



Artificial Intelligence for Diagnosis in Medical Imaging and Pathology: A Systematic Review of Reported Discriminative Performance with an Exploratory Diagnostic Test Accuracy Meta-Analysis

Dmytro Maltsev^{1*}, Hanna Fedoryshyn², Svitlana Myslovska³, Volodymyr Yakymenko⁴, Oleksandr Aleksandruk⁵

¹Department of Modern Technologies of Medical Diagnostics and Treatment, Bogomolets National Medical University, Kyiv, Ukraine

²Diachuk Clinic, LLC “Cera Med”, Kyiv, Ukraine

³Department of Biophysics, Informatics and Medical Equipment, National Pirogov Memorial Medical University, Vinnytsia, Ukraine

⁴Department of Propaedeutics of Internal Medicine, Radiation Diagnostics and Radiation Therapy, Medical Faculty No. 1, Zaporizhzhia State Medical and Pharmaceutical University, Zaporizhzhia, Ukraine

⁵Department of Dermatology and Venereology, Ivano-Frankivsk National Medical University, Ivano-Frankivsk, Ukraine

*Corresponding author: Dmytro Maltsev (e-mail: dmaltsev@ukr.net).

©2026 the Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)

Abstract Background: Artificial intelligence (AI) models are increasingly evaluated for diagnosis in medical imaging and pathology, but reported performance depends on validation strategy. **Methods:** PubMed, Embase, Cochrane Library, IEEE Xplore and Web of Science were searched for open-access English-language studies from January 2021 to February 2026 reporting the area under the receiver operating characteristic curve (AUC) of a diagnostic AI model. AUC values were summarized descriptively by validation strategy, architecture and modality. Studies with 2×2 data from an external validation set were pooled with a bivariate random-effects model. Risk of bias was assessed with the Quality Assessment of Diagnostic Accuracy Studies-2 tool. **Results:** Of 1,503 records screened, 170 studies (one model each) were included. AUC was ≥0.90 in 119 studies (70.0%); mean AUC was 0.945 (internal and external validation), 0.929 (internal), 0.912 (external) and 0.798 (four randomized trials). Seven externally validated studies of unrelated conditions provided 2×2 data (summary sensitivity 92.1% (95% confidence interval 88.5–94.6), specificity 90.1% (81.3–95.0); wide prediction region). Risk of bias was frequently unclear (patient selection 71.8%, flow and timing 64.7%). **Conclusions:** High reported AUC values reflect discrimination in selected, mostly retrospective datasets and do not establish clinical effectiveness or readiness for routine use.

Key Words Artificial Intelligence, Deep Learning, Diagnostic Imaging, Pathology, ROC Curve, Sensitivity and Specificity, Systematic Review, Meta-Analysis

INTRODUCTION

Accurate and timely diagnosis underpins effective patient management because it enables earlier treatment, can improve outcomes and reduces unnecessary investigations and costs [1,2]. Diagnostic error nevertheless remains common: missed, incorrect or delayed diagnoses are reported in 10–15% of cases, autopsy series reveal diagnostic discrepancies in about 20%, and 60–80% of radiological errors are perceptual and associated with fatigue, interruptions or satisfaction of search, while communication failures also contribute substantially [3]. Variability in interpretation between professionals and time-

consuming manual procedures further limit conventional diagnostic workflows, particularly for severe and rapidly progressing conditions [4].

Artificial intelligence (AI) has been proposed as a complement to human interpretation. AI systems can perform repetitive, high-volume image-analysis tasks consistently, whereas clinicians remain better suited to integrating clinical context, managing ambiguity and adapting to unfamiliar situations [5]. In radiology, machine learning (ML) and, in particular, deep learning (DL) with convolutional neural networks (CNNs) have been applied to computed tomography (CT), magnetic resonance imaging

(MRI), radiography, mammography and positron emission tomography to detect and characterize lesions such as tumors and fractures [6,7]; the long-established digital imaging infrastructure of radiology has facilitated this development [4]. In pathology, whole-slide imaging allows AI models to identify cellular and tissue patterns associated with disease [8,9]. Digital pathology has, however, been adopted more slowly because of practical and financial constraints, the continued need to produce and store glass slides, additional staffing and equipment requirements, and image files that are at least ten times larger than radiological images [4,10].

Computer-aided diagnosis (CAD) was introduced to support rather than replace the reader, for example in mammography and in the detection of pulmonary nodules on CT [11], whereas conventional interpretation remains affected by inter-observer variability [12]. Contemporary DL models learn image features directly through stacked convolution and pooling layers [6,7,13], and newer architectures, including residual networks, Inception-type networks, U-Net and vision transformers (ViT), have extended these capabilities [14,15]. Recent examples include a hybrid ViT–deformable CNN model for lung and colon cancer histopathology [16], a CNN framework for multi-view chest radiographs [17] and a fine-tuned ResNet34 model for brain tumor classification [18]. Applications have also been described in neuroradiology and musculoskeletal radiology [19] and in liver, skin and kidney pathology [20–22]. A meta-analysis of AI in digital pathology reported high summary sensitivity and specificity but also substantial heterogeneity and a high prevalence of risk-of-bias concerns [23].

High reported performance should not be equated with clinical benefit. In a meta-analysis of 83 studies, generative AI models achieved an overall diagnostic accuracy of 52.1%, with no significant difference from physicians overall ($p = 0.10$) but significantly lower accuracy than expert physicians ($p = 0.007$) [24]. Methodological reviews of AI diagnostic research have repeatedly identified retrospective designs, selection bias, homogeneous development datasets with a risk of overfitting, limited external validation and incomplete reporting [23,25], and the risks of over-reliance on AI in diagnostic practice have been highlighted [26]. Independent validation frameworks have been proposed to screen AI models before clinical use [27], and AI-specific reporting extensions, prospective registration and external validation have been recommended [25]. Previous reviews have also differed in their approach to synthesis: a review of externally validated CNN models for emergency head CT did not pool results because of heterogeneity [28], whereas a review of AI in digital pathology applied bivariate pooling [23].

It therefore remains unclear how reported discrimination varies with validation strategy across the rapidly growing literature on diagnostic AI in imaging and pathology, and how much of this literature provides the data required for formal diagnostic test accuracy (DTA) synthesis. The primary objective of this systematic review

was to describe the reported discriminative performance, expressed as the area under the receiver operating characteristic curve (AUC), of AI models used for diagnosis in medical imaging and pathology and published between January 2021 and February 2026, stratified by validation strategy, model architecture and imaging modality. The secondary objective was to estimate, in an exploratory DTA meta-analysis, the summary sensitivity and specificity of AI models evaluated on external validation sets with extractable 2×2 data. External validation was therefore a stratification variable for the primary objective and an eligibility criterion for the secondary objective.

METHODS

Protocol and Registration

This systematic review is reported in accordance with the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) statement [29]; the completed checklist is provided as Supplementary File 2. The review was not registered in PROSPERO or in any other registry, and the review protocol was not published. The eligibility criteria, outcomes and methods are reported in full in this article and its Supplementary Material.

Review Question

The review question was structured according to the population, index test, reference standard and target condition framework. The population comprised patients, or specimens obtained from patients, who underwent diagnostic imaging or pathological examination, without restriction by age, sex, disease severity or clinical setting. The index test was an AI model (conventional ML, DL including CNN and transformer-based architectures, or a combination) applied to a diagnostic classification task. The reference standard was that used in each primary study, such as histopathology, expert or consensus reading, clinical follow-up or a confirmed clinical diagnosis. The target condition was the disease or abnormality defined by each primary study. The primary outcome was the AUC; the secondary outcomes were sensitivity and specificity derived from 2×2 tables for an external validation set and the summary receiver operating characteristic (SROC) curve.

