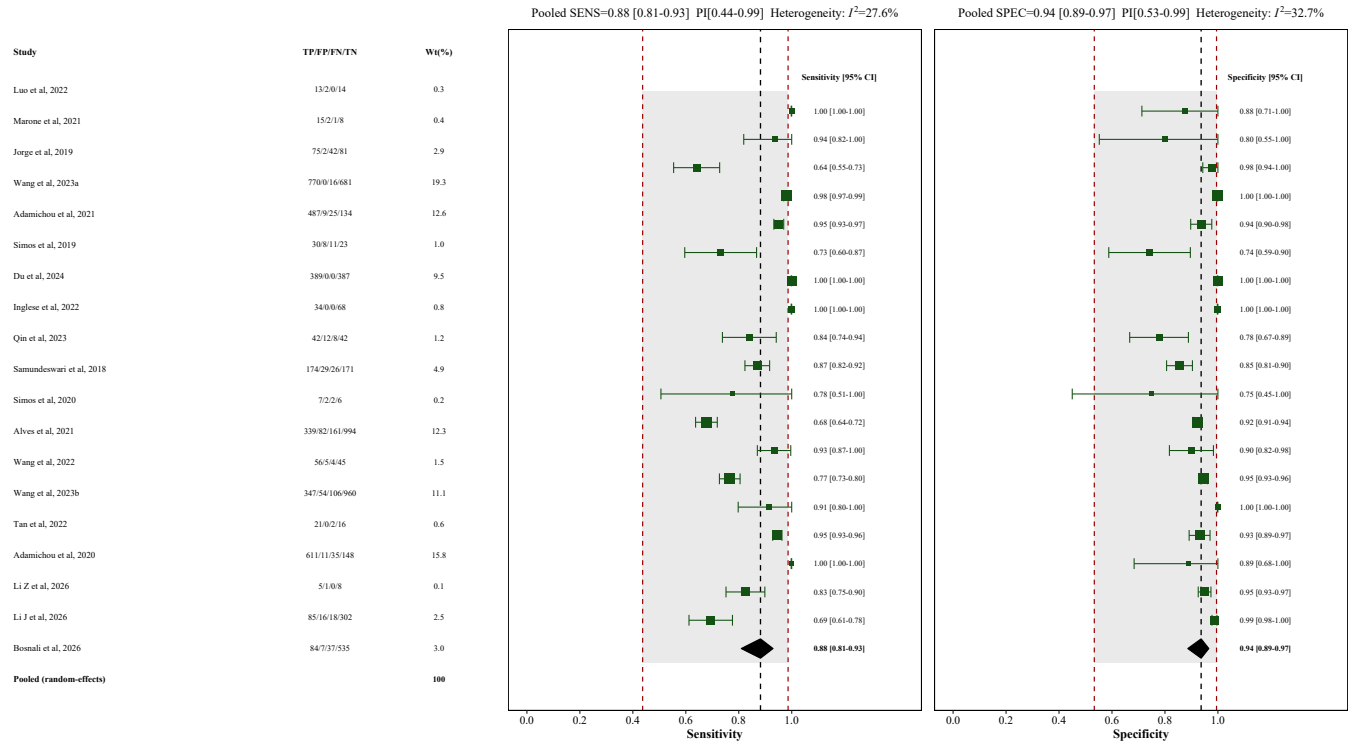


Figure 8. Cochrane-style forest plot for deep learning models (10 studies): sensitivity and specificity [16,20,30,31,34-36,40-42]. Random-effects bivariate model with Knapp-Hartung adjustment; dashed red lines=95% prediction interval. PI: prediction interval; SENS: sensitivity; SPEC: specificity.