Eligibility Criteria

Peer-reviewed primary studies were eligible if they were published in English between January 2021 and February 2026, had freely accessible full text, evaluated a diagnostic AI model in the population defined above and reported the AUC of the model. The start date of 2021 was chosen to capture recent deep learning architectures. Reports were excluded when diagnostic accuracy could not be determined, when the application was not related to imaging or pathology, when the study design was inappropriate, or when the AI model was unavailable or had been withdrawn. Studies included in the review were eligible for the DTA meta-analysis if they evaluated the model on an external

validation set (Section 2.4) for a binary target condition and reported sufficient information to construct a 2×2 table of true-positive (TP), false-positive (FP), false-negative (FN) and true-negative (TN) results against the reference standard for that set.

English was selected because most AI diagnostic studies are published in English and to avoid translation errors. Only articles with freely accessible full text were included, because the full text was required to extract the AUC and the validation design. Open-access status is not related to diagnostic validity, and this criterion may have introduced availability and publication bias, because studies published in subscription journals may differ in design, setting and reported performance; its potential impact is considered in Section 4.6.

Operational Definition of Validation Strategy

External validation was defined as evaluation of a trained model on data that were not used for model development (training, tuning or internal testing) and that originated from a different institution or site, a different time period, or an independent dataset. Internal validation comprised split-sample, cross-validation or bootstrap evaluation within the development dataset. Each study was classified into one of four categories according to the validation reported in its methods: internal validation, external validation, internal and external validation, or evaluation within a prospective randomized controlled trial (RCT). The same definition is applied throughout the Results, tables and Supplementary Material.

Information Sources and Search Strategy

PubMed (MEDLINE), Embase, the Cochrane Library, IEEE Xplore and Web of Science were searched for records published from January 2021 to February 2026. The search strings combined free-text terms for AI (artificial intelligence, machine learning, deep learning, neural network*, convolutional neural network, CNN), for imaging and pathology (radiology, medical imaging, imaging, medical image, pathology, histopathology, cytology), for diagnostic accuracy (diagnosis, diagnostic accuracy, sensitivity, specificity) and for the diagnostic procedure (pathological examination, diagnostic imaging), linked with Boolean operators. Controlled vocabulary (MeSH or Emtree) was not used. The terms differed between databases: the Embase and Web of Science strings did not include the pathology, histopathology, cytology, sensitivity or specificity terms, and the IEEE Xplore string used the All Metadata field.

Database limits were applied as follows. In PubMed, the limits were publication within the last five years, free full text, English language, humans, MEDLINE, and the article types clinical study, clinical trial, observational study and randomized controlled trial. In the Cochrane Library, the limits were the last five years, trials, English language and humans. Embase, IEEE Xplore and Web of Science were limited to 2021–2026, and the Web of Science search was

restricted to six journals (BMC Cancer, Cancer Medicine, EBioMedicine, Insights into Imaging, Diagnostic and Interventional Radiology and JMIR Medical Informatics). The complete search strings, limits and numbers of retrieved records for each database are reported in Supplementary Table S1.

Study Selection

Before screening, duplicate records, records marked as ineligible by automation tools and records that could not be accessed were removed. Titles and abstracts were then screened with AUC as the principal screening outcome; some records were excluded with an AI-based screening tool and the remainder by the reviewers. Potentially eligible reports were assessed in full text against the eligibility criteria.

Data Extraction

Two reviewers independently extracted data using a standardized Microsoft Excel form: first author, year of publication, study design, target condition, imaging or pathology modality, index test characteristics (AI architecture and validation strategy), reference standard, flow and timing, AUC and, where available, sensitivity, specificity and 2×2 data. AUC was extracted for the external validation set when one was reported and otherwise for the internal validation set. Disagreements were resolved by discussion between the two reviewers until consensus was reached. The F-score was initially included in the extraction form but was removed because it was inconsistently reported. When key metrics were missing, the corresponding authors were contacted by e-mail; studies without a reply after two attempts were retained in the review with the data available. Data were managed in Microsoft Excel (Power Query and PivotTables).

During the revision, the data of the nine studies considered for the meta-analysis were re-extracted from the full texts. TP, FP, FN and TN were taken from reported confusion matrices or calculated from the reported sensitivity, specificity and numbers of units with and without the target condition, and each calculated table was checked against the sensitivity, specificity, accuracy and confidence intervals reported in the study. When a study reported more than one external validation set, the largest set with extractable 2×2 data was used, and the AUC recorded for the study was taken from the same set. This re-extraction corrected the validation category of five studies, the AUC of three studies and the architecture or modality label of seven studies; the corrected values are used throughout this article.

Unit of Analysis

The unit of analysis was the study. Each included study contributed a single record, consisting of one AI model evaluated on one validation set; no publication contributed more than one model, dataset, validation set or outcome, so no participant or image set was counted more than once. When a publication reported several validation sets, the

result for the external validation set was extracted. In the meta-analysis, each study contributed one 2×2 table. The unit of observation within studies (patients, lesions, images, slides or smears) differed between studies and is reported in Table 3.

Outcome Definitions and AUC Categories

For descriptive purposes, AUC values were grouped into four study-defined categories: <0.80, 0.80–0.89, 0.90–0.94 and ≥0.95. These cut-offs are not validated clinical thresholds; they were chosen to describe the distribution of the reported values, most of which exceeded 0.80, and only the highest category corresponds to the value (≥0.90) that a commonly cited interpretive scale describes as excellent discrimination [30]. The AUC summarizes discrimination across all thresholds and does not describe calibration, performance at the operating threshold used in practice, clinical utility or performance relative to clinicians.

Risk of Bias and Applicability

Both reviewers independently assessed each included study with the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool [31] in four domains (patient selection, index test, reference standard, and flow and timing), each judged as low, high or unclear risk of bias on the basis of the signaling questions. For the reference standard domain, these questions address whether the reference standard is likely to classify the target condition correctly and whether its results were interpreted without knowledge of the index test. Concerns regarding applicability were judged separately from risk of bias and were recorded for the patient-selection domain. Inter-reviewer agreement on the QUADAS-2 judgments was $\kappa = 0.70$, indicating substantial agreement [32]; disagreements were resolved by discussion.

Statistical Analysis

Systematic review. Reported AUC values were summarized as frequencies and percentages per category and as mean (standard deviation, SD) and median (interquartile range, IQR) by validation strategy, architecture group and modality. No formal statistical comparisons were made between validation strategies, architectures or modalities, because each group combined different target conditions, datasets, reference standards and study designs; differences between groups are therefore reported descriptively. Study-reported AUC values were not pooled, because they were obtained for different tasks, thresholds and datasets and were not accompanied by consistently reported measures of precision.