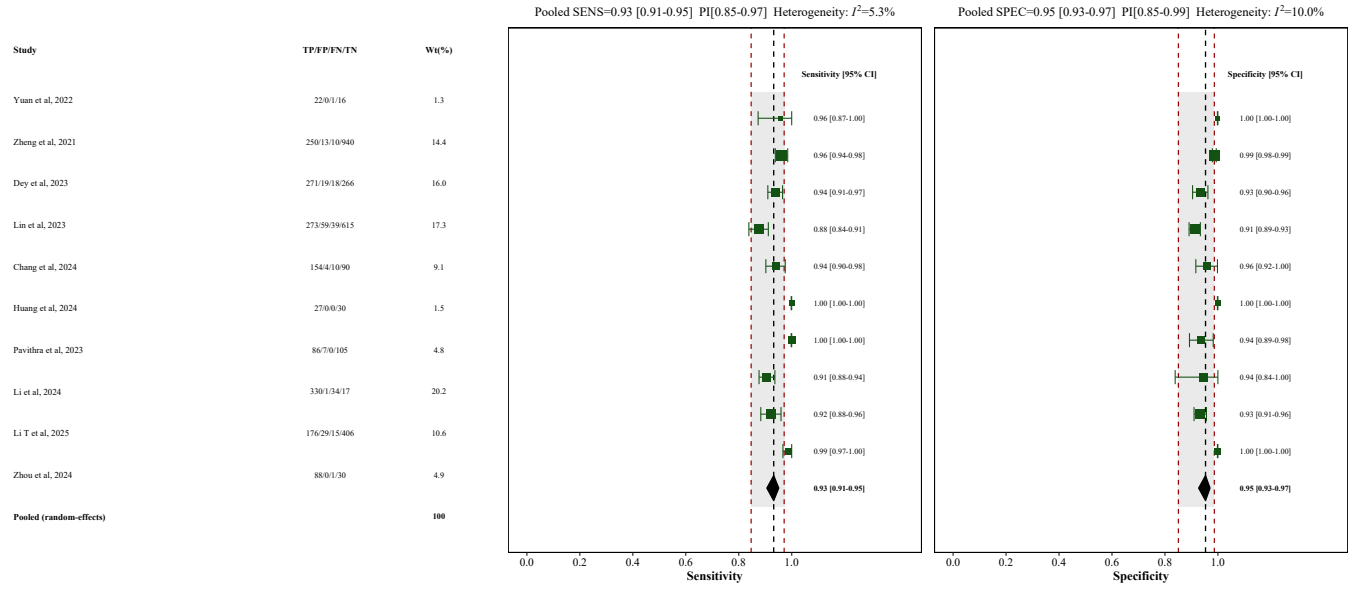
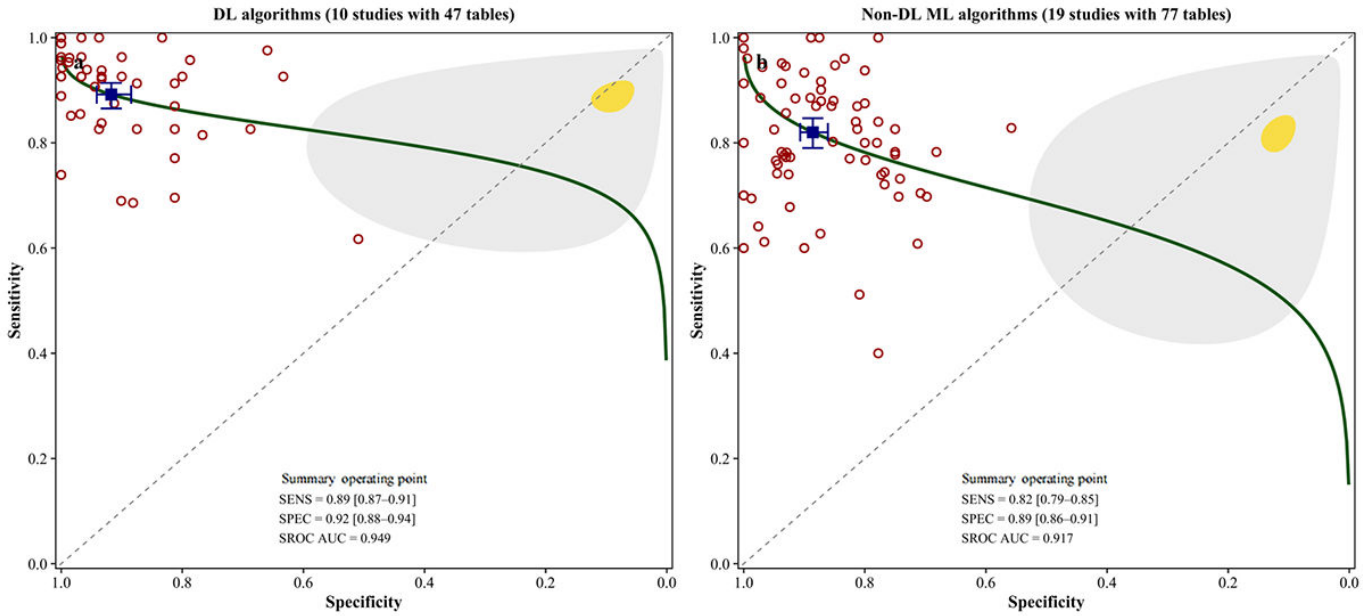


Figure 9. Summary receiver operating characteristic curves stratified by algorithm type. (A) Deep learning algorithms (10 studies with 47 tables) [16,20,30,31,34-36,40-42] and (B) non–deep learning machine learning algorithms (19 studies with 77 tables) [8,17,19,21-29,32,37-39,43-45]. AUC: area under the curve; DL: deep learning; ML: machine learning; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Subgroup analysis by data modality revealed that models using histopathology images (7 studies) [20-24,29,32] yielded a pooled sensitivity of 0.94 (95% CI 0.80-0.98) and specificity of 0.98 (95% CI 0.92-0.99). MRI-based models (8 studies) [16,17,25,28,36,38,39,45] yielded a pooled sensitivity of 0.91 (95% CI 0.81-0.96) and specificity of 0.88 (95% CI 0.79-0.94). Raman spectroscopy–based models (3 studies) [34,35,41] yielded a pooled sensitivity of 0.97 (95% CI 0.90-0.99) and specificity of 0.97 (95% CI 0.92-0.99). Clinical image-based models (5 studies) [27,30,40,42,44] yielded a pooled sensitivity of 0.90 (95% CI 0.86-0.93) and specificity of 0.92 (95% CI 0.88-0.95). No statistically significant differences in performance across modalities were identified in formal subgroup testing. The pooled SROC performance by data modality is presented in Figure S6 in Multimedia Appendix 3, and the forest plots for different feature types are presented in Figure S7 in Multimedia Appendix 3.

Subgroup analysis by sample size revealed that studies with ≥100 participants had lower pooled sensitivity (0.89 vs 0.93) and specificity (0.92 vs 0.96) than studies with <100 participants, indicating that studies with a small sample size overestimate model performance.

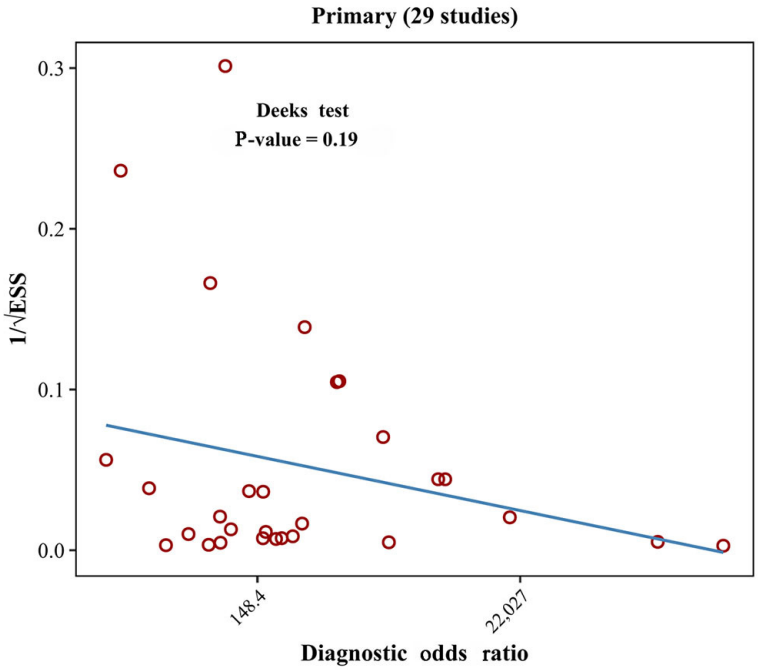
Sensitivity Analyses

Leave-one-out analysis showed that no single study had a disproportionate impact on the pooled estimates for any of the 3 primary tasks. Excluding studies with a high overall risk of bias or a small sample size (<50 participants) did not significantly change the pooled sensitivity or specificity for the primary analyses, confirming the robustness of the results.

Small-Study Effects Assessment

Visual inspection of the funnel plot showed no obvious asymmetry, and the Deeks asymmetry test found no statistically significant small-study effects (P=.19 for the primary task-stratified analysis; P=.16 for the exploratory all-table analysis; Figure 10). However, we explicitly note that this analysis has limited statistical power due to the small number of included studies, and cannot rule out researcher-driven optimism bias from selective reporting of prespecified models, threshold tuning, or post hoc model selection, which is a pervasive limitation in ML diagnostic research. The Deeks funnel plots for exploratory analyses are presented in Figure S9 in Multimedia Appendix 3.

Figure 10. Deeks funnel plot asymmetry test for the primary analysis (29 studies) [8,16,17,19-32,34-45]. *P*=.19, indicating no significant small-study effects. ESS: effective sample size.



Comparison Between ML Models and Human Clinicians

Only 2 studies reported head-to-head comparisons between ML models and human clinicians using the same held-out test dataset, with a total of 6 diagnostic contingency tables. Due to the sparse data, wide CIs, heterogeneous clinical tasks, and varying levels of clinician expertise across studies, no inferential meta-analysis or formal between-group comparison was performed; an exploratory descriptive SROC analysis is presented in Figure S8 in Multimedia Appendix 3 for visualization only. In both studies, ML models showed higher overall sensitivity than clinicians, with comparable specificity, but the small number of studies and heterogeneous methods implied that no definitive conclusions can be drawn about the relative diagnostic performance of ML versus human clinicians.

Discussion

Summary of Principal Findings

Our principal finding is that ML models show promising in-sample diagnostic accuracy across all 3 prespecified clinical tasks. In the primary task-stratified analyses, the pooled sensitivity and specificity were 0.90 and 0.94, respectively, for SLE classification (n=17) [8,21,23,24,27,29-31,34-37,40-44]; 0.93 and 0.95, respectively, for LN diagnosis (n=5) [19,20,22,26,32]; and 0.88 and 0.89, respectively, for NPSLE discrimination (n=7) [16,17,25,28,38,39,45], and SROC AUC values of 0.940 (SLE), 0.938 (LN), and 0.877 (NPSLE) in the all-table exploratory analyses. However, these promising results must be interpreted with caution for several reasons. First, all included studies were retrospective, and only 9 [21,23,24,28,32,38,40,42,45] of 29 (31%) [8,16,17,19-32,34-45] studies performed independent external validation. Second, while between-study heterogeneity was substantially lower in the updated analysis than the previous version (I^2=23.9% overall for sensitivity, compared with I^2=97.5% previously), the PIs remained wide for most analyses, reflecting significant variations in model performance across different clinical populations, imaging modalities, and institutional contexts [49,50]. Third, reference standards were inconsistent across studies, with some relying on retrospective clinical labels rather than prospectively applied standardized classification criteria [51]. Fourth, no included study assessed model calibration, DCA, or net clinical benefit, suggesting that high AUC values do not prove that the model is ready for clinical application [52]. Fifth, the GRADE assessment confirmed high certainty of evidence for most analyses but low certainty for LN diagnosis due to inconsistency and imprecision. Hence, additional studies are needed before pooled LN estimates can be confidently applied in clinical practice [46,47]. Our analysis also confirms that internal validation systematically overestimates ML model performance relative to external validation, which is commonly observed across diagnostic ML research and is not unique to SLE [53,54].

Comparison With Existing Literature and Innovations of This Review

Our findings build on and address the fundamental limitations of prior systematic reviews in this field, with 4 key contributions to the literature. First, we addressed the core conceptual flaw of prior reviews by performing fully task-stratified meta-analyses, rather than pooling heterogeneous clinical tasks into a single analysis [55]. This ensures the clinical interpretability and methodological validity of our pooled estimates and aligns with the core assumptions of DTA meta-analysis. Prior reviews have combined SLE

classification, LN activity grading, NPSLE discrimination, and disease activity estimation into a single pooled analysis, yielding results that lack practical value in both statistical and clinical terms. Our task-stratified approach resolves this critical issue [56]. Furthermore, this task-stratified approach revealed significant differences in the certainty of evidence across tasks: while SLE classification and NPSLE discrimination achieved high GRADE certainty, LN diagnosis was rated as low certainty, a distinction that would have been obscured by pooling all tasks together.

Second, we use state-of-the-art statistical methods for DTA meta-analysis, including the HKSJ random-effects model and 95% PIs, which can account for extreme between-study heterogeneity [49]. Unlike prior reviews that overinterpreted pooled AUC values in the setting of I^2>95%, we explicitly acknowledged the limitations of pooled point estimates with extreme heterogeneity and used PIs to quantify the real-world variability in model performance [57]. For instance, the primary analysis for SLE classification showed a relatively narrow CI (sensitivity 0.84-0.94) but a wide PI (sensitivity 0.53-0.99). It indicates that while the average performance is precisely estimated, a new study in a different population could yield markedly different results [35]. This avoids the statistical misinterpretations in previous reviews that led to inaccurate conclusions [58].