Meta-analysis. Sensitivity and specificity were pooled jointly with the bivariate random-effects model of Reitsma *et al.* [33], which models logit-transformed sensitivity and false-positive rate (FPR, $1 - \text{specificity}$) with their between-study variances and correlation, fitted by restricted maximum likelihood. A continuity correction of 0.5 was added to all cells of all studies because two studies contained

a zero cell. The SROC curve was derived from the bivariate model parameters with the Rutter–Gatsonis parameterization [34] and is presented with the 95% confidence region for the summary point and the 95% prediction region; the area under the SROC curve (SROC AUC) is reported only as a descriptive summary of the fitted curve and was not obtained by pooling study-level AUC values. Heterogeneity was quantified by the between-study standard deviations (τ) of logit sensitivity and logit FPR, their correlation and the prediction region; univariate I^2 values are reported descriptively for the forest plots only. Forest plots show study-level estimates with exact binomial 95% confidence intervals (CIs).

The pooled estimate combines studies of different target conditions, modalities, reference standards and architectures and was therefore calculated as an exploratory summary of the accuracy of externally validated AI models in this set of studies, not as an estimate for any specific clinical application. Two sensitivity analyses were performed: restriction to studies whose external validation set came from institutions not involved in model development, and leave-one-out analysis. Subgroup analyses and meta-regression were not performed because no target condition was represented by more than one study, and Deeks' funnel plot asymmetry test was not performed because fewer than ten studies were included [35].

The meta-analysis was performed in R 4.3.3 with the metafor package 4.4.0 (rma.mv function); the code is provided as Supplementary File 4. Descriptive statistics were calculated in Python 3.12 with pandas 3.0 and cross-checked with the spreadsheet formulas in the study-level data file (Supplementary File 3). Figures were drawn with matplotlib 3.10.

RESULTS

Study Selection

The searches identified 1,540 records: 691 in PubMed, 606 in Embase, 209 in the Cochrane Library, 23 in IEEE Xplore and 11 in Web of Science. Before screening, 37 records were removed (11 duplicates, 10 records marked as ineligible by automation tools and 16 records that could not be accessed). Of the 1,503 records screened, 736 were excluded (229 with the AI-based screening tool and 507 by the reviewers). Of the 767 reports sought for retrieval, 2 could not be retrieved, and 765 were assessed in full text. After exclusion of 595 reports (diagnostic accuracy not determined, $n = 413$; not related to imaging or pathology, $n = 145$; inappropriate study design, $n = 34$; AI model unavailable or withdrawn, $n = 3$), 170 studies were included in the review (Figure 1). Of the nine studies considered for the meta-analysis, seven provided 2×2 data from an external validation set and were included; the remaining 163 studies contributed to the descriptive synthesis only (Section 3.6).

Characteristics of the Included Studies

The 170 studies were published between 2021 and 2026 (2021, $n = 3$; 2022, $n = 17$; 2023, $n = 20$; 2024, $n = 45$; 2025,

Table 1: Reported AUC Category by Validation Strategy (n = 170 Studies; One Record Per Study)

Validation strategy	n	AUC ≥ 0.95	AUC 0.90–0.94	AUC 0.80–0.89	AUC < 0.80	Mean $\pm$ SD	Median (IQR)
Internal validation	78	47 (60.3)	11 (14.1)	14 (17.9)	6 (7.7)	0.929 $\pm$ 0.077	0.960 (0.893–0.982)
External validation	72	28 (38.9)	19 (26.4)	20 (27.8)	5 (6.9)	0.912 $\pm$ 0.075	0.920 (0.860–0.980)
Internal and external validation	15	9 (60.0)	3 (20.0)	2 (13.3)	1 (6.7)	0.945 $\pm$ 0.055	0.970 (0.925–0.982)
Prospective RCT	4	1 (25.0)	0 (0.0)	0 (0.0)	3 (75.0)	0.798 $\pm$ 0.126	0.760 (0.735–0.823)
Not recorded	1	1 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0.983	–
Total	170	86 (50.6)	33 (19.4)	36 (21.2)	15 (8.8)	0.921 $\pm$ 0.078	0.950 (0.880–0.980)

Values are n (row %) unless stated otherwise. AUC categories are study-defined descriptive groupings, not clinical thresholds. AUC, area under the receiver operating characteristic curve; IQR, interquartile range; RCT, randomized controlled trial; SD, standard deviation

Table 2: QUADAS-2 Risk of Bias and Concerns Regarding Applicability (n = 170 studies)

Domain	Low	Unclear	High
Risk of bias			
Patient selection	48 (28.2)	122 (71.8)	0 (0.0)
Index test	162 (95.3)	8 (4.7)	0 (0.0)
Reference standard	169 (99.4)	1 (0.6)	0 (0.0)
Flow and timing	60 (35.3)	110 (64.7)	0 (0.0)
Concerns regarding applicability			
Patient selection	0 (0.0)	170 (100.0)	0 (0.0)
Index test	Not assessed	Not assessed	Not assessed
Reference standard	Not assessed	Not assessed	Not assessed

Values are n (%). Per-study judgments are reported in Supplementary Table S4 and in the traffic-light plot (Supplementary Figure S2). QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2

Table 3: Characteristics of the Seven Studies Included in the Diagnostic Test Accuracy Meta-Analysis

Study	Target condition	Modality	Index test	External validation set; units	Reference standard
Shi <i>et al.</i> [39]	Nasopharyngeal carcinoma (malignant vs non-malignant)	Endoscopic images	Swin Transformer (STND)	Prospective, 10 hospitals; 1,885 images from 945 patients	Biopsy (carcinoma, hyperplasia)
Yang <i>et al.</i> [40]	Adhesive capsulitis	MRI (PD and T2)	Radiomics with LightGBM	Second hospital; 57 patients	Clinical diagnostic criteria
Valesano <i>et al.</i> [41]	STIC or STIL	WSI	Attention-based MIL (UNI encoder)	Independent cohort, development institution; 43 cases	Pathology with p53 and Ki-67 IHC
Liu H <i>et al.</i> [42]	Breast cancer	ABVS	ResNet50, three-plane fusion	Third hospital; 119 lesions	Histopathology (biopsy or surgery)
Lee <i>et al.</i> [43]	Myocardial infarction	ECG images	Dedicated ECG AI (STEMI output)	Public dataset from another country; 928 ECG images	Dataset labels (clinical diagnosis)
Xu <i>et al.</i> [44]	DDH (Graf type III or IV)	Ultrasound (videos)	HSNN (YOLOv11m and ResNet50)	Prospective, 4 hospitals; 151 infants	Graf classification, expert consensus
Wang J <i>et al.</i> [45]	Thyroid nodules (TBSRTC III+ vs benign)	FNA cytology WSI	ThyroPower ensemble	Prospective temporal set; 537 smears from 469 patients	TBSRTC diagnosis by senior cytopathologists

Units of analysis are those of the external validation set used for the 2x2 table. ABVS, automated breast volume scanner; DDH, developmental dysplasia of the hip; ECG, electrocardiogram; FNA, fine-needle aspiration; HSNN, HipSonoNeuNet; IHC, immunohistochemistry; MIL, multiple-instance learning; MRI, magnetic resonance imaging; PD, proton density; STEMI, ST-elevation myocardial infarction; STIC, serous tubal intraepithelial carcinoma; STIL, serous tubal intraepithelial lesion; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology; WSI, whole-slide image

n = 81; 2026, n = 4). The most frequent modalities were CT (n = 35), MRI (n = 21), radiography (n = 19), ultrasound (n = 16) and whole-slide imaging (n = 12). Eleven records were based primarily on non-image data or had no recorded modality (electronic health record [EHR] reports, natural language processing, plasma lipidomics, clinical metadata or other non-imaging data), including multimodal models that fused EHR data with radiographs [36]; these records are retained in the descriptive synthesis and are identified in Supplementary Table S2. Internal validation was reported in 78 studies (45.9%), external validation in 72 (42.4%), internal and external validation in 15 (8.8%) and evaluation within a prospective RCT in 4 (2.4%); the validation strategy was not recorded for 1 study (0.6%). Thus, 87 studies (51.2%) reported any external validation. The characteristics of all included studies are listed in Supplementary Table S2.

Distribution of Reported AUC Values by Validation Strategy

The median reported AUC was 0.950 (IQR 0.880–0.980; range 0.680–1.000). AUC was ≥ 0.95 in 86 studies (50.6%), 0.90–0.94 in 33 (19.4%), 0.80–0.89 in 36 (21.2%) and < 0.80 in 15 (8.8%); overall, 119 studies (70.0%) reported an AUC ≥ 0.90 . These values are study-level descriptive summaries of reported results obtained for different tasks, datasets and validation designs; they are not independent estimates of diagnostic accuracy and were not pooled.

Validation strategy and AUC category are different variables and are cross-tabulated in Table 1. Of the 86 studies with an AUC ≥ 0.95 , 47 reported internal validation, 28 external validations, 9 internal and external validation, 1 an RCT and 1 did not record the validation strategy; thus 37 of these 86 studies (43.0%) reported any external validation. An AUC ≥ 0.95 was reported by 60.3% of studies with

Table 4: 2x2 Data, Study-Level Accuracy and Summary Estimates of the Diagnostic Test Accuracy Meta-Analysis

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Shi <i>et al.</i> [39]	612	169	56	1048	0.916 (0.893–0.936)	0.861 (0.840–0.880)
Yang <i>et al.</i> [40]	15	14	1	27	0.938 (0.698–0.998)	0.659 (0.494–0.799)
Valesano <i>et al.</i> [41]	3	1	0	39	1.000 (0.292–1.000)	0.975 (0.868–0.999)
Liu H <i>et al.</i> [42]	53	11	4	51	0.930 (0.830–0.981)	0.823 (0.705–0.908)
Lee <i>et al.</i> [43]	231	20	8	669	0.967 (0.935–0.985)	0.971 (0.956–0.982)
Xu <i>et al.</i> 2025	6	12	0	133	1.000 (0.541–1.000)	0.917 (0.860–0.957)
Wang J <i>et al.</i> [44]	235	16	27	259	0.897 (0.854–0.931)	0.942 (0.907–0.966)
Summary estimate (7 studies)	-	-	-	-	0.921 (0.885–0.946)	0.901 (0.813–0.950)
External institutions only (5 studies)	-	-	-	-	0.925 (0.881–0.953)	0.880 (0.746–0.948)
Leave-one-out range	-	-	-	-	0.912–0.928	0.878–0.921

Study-level 95% CIs are exact (Clopper–Pearson). Summary estimates are from the bivariate random-effects model (restricted maximum likelihood) with a continuity correction of 0.5 added to all cells. τ , logit sensitivity 0.376; τ , logit FPR 0.931; correlation -0.700 ; SROC AUC 0.957. CI, confidence interval; FN, false negative; FP, false positive; FPR, false-positive rate; SROC, summary receiver operating characteristic; TN, true negative; TP, true positive

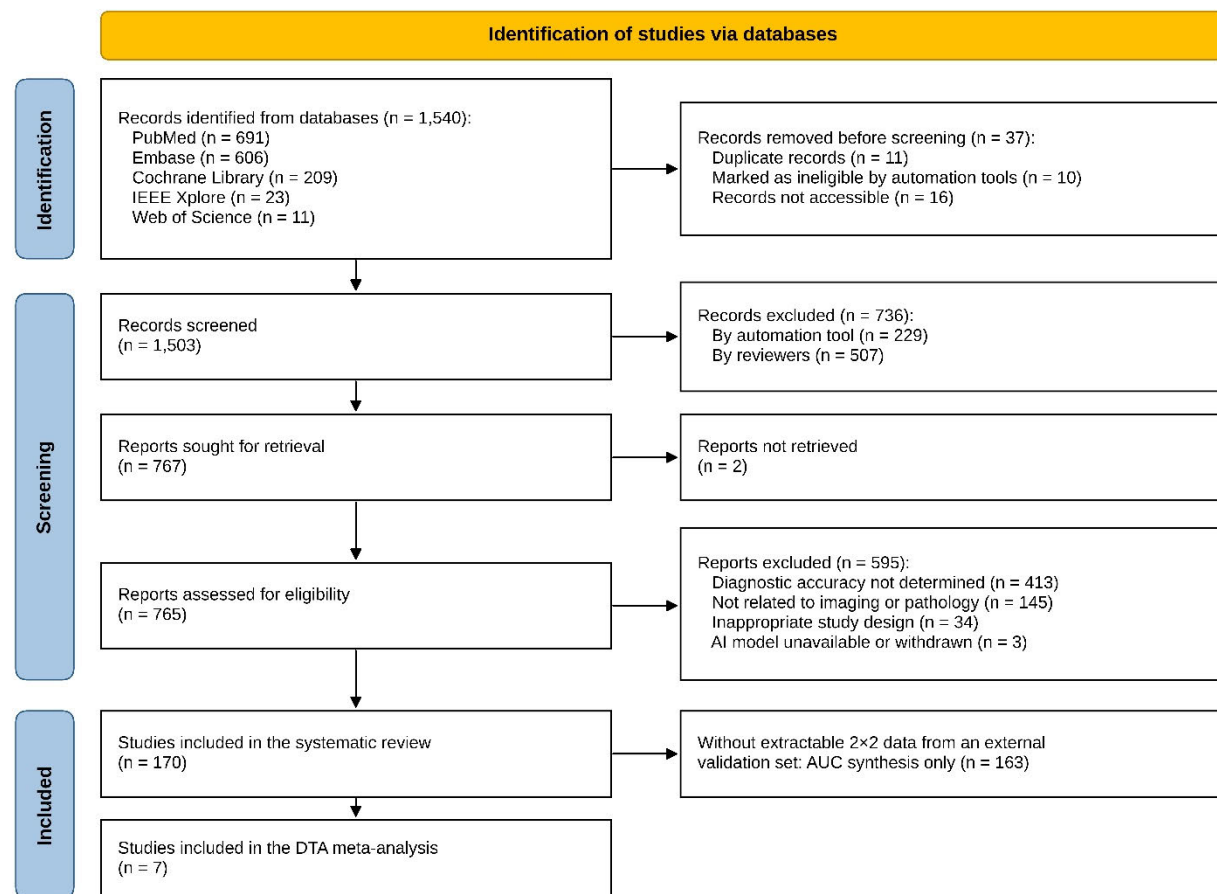


Figure 1: PRISMA-DTA flow Diagram of Study Selection. AI, Artificial Intelligence; DTA, Diagnostic Test Accuracy; PRISMA-DTA, Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies

internal validation, 38.9% with external validation, 60.0% with internal and external validation and 1 of the 4 RCTs. The mean AUC was 0.945 ± 0.055 for studies with internal and external validation, 0.929 ± 0.077 for internal validation, 0.912 ± 0.075 for external validation and 0.798 ± 0.126 for RCTs.