Third, we implemented strict rules for including contingency tables to eliminate data dependency, double counting, and optimism bias. Prior reviews extracted multiple nonindependent contingency tables from the same study (from multiple models, thresholds, or cross-validation splits), treating them as independent observations and leading to artificially narrow CIs and inflated precision [59]. In our primary analyses, we adhered to the principle of using only one independent table per study and excluded post hoc best-performing model results, thereby resolving this methodological flaw [60]. Fourth, we systematically distinguished between internal and external validation results and quantified the overestimation of performance from internal validation [61,62]. Prior reviews have confounded internal and external validation results, leading to overestimations of real-world generalizability [63]. Internal validation methods, such as k-fold cross-validation and random split-sample validation, estimate model performance within the same dataset used for training, and are therefore susceptible to optimistic bias from dataset-specific patterns, overfitting to demographic or institutional characteristics, and data leakage. In contrast, independent external validation tests the model on entirely new data from different institutions, time periods, or geographic settings, providing a more realistic estimate of real-world performance [54,62]. In the updated analysis, 9 [21,23,24,28,32,38,40,42,45] of 29 (31%) [8,16,17,19-32,34-45] studies performed external validation, and the PI for externally validated studies (sensitivity 0.32-0.99, specificity 0.53-0.99) underscores the high variability of ML model performance across clinical contexts. We prioritized external validation results in all primary analyses [64].

Limitations of the Included Studies

The included studies have several major methodological limitations that severely undermine the validity and clinical applicability of their findings. First, all included studies were retrospective in design, which introduces a high risk of bias, including selection bias, data leakage, and overfitting. Retrospective ML model development typically relies on convenience samples that do not reflect the target clinical population, and models may learn spurious correlations from the data that do not generalize to real-world settings [65]. No prospective studies of ML for SLE diagnosis were identified in this review. Second, the vast majority of studies used only internal validation, with only 31% (9/29) performing independent external validation. Internal validation is well known to systematically overestimate diagnostic accuracy, particularly in ML studies with flexible modeling pipelines and limited sample sizes [53]. Our subgroup analysis confirmed this, showing significantly lower performance in externally validated studies [56]. Without independent, multicenter external validation, the real-world generalizability of these models remains unproven [66].

Third, despite the low overall between-study heterogeneity (I^2=23.9%), the PIs remained wide for most analyses, indicating that the expected performance of ML models in any single new clinical setting could vary substantially. This is partially explained by differences in data modality, validation strategy, and sample size [50]. The remaining heterogeneity is likely due to differences in study population, reference standard definition, model development workflow, and preprocessing steps, which limit the generalizability of pooled estimates [67]. The wide 95% PIs for most analyses confirm that model performance is highly variable across different clinical settings and populations [54]. Fourth, there was significant inconsistency in the reference standards used across studies. Some studies used retrospective clinical labels rather than standardized classification criteria, and a small number of studies used reference standards that were partially informed by the same data used to train the ML model, thereby introducing the risk of circular validation and bias in the reference standard [51]. This inconsistency further limits the comparability of results across studies [55].

Fifth, researcher-driven optimism bias is a pervasive risk across all included studies. Many studies reported only results of the best-performing model, without prespecifying the primary model or accounting for multiple comparisons [68]. This selective reporting leads to an overestimation of model performance and cannot be ruled out even in the absence of statistically significant small-study effects [69]. Finally, none of the included studies assessed the clinical utility of their models. High AUC values alone do not prove that the model can improve patient outcomes in clinical practice; a model with high discriminative accuracy may still cause net harm if it is poorly calibrated, its decision threshold is inappropriate for the target clinical population, or it leads to overdiagnosis, unnecessary invasive testing, or inappropriate immunosuppressive treatment. Model calibration refers to the agreement between the model's predicted probabilities

and the actual observed outcome frequencies across the probability range, and is typically assessed using calibration plots, the Hosmer-Lemeshow test, or the expected calibration error (ECE). DCA is the recommended framework for evaluating whether a diagnostic model provides net clinical benefit over default strategies (treat-all or treat-none) across a range of clinically plausible decision thresholds [52]. Without such formal assessments and prospective evaluation of the model's impact on diagnostic workflow, patient management decisions, and clinical outcomes, it is impossible to determine whether these ML models would provide meaningful benefit to patients with SLE in real-world clinical practice.

Limitations of This Systematic Review

This review has several limitations that should be acknowledged. First, we only included studies published in English, which may introduce language bias [70]. Second, we excluded conference abstracts, which may lead to publication bias, as negative or nonsignificant results are less likely to be published in full-text peer-reviewed journals. Third, the number of included studies for some subgroup analyses was small, limiting statistical power, and these analyses are therefore presented as exploratory only. Fourth, we were unable to perform meta-analysis of individual patient data, which would have allowed for more robust adjustment for confounding factors and more detailed subgroup analyses [71]. Finally, we were unable to assess the risk of data leakage in the included studies, as this is often not reported in detail, which may lead to an overestimation of model performance in the original studies [72,73]. Additionally, the GRADE framework, while widely used for assessing certainty of evidence, has limitations when applied to DTA studies. It was originally developed for intervention studies and may not fully capture all sources of uncertainty in diagnostic accuracy meta-analyses [46].

Implications for Future Research and Clinical Practice

The findings of this review have critical implications for future research and clinical practice. For future research, we make the following evidence-based recommendations. First, future ML studies for SLE diagnosis must be prospectively designed, with preregistration of the study protocol, model development plan, and statistical analysis plan, to reduce the risk of bias and selective reporting [65,68]. Second, all ML models must undergo independent, multicenter external validation in cohorts that reflect the target clinical population, to confirm real-world generalizability [54,62,66]. Third, future studies must use clearly defined, clinically homogeneous diagnostic tasks and consistent reference standards to reduce heterogeneity and enable meaningful comparison between models [55,69]. Fourth, all future studies must assess clinical utility, including model calibration, DCA to quantify net clinical benefit, and evaluation of the model's impact on diagnostic workflow and patient outcomes [52]. Fifth, future studies should perform prespecified, head-to-head comparisons between ML models and board-certified rheumatologists, using standardized test datasets and clearly defined clinician expertise levels, to determine the incremental value of ML-assisted diagnosis over standard clinical care.