Reported AUC Values by AI Architecture

The median reported AUC was 0.959 (IQR 0.900–0.985) for CNN-based DL models ($n = 65$), 0.970 (0.915–0.982) for

other or unspecified DL models ($n = 35$), 0.950 (0.920–0.990) for transformer-based or hybrid CNN–transformer models ($n = 9$), 0.965 (0.960–0.980) for combined DL and ML approaches ($n = 9$) and 0.900 (0.840–0.953) for conventional ML models, including radiomics ($n = 45$). Two studies evaluated GPT-4 (AUC 0.770 and 1.000), and the two conventional CAD systems, both evaluated within RCTs, reported AUCs of 0.69 and 0.77. DL-based models therefore reported higher median AUC values than conventional ML models. Because the groups differed in

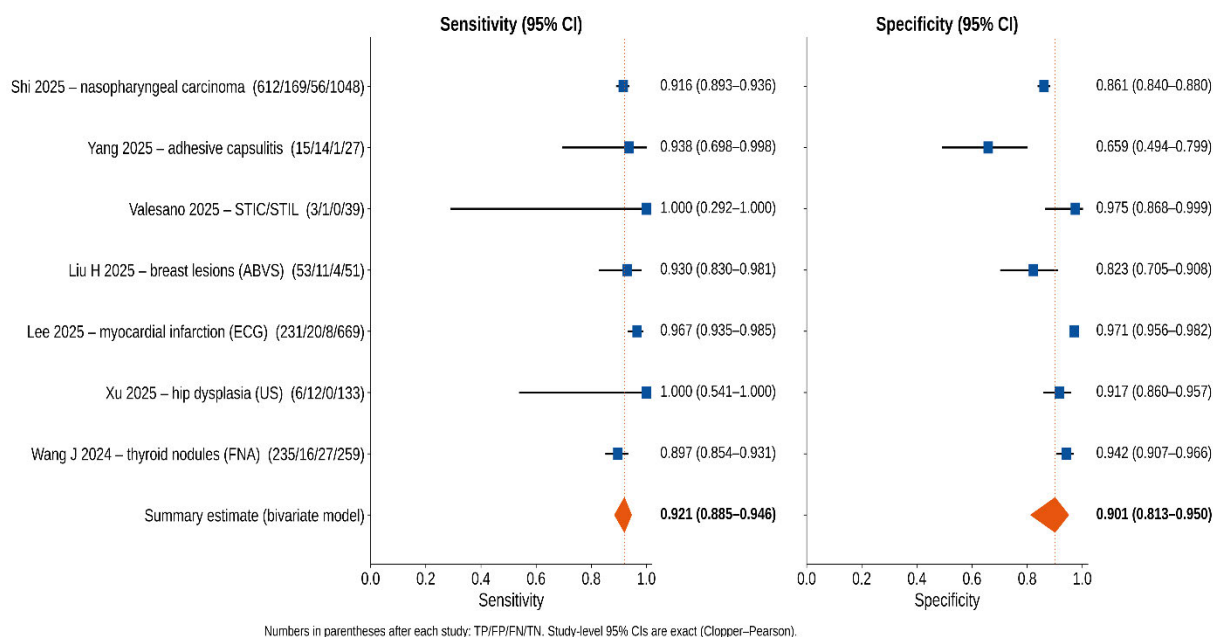


Figure 2: Paired Forest Plots of Sensitivity (Left) and Specificity (Right) with Exact 95% Confidence Intervals for the Seven Studies in the Meta-Analysis. Studies are Identified by First Author, Year and Target Condition, followed by TP/FP/FN/TN; The Diamonds Show the Summary Estimates from the Bivariate Random-Effects Model. CI, Confidence Interval; FN, False Negative; FP, False Positive; TN, True Negative; TP, True Positive

target conditions, datasets and validation strategies and no formal comparative analysis was performed, this difference does not demonstrate the superiority of any architecture. The results for each architecture, as extracted, and the grouping used are reported in Supplementary Table S3.

Risk of Bias and Applicability

QUADAS-2 judgments for all 170 studies are summarized in Table 2 and reported per study in Supplementary Table S4 and Supplementary Figures S1 and S2. No study was judged to be at high risk of bias in any domain. Risk of bias was unclear for patient selection in 122 studies (71.8%), mainly because of retrospective designs, the use of publicly available datasets and the absence of information on consecutive or random sampling. Risk of bias for flow and timing was unclear in 110 studies (64.7%) because of multiple cohorts or datasets and incomplete reporting of exclusions and of the interval between the index test and the reference standard. Risk of bias for the index test was low in 162 studies (95.3%) and unclear in 8 (4.7%). Risk of bias for the reference standard was low in 169 studies (99.4%) and unclear in 1 (0.6%).

Reference standards were heterogeneous and included histopathology, radiologist consensus, clinical diagnosis and EHR-confirmed outcomes. Reference standards based on clinical diagnosis, EHR records or reader consensus may misclassify the target condition, so the predominance of low-risk judgments in this domain should be interpreted with caution and is considered a limitation (Section 4.6). Concerns regarding applicability of patient selection were judged unclear in all 170 studies.

Diagnostic Test Accuracy Meta-Analysis

Nine studies had been considered for the meta-analysis. After re-extraction from the full texts, two were excluded: in one, the test set was a random split of the single-centre development dataset (internal validation) [37], and in the other, the test set was a random split of public datasets and the task was multiclass classification of blood-cell types without a binary target condition [38]. Seven studies were included, evaluating AI models for nasopharyngeal carcinoma on endoscopic images [39], adhesive capsulitis on MRI [40], serous tubal intraepithelial lesions on whole-slide images [41], breast cancer on automated breast volume scanner images [42], myocardial infarction on electrocardiogram images [43], developmental dysplasia of the hip on ultrasound [44] and thyroid nodules on fine-needle aspiration slides [45] (Table 3). The external validation sets comprised 3,720 units of analysis (images, patients, lesions, cases or smears); five came from institutions not involved in model development, one was an independent cohort from the development institution and one was a prospective temporal validation set. Risk of bias was low for patient selection in six of the seven studies and unclear in one, and risk of bias for flow and timing was unclear in all seven.

Study-level sensitivity ranged from 0.897 to 1.000 and specificity from 0.659 to 0.975 (Table 4, Figure 2). The summary sensitivity was 92.1% (95% CI 88.5–94.6) and the summary specificity was 90.1% (95% CI 81.3–95.0). The between-study standard deviation was 0.376 for logit sensitivity and 0.931 for logit FPR, indicating greater heterogeneity in specificity than in sensitivity, and the correlation between logit sensitivity and logit FPR was