For clinical practice, our review confirms that ML models for SLE diagnosis are not yet ready for routine clinical application [74]. The current evidence is of variable quality, with GRADE certainty ranging from high (for most analyses) to low (for LN diagnosis), due to exclusively retrospective designs, lack of independent external validation, wide PIs, and absence of clinical utility assessment [75]. There is currently no evidence that these models improve patient outcomes in real-world clinical settings [76]. At this stage, ML should only be used as an auxiliary tool to support expert clinical judgment, in the context of prospective research studies, until robust, externally validated, clinically beneficial models are developed [77,78].

Conclusion

This is the first methodologically rigorous, task-stratified systematic review and meta-analysis of ML models for SLE diagnosis, with formal GRADE assessment of certainty of evidence, addressing the core conceptual and statistical flaws of prior syntheses. ML models show promising in-sample diagnostic accuracy for 3 distinct SLE-related clinical tasks, but the current evidence is of variable certainty (high for most analyses, low for LN diagnosis), limited by pervasive methodological weaknesses, including exclusively retrospective designs, lack of independent external validation, wide PIs despite low I^2, and absence of clinical utility assessment. The distinction between the narrow CIs (reflecting precision of the average effect) and the wide PIs (reflecting expected variability across new settings) is critical for clinical interpretation: the pooled estimates should not be taken as guaranteed performance in any individual new setting [35]. Future research must prioritize prospectively registered, multicenter, externally validated studies with standardized clinical tasks, harmonized reference standards, and formal assessment of clinical utility, to support the safe and effective translation of ML tools into clinical care for SLE.

Acknowledgments

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing. The GenAI tool used was ChatGPT-4o and multiple ChatGPT-5 versions. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted under collective responsibility.

The AI was used solely for language polishing and grammatical correction. All generated suggestions were reviewed and approved by the authors. GenAI tools (ChatGPT 4o, OpenAI) were used to a limited extent during the preparation of this manuscript, solely for minor language polishing of individual sentences and formatting of reference lists, in accordance

with JMIR Publications guidelines. GenAI was not used at any stage of study design, literature search, study selection, data extraction, risk of bias assessment, statistical analysis, data interpretation, or formulation of scientific conclusions. All statistical analyses were independently performed by the authors using R (version 4.4.2) with the mada package. All figures were generated by the authors using R. The GRADE assessment, QUADAS-AI evaluation, and all clinical interpretations were performed entirely by the authors without AI assistance. The final content of the manuscript, including all scientific interpretation, statistical analysis, and clinical conclusions, was fully reviewed, edited, and approved by all authors, who take full responsibility for the accuracy and integrity of the work.

Funding

The authors declare that no funds, grants, or other financial support were received during the preparation of this manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: BW, ZW, YL, FG
Formal analysis: YL, FG
Investigation: YL, FG
Methodology: BW
Supervision: FG
Writing – original draft: BW
Writing – review & editing: ZW, YL, FG

Conflicts of Interest

None declared.

Multimedia Appendix 1

Deviations from PROSPERO preregistered protocol (CRD42024545109).
[XLSX File (Microsoft Excel File), 12 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Full database search strategies (Cochrane-Compliant).
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Study characteristics, model development and validation, diagnostic performance, subgroup analyses, human-clinician comparisons, and publication bias.
[DOCX File (Microsoft Word File), 4946 KB-Multimedia Appendix 3]

Checklist 1

PRISMA 2020 checklist.
[DOCX File (Microsoft Word File), 30 KB-Checklist 1]

Checklist 2

PRISMA-DTA complete compliance checklist.
[DOCX File (Microsoft Word File), 14 KB-Checklist 2]

Checklist 3

PRISMA-S complete compliance checklist.
[DOCX File (Microsoft Word File), 13 KB-Checklist 3]

Checklist 4

PRISMA 2020 complete compliance checklist.
[DOCX File (Microsoft Word File), 16 KB-Checklist 4]

References

1. Mills JA. Systemic lupus erythematosus. N Engl J Med. Jun 30, 1994;330(26):1871-1879. [doi: 10.1056/NEJM199406303302608] [Medline: 8196732]