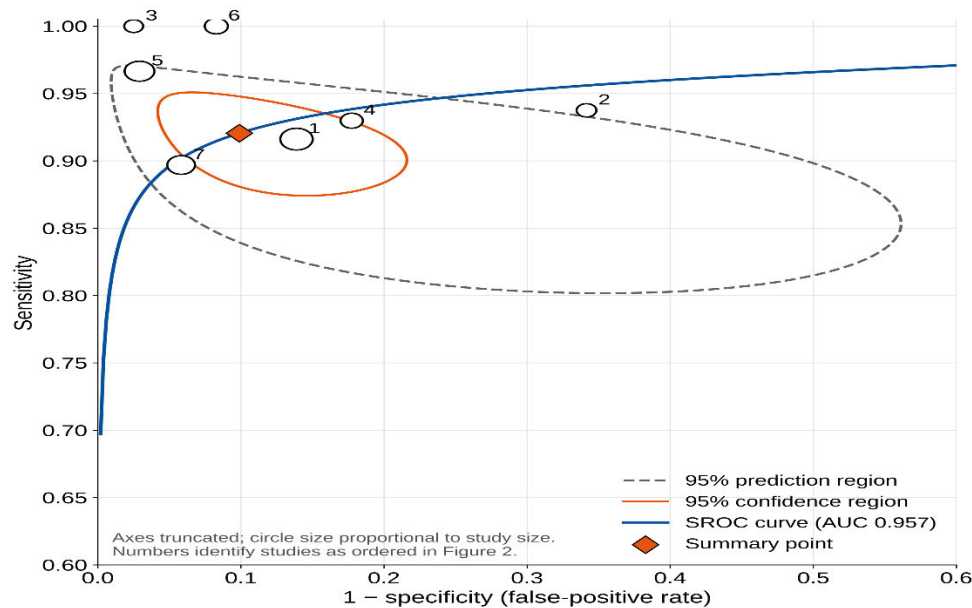


Figure 3: Summary Receiver Operating Characteristic (SROC) Curve from the Bivariate Random-Effects Model, with the Study Estimates (Circle Size Proportional to Study Size), the Summary Point, the 95% Confidence Region and the 95% Prediction Region. Axes are Truncated. AUC, Area under the Curve

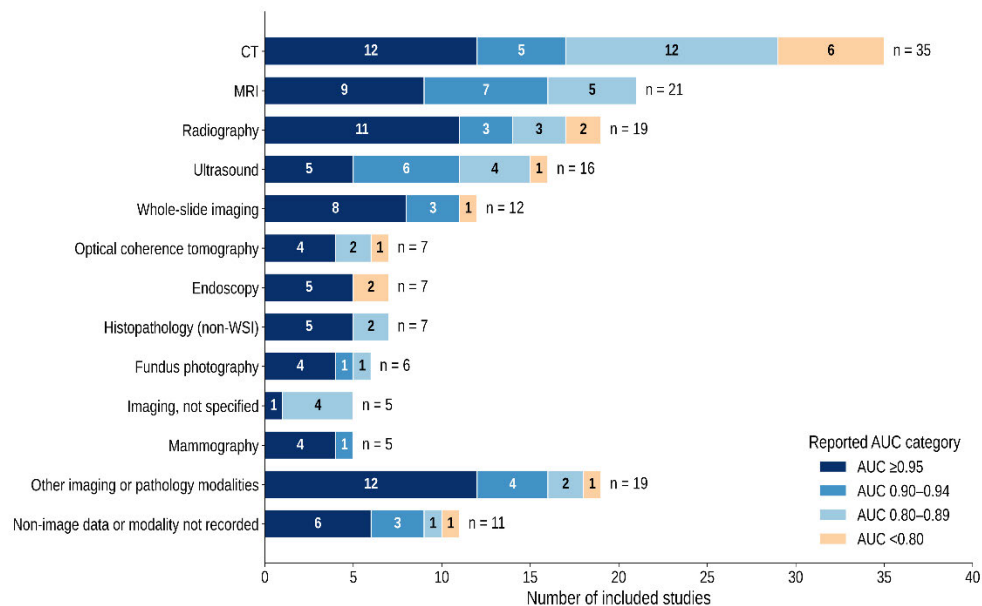


Figure 4: Number of Included Studies by Modality and Reported AUC Category (N = 170). Modalities with Fewer than Five Studies are Combined as Other Imaging or Pathology Modalities; Records Based Primarily on Non-Image Data or Without a Recorded Modality are Shown Separately. AUC Categories are Study-Defined Descriptive Groupings. AUC, Area under the Receiver Operating Characteristic Curve; Ct, Computed Tomography; Mri, Magnetic Resonance Imaging; Wsi, Whole-Slide Imaging

–0.700, giving no indication of a threshold effect. The 95% prediction region extended from about 80.2% to 97.1% for sensitivity and from 43.9% to 99.1% for specificity (Figure 3), so the accuracy of a new externally validated model could differ substantially from the summary point. The SROC AUC was 0.957. Univariate I^2 was 50.5% for sensitivity and 94.1% for specificity. Restricting the analysis to the five studies

validated at institutions not involved in model development gave a summary sensitivity of 92.5% (95% CI 88.1–95.3) and specificity of 88.0% (74.6–94.8); omitting one study at a time gave summary sensitivities of 91.2–92.8% and specificities of 87.8–92.1% (Supplementary Table S8). In two studies, the operating threshold was selected in the validation data with the Youden index [40,43].

Reported AUC Values by Imaging Modality

The distribution of AUC categories by modality is shown in Figure 4 and in Supplementary Tables S6 and S7. CT contributed the largest number of studies (AUC ≥ 0.95 , $n = 12$; 0.90–0.94, $n = 5$; 0.80–0.89, $n = 12$; < 0.80 , $n = 6$), followed by MRI (9, 7, 5 and 0), radiography (11, 3, 3 and 2) and ultrasound (5, 6, 4 and 1). Among the 86 studies with an AUC ≥ 0.95 , the most frequent modalities were CT ($n = 12$), radiography ($n = 11$), MRI ($n = 9$) and whole-slide imaging ($n = 8$); the remaining 46 studies were distributed across 20 other categories, each containing one to five studies. This distribution largely reflects the number of studies available for each modality, and the data do not support conclusions about modality-specific performance.

DISCUSSION

Principal Findings

This review identified 170 studies published between 2021 and 2026 that reported the AUC of AI models for diagnosis in medical imaging and pathology. Reported AUC values were high, with 70.0% of studies reporting an AUC ≥ 0.90 . These values were, however, obtained for heterogeneous target conditions and datasets, most studies were retrospective, and risk of bias was unclear for patient selection in 71.8% and for flow and timing in 64.7% of studies. Only seven studies provided 2x2 data from an external validation set, and these addressed clinically unrelated target conditions. The findings therefore describe reported discrimination in selected datasets rather than the performance expected in clinical practice. Similar concerns about the quality of evidence and the scarcity of external validation were raised in an earlier systematic comparison of DL models with health-care professionals [46].

Validation Strategy and Generalizability

Studies with only internal validation reported an AUC ≥ 0.95 more often than studies with external validation (60.3% vs 38.9%) and had a higher mean AUC (0.929 vs 0.912), and the four studies conducted within RCTs reported the lowest values (mean 0.798). The highest mean AUC was observed in the small group of 15 studies that reported both internal and external validation (0.945). These differences are descriptive and are not adjusted for task or dataset, but their direction is consistent with previous evidence: in a systematic review of DL algorithms for radiological diagnosis, 81% of algorithms showed some decline in performance on external datasets, 49% at least a modest decline and 24% a substantial decline [47]. Internal validation can overestimate performance through overfitting and data leakage, whereas external data introduce differences in populations, equipment and acquisition protocols that more closely reflect real-world use [47]. Generalizability is also difficult to assess because many DL studies provide limited demographic and clinical information about their development datasets and because the decision processes of DL models are not fully transparent [47]. With only four RCTs, no conclusion can be drawn about prospective performance.