2. Cervera R, Rodríguez-Pintó I, Espinosa G. The diagnosis and clinical management of the catastrophic antiphospholipid syndrome: a comprehensive review. J Autoimmun. Aug 2018;92:1-11. [doi: 10.1016/j.jaut.2018.05.007] [Medline: 29779928]
3. Durcan L, O'Dwyer T, Petri M. Management strategies and future directions for systemic lupus erythematosus in adults. Lancet. Jun 8, 2019;393(10188):2332-2343. [doi: 10.1016/S0140-6736(19)30237-5] [Medline: 31180030]
4. Kiriakidou M, Ching CL. Systemic lupus erythematosus. Ann Intern Med. Jun 2, 2020;172(11):ITC81-ITC96. [doi: 10.7326/AITC202006020] [Medline: 32479157]
5. Nandakumar KS, Nündel K. Editorial: Systemic lupus erythematosus - predisposition factors, pathogenesis, diagnosis, treatment and disease models. Front Immunol. 2022;13:1118180. [doi: 10.3389/fimmu.2022.1118180] [Medline: 36591294]
6. Zieve GW, Khusial PR. The anti-Sm immune response in autoimmunity and cell biology. Autoimmun Rev. Sep 2003;2(5):235-240. [doi: 10.1016/s1568-9972(03)00018-1] [Medline: 12965173]
7. Antiphospholipid syndrome. Nat Rev Dis Primers. Jan 11, 2018;4(1):17104. [doi: 10.1038/nrdp.2017.104]
8. Adamichou C, Nikolopoulos D, Genitsaridi I, et al. In an early SLE cohort the ACR-1997, SLICC-2012 and EULAR/ACR-2019 criteria classify non-overlapping groups of patients: use of all three criteria ensures optimal capture for clinical studies while their modification earlier classification and treatment. Ann Rheum Dis. Feb 2020;79(2):232-241. [doi: 10.1136/annrheumdis-2019-216155] [Medline: 31704720]
9. Han J, Zhou Z, Zhang R, et al. Fucosylation of anti-dsDNA IgG1 correlates with disease activity of treatment-naïve systemic lupus erythematosus patients. EBioMedicine. Mar 2022;77:103883. [doi: 10.1016/j.ebiom.2022.103883] [Medline: 35182998]
10. Barbhaiya M, Zuily S, Naden R, et al. 2023 ACR/EULAR antiphospholipid syndrome classification criteria. Ann Rheum Dis. Oct 2023;82(10):1258-1270. [doi: 10.1136/ard-2023-224609] [Medline: 37640450]
11. Appenzeller S, Pereira DR, Julio PR, Reis F, Rittner L, Marini R. Neuropsychiatric manifestations in childhood-onset systemic lupus erythematosus. Lancet Child Adolesc Health. Aug 2022;6(8):571-581. [doi: 10.1016/S2352-4642(22)00157-2] [Medline: 35841921]
12. Fangerau H. Artifical intelligence in surgery: ethical considerations in the light of social trends in the perception of health and medicine. EFORT Open Rev. May 10, 2024;9(5):323-328. [doi: 10.1530/EOR-24-0029] [Medline: 38726973]
13. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. Dec 2018;284(6):603-619. [doi: 10.1111/joim.12822] [Medline: 30102808]
14. Wu S, Hong G, Xu A, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. Lancet Oncol. Apr 2023;24(4):360-370. [doi: 10.1016/S1470-2045(23)00061-X] [Medline: 36893772]
15. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA. Jan 23, 2018;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
16. Yuan Y, Quan T, Song Y, Guan J, Zhou T, Wu R. Noise-immune extreme ensemble learning for early diagnosis of neuropsychiatric systemic lupus erythematosus. IEEE J Biomed Health Inform. Jul 2022;26(7):3495-3506. [doi: 10.1109/JBHI.2022.3164937] [Medline: 35380977]
17. Luo X, Piao S, Li H, et al. Multi-lesion radiomics model for discrimination of relapsing-remitting multiple sclerosis and neuropsychiatric systemic lupus erythematosus. Eur Radiol. Aug 2022;32(8):5700-5710. [doi: 10.1007/s00330-022-08653-2] [Medline: 35243524]
18. Xu Y, Liu X, Cao X, et al. Artificial intelligence: a powerful paradigm for scientific research. Innovation (Camb). Nov 28, 2021;2(4):100179. [doi: 10.1016/j.xinn.2021.100179] [Medline: 34877560]
19. Marone A, Tang W, Kim Y, et al. Evaluation of SLE arthritis using frequency domain optical imaging. Lupus Sci Med. Aug 2021;8(1):e000495. [doi: 10.1136/lupus-2021-000495] [Medline: 34462335]
20. Zheng Z, Zhang X, Ding J, et al. Deep learning-based artificial intelligence system for automatic assessment of glomerular pathological findings in lupus nephritis. Diagnostics (Basel). Oct 26, 2021;11(11):1983. [doi: 10.3390/diagnostics11111983] [Medline: 34829330]
21. Jorge A, Castro VM, Barnado A, et al. Identifying lupus patients in electronic health records: development and validation of machine learning algorithms and application of rule-based algorithms. Semin Arthritis Rheum. Aug 2019;49(1):84-90. [doi: 10.1016/j.semarthrit.2019.01.002] [Medline: 30665626]
22. Wang DC, Xu WD, Wang SN, et al. Lupus nephritis or not? A simple and clinically friendly machine learning pipeline to help diagnosis of lupus nephritis. Inflamm Res. Jun 2023;72(6):1315-1324. [doi: 10.1007/s00011-023-01755-7] [Medline: 37300586]

23. Adamichou C, Genitsaridi I, Nikolopoulos D, et al. Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. Ann Rheum Dis. Jun 2021;80(6):758-766. [doi: 10.1136/annrheumdis-2020-219069] [Medline: 33568388]
24. Du J, Huang H, Pang L, et al. A machine learning model for identifying systemic lupus erythematosus through laboratory information system and electronic medical record. Clin Exp Rheumatol. Mar 2024;42(3):702-712. [doi: 10.55563/clinexprheumatol/jvdrpc] [Medline: 37976115]
25. Inglese F, Kim M, Steup-Beekman GM, et al. MRI-based classification of neuropsychiatric systemic lupus erythematosus patients with self-supervised contrastive learning. Front Neurosci. 2022;16:695888. [doi: 10.3389/fnins.2022.695888] [Medline: 35250439]
26. Qin X, Xia L, Zhu C, et al. Noninvasive evaluation of lupus nephritis activity using a radiomics machine learning model based on ultrasound. J Inflamm Res. 2023;16:433-441. [doi: 10.2147/JIR.S398399] [Medline: 36761904]
27. Samundeswari S, Ramalinga V, Latha B, Palanivel S. Pattern classification techniques for the classification of cutaneous manifestations of systemic lupus erythematosus. Pak J Biotechnol. Jun 2018;15(2):333-337. URL: https://pjbt.org/index.php/pjbt/article/view/400 [Accessed 2025-01-12]
28. Simos NJ, Dimitriadis SI, Kavroulakis E, et al. Quantitative identification of functional connectivity disturbances in neuropsychiatric lupus based on resting-state fMRI: a robust machine learning approach. Brain Sci. Oct 25, 2020;10(11):777. [doi: 10.3390/brainsci10110777] [Medline: 33113768]
29. Alves P, Bandaria J, Leavy MB, et al. Validation of a machine learning approach to estimate Systemic Lupus Erythematosus Disease Activity Index score categories and application in a real-world dataset. RMD Open. May 2021;7(2):e001586. [doi: 10.1136/rmdopen-2021-001586] [Medline: 34016712]
30. Dey SB, Mansoor N. A butterfly malar rash detection model for early systemic lupus erythematosus diagnosis. Presented at: 2023 26th International Conference on Computer and Information Technology (ICCIT); Dec 13-15, 2023:1-6; Cox's Bazar, Bangladesh. [doi: 10.1109/ICCIT60459.2023.10441593]
31. Lin S, Masood A, Li T, Huang G, Dai R. Deep learning-enabled automatic screening of SLE diseases and LR using OCT images. Vis Comput. Aug 2023;39(8):3259-3269. [doi: 10.1007/s00371-023-02945-4] [Medline: 37361461]
32. Wang M, Liang Y, Hu Z, et al. Lupus nephritis diagnosis using enhanced moth flame algorithm with support vector machines. Comput Biol Med. Jun 2022;145:105435. [doi: 10.1016/j.compbiomed.2022.105435] [Medline: 35397339]
33. Silaide de Araújo Júnior A, Sato EI, Silva de Souza AW, et al. Development of an instrument to predict proliferative histological class in lupus nephritis based on clinical and laboratory data. Lupus. Feb 2023;32(2):216-224. [doi: 10.1177/09612033221143933] [Medline: 36461171]
34. Chang C, Liu H, Chen C, et al. Rapid diagnosis of systemic lupus erythematosus by Raman spectroscopy combined with spiking neural network. Spectrochim Acta A Mol Biomol Spectrosc. Apr 5, 2024;310:123904. [doi: 10.1016/j.saa.2024.123904] [Medline: 38262298]
35. Huang Y, Chen C, Chang C, et al. SLE diagnosis research based on SERS combined with a multi-modal fusion method. Spectrochim Acta A Mol Biomol Spectrosc. Jul 2024;315:124296. [doi: 10.1016/j.saa.2024.124296]
36. U P, M S, E SS, P S, J C. Systemic lupus erythematosus detection using deep learning with auxiliary parameters. Presented at: 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT); Apr 5-7, 2023:1-7; Trichirappalli, India. [doi: 10.1109/ICEEICT56924.2023.10157562]
37. Wang DC, Xu WD, Qin Z, et al. Systemic lupus erythematosus with high disease activity identification based on machine learning. Inflamm Res. Sep 2023;72(9):1909-1918. [doi: 10.1007/s00011-023-01793-1] [Medline: 37725103]
38. Simos NJ, Manikis GC, Papadaki E, Kavroulakis E, Bertsias G, Marias K. Machine learning classification of neuropsychiatric systemic lupus erythematosus patients using resting-state fmri functional connectivity. Presented at: 2019 IEEE International Conference on Imaging Systems and Techniques (IST); Dec 9-10, 2019:1-6; Abu Dhabi, United Arab Emirates. [doi: 10.1109/IST48021.2019.9010078]
39. Tan G, Huang B, Cui Z, Dou H, Zheng S, Zhou T. A noise-immune reinforcement learning method for early diagnosis of neuropsychiatric systemic lupus erythematosus. Math Biosci Eng. Jan 4, 2022;19(3):2219-2239. [doi: 10.3934/mbe.2022104] [Medline: 35240783]
40. Li Q, Yang Z, Chen K, et al. Human-multimodal deep learning collaboration in "precise" diagnosis of lupus erythematosus subtypes and similar skin diseases. J Eur Acad Dermatol Venereol. Dec 2024;38(12):2268-2279. [doi: 10.1111/jdv.20031] [Medline: 38619440]
41. Zhou X, Chen C, Lv X, et al. CMACF: Transformer-based cross-modal attention cross-fusion model for systemic lupus erythematosus diagnosis combining Raman spectroscopy, FTIR spectroscopy, and metabolomics. Inf Process Manag. Nov 2024;61(6):103804. [doi: 10.1016/j.ipm.2024.103804]
42. Li T, Lin S, Guan Z, et al. A deep learning system for detecting systemic lupus erythematosus from retinal images. Cell Rep Med. Jul 15, 2025;6(7):102203. [doi: 10.1016/j.xcrm.2025.102203] [Medline: 40570853]

43. Bosnalı B, Türk E, Öğüt TS, et al. Effectiveness of artificial intelligence in classification of connective tissue diseases in patients with anti-nuclear antibody (ANA) positivity. Comput Biol Chem. Feb 2026;120(Pt 1):108679. [doi: 10.1016/j.compbiolchem.2025.108679] [Medline: 40945129]
44. Li J, Jian C, Zhao J, et al. Machine learning-based multiclass model for autoimmune disease diagnosis and classification through nailfold videocapillaroscopy features. RMD Open. Mar 4, 2026;12(1):e006393. [doi: 10.1136/rmdopen-2025-006393] [Medline: 41781159]
45. Li Z, Li H, Tian B, et al. Free water in the hippocampal cingulum as a Radiomic biomarker for Identifying inflammatory neuropsychiatric Lupus: a cross-sectional case-control study. J Autoimmun. May 2026;160:103560. [doi: 10.1016/j.jaut.2026.103560] [Medline: 41974094]
46. Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. Jun 2020;122:129-141. [doi: 10.1016/j.jclinepi.2019.12.020] [Medline: 32060007]
47. Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. Jun 2020;122:142-152. [doi: 10.1016/j.jclinepi.2019.12.021] [Medline: 32058069]
48. Aliyev E, Ugur Y, Cam V, et al. Closed circuit artificial ıntelligence model named morgaf for childhood onset systemic lupus erythematosus diagnosis. Sci Rep. Jul 1, 2025;15(1):20868. [doi: 10.1038/s41598-025-92964-z] [Medline: 40595005]
49. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. Apr 12, 2022;5(1):48. [doi: 10.1038/s41746-022-00592-y] [Medline: 35413988]
50. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak. Jan 24, 2023;23(1):18. [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
51. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA. Jan 28, 2020;323(4):305-306. [doi: 10.1001/jama.2019.20866] [Medline: 31904799]
52. Hofer IS, Burns M, Kendale S, Wanderer JP. Realistically integrating machine learning into clinical practice: a road map of opportunities, challenges, and a potential future. Anesth Analg. May 2020;130(5):1115-1118. [doi: 10.1213/ANE.0000000000004575] [Medline: 32287118]
53. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test. 2018;2(3):249-262. [doi: 10.1007/s41664-018-0068-2] [Medline: 30842888]
54. Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. Sep 2021;208:106288. [doi: 10.1016/j.cmpb.2021.106288] [Medline: 34352688]
55. Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: a comprehensive review. Health Care (Don Mills). ;10(3):541. [doi: 10.3390/healthcare10030541]
56. Xue P, Wang J, Qin D, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. NPJ Digit Med. Feb 15, 2022;5(1):19. [doi: 10.1038/s41746-022-00559-z] [Medline: 35169217]
57. Arbet J, Brokamp C, Meinzen-Derr J, Trinkley KE, Spratt HM. Lessons and tips for designing a machine learning study using EHR data. J Clin Transl Sci. Jul 24, 2020;5(1):e21. [doi: 10.1017/cts.2020.513] [Medline: 33948244]
58. Kiani AK, Naureen Z, Pheby D, et al. Methodology for clinical research. J Prev Med Hyg. Jun 2022;63(2 Suppl 3):E267-E278. [doi: 10.15167/2421-4248/jpmh2022.63.2S3.2769] [Medline: 36479476]
59. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging. Aug 2017;30(4):392-399. [doi: 10.1007/s10278-017-9976-3] [Medline: 28516233]
60. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal. Oct 2020;65:101759. [doi: 10.1016/j.media.2020.101759] [Medline: 32623277]
61. Grannis SJ, Xu H, Vest JR, et al. Evaluating the effect of data standardization and validation on patient matching accuracy. J Am Med Inform Assoc. May 1, 2019;26(5):447-456. [doi: 10.1093/jamia/ocy191] [Medline: 30848796]
62. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. Apr 7, 2021;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]