Interpretation of the Pooled Estimate and Clinical Heterogeneity

The seven studies in the meta-analysis addressed unrelated target conditions (nasopharyngeal carcinoma, adhesive capsulitis, serous tubal intraepithelial lesions, breast cancer, myocardial infarction, developmental dysplasia of the hip and thyroid nodules) with different modalities, reference standards, units of analysis and architectures. The summary sensitivity of 92.1% and specificity of 90.1% therefore describe the average accuracy of this particular set of externally validated models and cannot be applied to any specific condition, test or setting. The prediction region, which covered specificities from about 43.9% to 99.1%, is more informative than the summary point: the specificity of a new externally validated model could be considerably lower. The large between-study variance of logit FPR and the high univariate I^2 for specificity reflect this heterogeneity, whereas sensitivity was more consistent. The summary estimates changed little in the sensitivity analyses, but with seven studies and one study per target condition, the analysis cannot identify sources of heterogeneity. A systematic review of externally validated CNN models for emergency head CT likewise considered meta-analysis inappropriate because of heterogeneity; scan-level sensitivity exceeded 90% in five of six studies, whereas specificity ranged from 58.0% to 97.7% [28], and a review of AI in digital pathology reported high summary accuracy together with substantial heterogeneity and frequent risk-of-bias concerns [23]. Two further issues may make the summary estimates optimistic: in two studies the operating threshold was selected in the validation data, and studies that report complete 2x2 data may differ systematically from those that report only AUC. The pooled estimate should therefore be regarded as exploratory.

AI Architectures and Imaging Modalities

DL models, particularly CNN-based and transformer-based architectures, reported higher median AUC values than conventional ML models. CNNs learn hierarchical image features directly from the data instead of relying on handcrafted features, which may partly explain this pattern [6,7,13]. Examples in the recent literature include AlexNet-based classification of femoral bone tumors on MRI [48], a multimodal system for eye emergency triage [49], hybrid ResNet-50–ViT models for acute lymphoblastic leukemia [50], a combined support vector machine and VGGNet-16 model for lung cancer [51], U-Net-based segmentation of spinal MRI and knee CT [52,53], YOLO-based fracture detection on CT [54], radiomics and hybrid radiomics–DL pipelines [55,56], and DL and ensemble models for lung cancer, chronic kidney disease and endometrial cancer on CT and for lumbar and brain disorders on MRI [57–62]. However, the architecture groups in this review differed in target conditions, datasets and validation strategies, and no formal comparison was performed, so higher reported AUC values do not establish the superiority of DL over ML.

Radiomics-based models have shown promising results, for example in hepatocellular carcinoma, but heterogeneity in methods and outcomes limits the transferability of such models between studies [63], and similar concerns regarding heterogeneity and applicability have been raised for CNN models in emergency head CT [28].

CT, MRI, radiography, ultrasound and whole-slide imaging accounted for most studies in every AUC category, which mainly reflects the number of studies per modality. Recent examples include CNN-based tumor classification on MRI [64,65], DenseNet-based and hybrid DL models for breast cancer [66,67], ensemble segmentation of chest radiographs [68], ultrasound applications in breast, thyroid, fetal and colorectal imaging [69–73] and fundus-based detection of retinal and ocular diseases [74–76]. In a direct comparison, a CNN model and pediatric radiologists did not differ significantly in the ultrasound diagnosis of biliary atresia [77]. The present data do not allow modality-specific conclusions; the quality of the dataset, annotation standards and validation rigor are likely to be more important determinants of reported performance than the modality itself [78].

Clinical Implementation, Ethical and Regulatory Considerations

High discrimination in selected datasets does not establish clinical effectiveness, cost-effectiveness or readiness for routine implementation; these require prospective evaluation of patient outcomes, workflow effects and costs. In current practice, AI is used mainly as a supportive tool that reduces diagnostic workload and prioritizes high-risk cases [5,79]. Clinicians' adoption decisions depend on transparency and on the evidence available for a given application [80], and over-reliance on AI may affect the training of young pathologists [26]. Ethical and legal issues include algorithmic bias, data privacy and security, patient consent and professional liability [81], and regulatory perspectives, including those of the United States Food and Drug Administration and the European Medicines Agency, emphasize independent evaluation and transparency in data collection and use [82,83]. Prospective validation studies [84], integration with picture archiving and communication systems and radiology information systems, user training and continuous monitoring of performance are prerequisites for safe clinical use [79].

Strengths and Limitations

The strengths of this review include a search of five databases, independent data extraction and risk-of-bias assessment by two reviewers, an explicit operational definition of external validation, and stratification of the reported AUC values by validation strategy.

The review has several limitations. First, the search did not use controlled vocabulary; the PubMed publication-type and free-full-text limits, the restriction of the Web of Science search to six journals and the omission of pathology terms from the Embase and Web of Science strings reduced

sensitivity, so relevant studies were probably missed. Second, the restriction to open-access English-language articles may have introduced availability, publication and language bias, and the direction of this bias is uncertain. Third, the review was not prospectively registered, which increases the risk of post hoc decisions. Fourth, inclusion was based on reported AUC, the AUC categories were defined by the authors, and each study contributed one model; if the best-performing model was preferentially reported or extracted, the AUC distribution will be optimistic. Fifth, most studies were retrospective, risk of bias was frequently unclear for patient selection and for flow and timing, the low-risk judgments for the reference standard may be optimistic given the heterogeneity of reference standards, and applicability was assessed only for patient selection. Sixth, only seven studies were eligible for the meta-analysis; they addressed clinically unrelated conditions and used different units of analysis, including image-level data from patients who contributed several images, one external set came from the development institution and one was a temporal validation set, and in two studies the threshold was selected in the validation data, so the summary estimates are exploratory; the 2×2 data were re-extracted during revision, which corrected errors in the submitted analysis. Seventh, 11 records were based primarily on non-image data or had no recorded modality. Eighth, an AI-based screening tool contributed to the exclusion of 229 records at title and abstract screening, which may have led to the loss of eligible studies. Finally, studies published before 2021 were outside the date range.

CONCLUSIONS

Most recent studies of AI models for diagnosis in medical imaging and pathology report high AUC values, but this evidence is dominated by retrospective studies with unclear risk of bias. Reported AUC values were lower in studies with external validation only and in the few prospective RCTs than in studies with internal validation. Only seven of 170 studies provided 2×2 data from an external validation set; in these studies of clinically unrelated conditions, the exploratory summary sensitivity was 92.1% and specificity 90.1%, with a wide prediction region for specificity. High reported performance does not establish clinical effectiveness, generalizability, cost-effectiveness or readiness for routine implementation. Future studies should report complete 2×2 data at prespecified thresholds, evaluate models on external and prospectively collected data from consecutive patients, compare AI with clinicians on identical cohorts and follow AI-specific reporting guidance.

Supplementary Materials

The following supporting information is provided. Supplementary File 1: Table S1: Search strategies, limits and numbers of records retrieved; Table S2: Characteristics of the 170 included studies; Table S3: Reported AUC by AI architecture; Table S4: QUADAS-2 risk of bias and applicability for each included study; Figure S1: Summary of QUADAS-2 risk of bias

judgments; Figure S2: QUADAS-2 traffic-light plot; Table S5: 2x2 data of the studies considered for the meta-analysis with their sources in the primary publications; Table S6: Imaging modality by reported AUC category and AI architecture; Table S7: Imaging modality by target condition or task and reported AUC category; Table S8: Sensitivity analyses of the meta-analysis. Supplementary File 2: PRISMA-DTA checklist. Supplementary File 3: Study-level data (Microsoft Excel workbook). Supplementary File 4: R code for the meta-analysis.