63. Tian J, Zhang D, Yao X, Huang Y, Lu Q. Global epidemiology of systemic lupus erythematosus: a comprehensive systematic analysis and modelling study. Ann Rheum Dis. Mar 2023;82(3):351-356. [doi: 10.1136/ard-2022-223035] [Medline: 36241363]
64. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med. Oct 9, 2023;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
65. Ueda D, Yamamoto A, Takashima T, et al. Training, validation, and test of deep learning models for classification of receptor expressions in breast cancers from mammograms. JCO Precis Oncol. Nov 2021;5:543-551. [doi: 10.1200/PO.20.00176] [Medline: 34994603]
66. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Oct 29, 2019;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
67. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. Sep 22, 2023;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
68. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. Dec 18, 2018;169(12):866-872. [doi: 10.7326/M18-1990] [Medline: 30508424]
69. Aringer M, Costenbader K, Daikh D, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol. Sep 2019;71(9):1400-1412. [doi: 10.1002/art.40930] [Medline: 31385462]
70. Chen R, Desai NR, Ross JS, et al. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. BMJ. Feb 17, 2016;352:i637. [doi: 10.1136/bmj.i637] [Medline: 26888209]
71. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (B Aires). May 16, 2021;57(5):503. [doi: 10.3390/medicina57050503]
72. Aydin OU, Taha AA, Hilbert A, et al. An evaluation of performance measures for arterial brain vessel segmentation. BMC Med Imaging. Jul 16, 2021;21(1):113. [doi: 10.1186/s12880-021-00644-x] [Medline: 34271876]
73. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. Feb 4, 2021;14(1):13. [doi: 10.1186/s13040-021-00244-z] [Medline: 33541410]
74. Di Matteo A, Smerilli G, Cipolletta E, et al. Imaging of joint and soft tissue involvement in systemic lupus erythematosus. Curr Rheumatol Rep. Jul 16, 2021;23(9):73. [doi: 10.1007/s11926-021-01040-8] [Medline: 34269905]
75. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J. 2018;16:34-42. [doi: 10.1016/j.csbj.2018.01.001] [Medline: 30275936]
76. Hanna MG, Ardon O, Reuter VE, et al. Integrating digital pathology into clinical practice. Mod Pathol. Feb 2022;35(2):152-164. [doi: 10.1038/s41379-021-00929-0] [Medline: 34599281]
77. Price WN II, Cohen IG. Privacy in the age of medical big data. Nat Med. Jan 2019;25(1):37-43. [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
78. Lu P, Oetjen KA, Bender DE, et al. IMC-Denoise: a content aware denoising pipeline to enhance imaging mass cytometry. Nat Commun. Mar 23, 2023;14(1):1601. [doi: 10.1038/s41467-023-37123-6] [Medline: 36959190]

Abbreviations

ACR: American College of Rheumatology
AUC: area under the curve
CT: computed tomography
DCA: decision-curve analysis
DL: deep learning
DTA: diagnostic test accuracy
ECE: expected calibration error
EEG: electroencephalogram
EHR: electronic health record
EULAR: European League Against Rheumatism
FN: false negative
FODI: finger optical diffusion imaging
FP: false positive
FTIR: Fourier transform infrared spectroscopy
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
HSROC: hierarchical summary receiver operating characteristic

ISN: International Society of Nephrology
LN: lupus nephritis
ML: machine learning
MRI: magnetic resonance imaging
NPSLE: neuropsychiatric systemic lupus erythematosus
NPV: negative predictive value
OCT: optical coherence tomography
PI: prediction interval
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: PRISMA for Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROC: receiver operating characteristic
RPS: Renal Pathology Society
SLE: systemic lupus erythematosus
SLICC: Systemic Lupus Collaborating Clinics
SROC: summary receiver operating characteristic
TN: true negative
TP: true positive

Edited by Stefano Brini; peer-reviewed by Iraj Abedi, Nguyen Quoc Khanh Le; submitted 23.Dec.2025; final revised version received 14.Jul.2026; accepted 15.Jul.2026; published 04.Sep.2026

Please cite as:
Wang B, Wang Z, Liu Y, Ge F
Diagnostic Performance of Machine Learning for Systemic Lupus Erythematosus: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e90209
URL: https://www.jmir.org/2026/1/e90209
doi: 10.2196/90209