Author Contributions

Conceptualization, D.M.; methodology, D.M. and V.Y.; software, S.M.; validation, V.Y. and O.A.; formal analysis, D.M. and S.M.; investigation, D.M., H.F., S.M., V.Y. and O.A.; data curation, H.F. and S.M.; writing – original draft preparation, D.M.; writing – review and editing, D.M., H.F., S.M., V.Y. and O.A.; visualization, S.M.; supervision, D.M. and O.A.; project administration, D.M. All authors have read and approved the final version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This systematic review analyzed data from previously published studies and did not involve human participants or identifiable personal information.

Informed Consent Statement

Not applicable, because the review analyzed only data from previously published studies.

Data Availability Statement

The study-level data extracted for this review, the QUADAS-2 judgments and the 2x2 data used in the meta-analysis are provided in Table 4 and Supplementary Files 1 and 3, and the analysis code is provided as Supplementary File 4.

Acknowledgments

The authors thank the investigators of the included studies whose published data made this review possible.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

- [1] Kwatra C.V. *et al.* "Harnessing ensemble deep learning models for precise detection of gynaecological cancers." *Clinical Epidemiology and Global Health*, vol. 32, 2025, pp. 101956. <https://doi.org/10.1016/j.cegh.2025.101956>
- [2] Yuan X. *et al.* "A novel early diagnostic framework for chronic diseases with class imbalance." *Scientific Reports*, vol. 12, no. 1, 2022, pp. 8614. <https://doi.org/10.1038/s41598-022-12574-x>
- [3] Bruno M.A. *et al.* "Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction." *RadioGraphics*, vol. 35, no. 6, 2015, pp. 1668-1676. <https://doi.org/10.1148/rg.2015150023>
- [4] Pesapane F. *et al.* "Errors in radiology: A standard review." *Journal of Clinical Medicine*, vol. 13, no. 15, 2024, pp. 4306. <https://doi.org/10.3390/jcm13154306>
- [5] Ahmad Z. *et al.* "Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: Present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review." *Diagnostic Pathology*, vol. 16, no. 1, 2021, pp. 24. <https://doi.org/10.1186/s13000-021-01085-4>
- [6] Oza P. *et al.* "Deep convolutional neural networks for computer-aided breast cancer diagnostic: A survey." *Neural Computing and Applications*, vol. 34, no. 3, 2022, pp. 1815-1836. <https://doi.org/10.1007/s00521-021-06804-y>
- [7] Li Z. *et al.* "A survey of convolutional neural networks: Analysis, applications, and prospects." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, 2022, pp. 6999-7019. <https://doi.org/10.1109/TNNLS.2021.3084827>
- [8] Di J. *et al.* "Utility of artificial intelligence in a binary classification of soft tissue tumors." *Journal of Pathology Informatics*, vol. 15, 2024, pp. 100368. <https://doi.org/10.1016/j.jpi.2024.100368>
- [9] Beyer D. *et al.* "Automating colon polyp classification in digital pathology by evaluation of a 'machine learning as a service' AI model: Algorithm development and validation study." *JMIR Formative Research*, vol. 9, 2025, pp. e67457. <https://doi.org/10.2196/67457>
- [10] Koelzer V.H. *et al.* "Precision immunoprofiling by image analysis and artificial intelligence." *Virchows Archiv*, vol. 474, no. 4, 2019, pp. 511-522. <https://doi.org/10.1007/s00428-018-2485-z>
- [11] Doi K. "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential." *Computerized Medical Imaging and Graphics*, vol. 31, nos. 4-5, 2007, pp. 198-211. <https://doi.org/10.1016/j.compmedimag.2007.02.002>
- [12] Albano D. *et al.* "Artificial intelligence for radiographic imaging detection of caries lesions: A systematic review." *BMC Oral Health*, vol. 24, no. 1, 2024, pp. 274. <https://doi.org/10.1186/s12903-024-04046-7>
- [13] Abut S. *et al.* "Paradigm shift from artificial neural networks (ANNs) to deep convolutional neural networks (DCNNs) in the field of medical image processing." *Expert Systems with Applications*, vol. 244, 2024, pp. 122983. <https://doi.org/10.1016/j.eswa.2023.122983>
- [14] Han K. *et al.* "A survey on vision transformer." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, 2023, pp. 87-110. <https://doi.org/10.1109/TPAMI.2022.3152247>
- [15] He K. *et al.* "Transformers in medical image analysis." *Intelligent Medicine*, vol. 3, no. 1, 2023, pp. 59-78. <https://doi.org/10.1016/j.imed.2022.07.002>
- [16] Pal A. *et al.* "ViT-DCNN: Vision transformer with deformable CNN model for lung and colon cancer detection." *Cancers*, vol. 17, no. 18, 2025, pp. 3005. <https://doi.org/10.3390/cancers17183005>
- [17] Giri L. *et al.* "Diagnostic accuracy of an offline CNN framework utilizing multi-view chest X-rays for screening 14 co-occurring communicable and non-communicable diseases." *Diagnostics*, vol. 16, no. 1, 2026, p. 66. <https://doi.org/10.3390/diagnostics16010066>

- [18] Shahin A. "Fine-tuned ResNet34 for efficient brain tumor classification." *Scientific Reports*, vol. 15, no. 1, 2025, pp. 36910. <https://doi.org/10.1038/s41598-025-20872-3>
- [19] Mello-Thoms C. and Mello C.A.B. "Clinical applications of artificial intelligence in radiology." *British Journal of Radiology*, vol. 96, no. 1150, 2023, pp. 20221031. <https://doi.org/10.1259/bjr.20221031>
- [20] Allalume P. et al. "Artificial intelligence-based opportunities in liver pathology: A systematic review." *Diagnostics*, vol. 13, no. 10, 2023, pp. 1799. <https://doi.org/10.3390/diagnostics13101799>
- [21] Clarke E.L. et al. "Image analysis of cutaneous melanoma histology: A systematic review and meta-analysis." *Scientific Reports*, vol. 13, no. 1, 2023, pp. 4774. <https://doi.org/10.1038/s41598-023-31526-7>
- [22] Girolami I. et al. "Artificial intelligence applications for pre-implantation kidney biopsy pathology practice: A systematic review." *Journal of Nephrology*, vol. 35, no. 7, 2022, pp. 1801-1808. <https://doi.org/10.1007/s40620-022-01327-8>
- [23] McGenity C. et al. "Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy." *npj Digital Medicine*, vol. 7, no. 1, 2024, p. 114. <https://doi.org/10.1038/s41746-024-01106-8>
- [24] Takita H. et al. "A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians." *npj Digital Medicine*, vol. 8, no. 1, 2025, pp. 175. <https://doi.org/10.1038/s41746-025-01543-z>
- [25] Kelly B.S. et al. "Radiology artificial intelligence: A systematic review and evaluation of methods (RAISE)." *European Radiology*, vol. 32, no. 11, 2022, pp. 7998-8007. <https://doi.org/10.1007/s00330-022-08784-6>
- [26] Bellahsen-Harrar Y. et al. "Exploring the risks of over-reliance on AI in diagnostic pathology: What lessons can be learned to support the training of young pathologists?" *PLOS ONE*, vol. 20, no. 8, 2025, pp. e0323270. <https://doi.org/10.1371/journal.pone.0323270>
- [27] Cai Y. et al. "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review." *BMC Medicine*, vol. 22, no. 1, 2024, pp. 56. <https://doi.org/10.1186/s12916-024-03273-7>
- [28] Mäenpää S.M. and Korja M. "Diagnostic test accuracy of externally validated convolutional neural network (CNN) artificial intelligence (AI) models for emergency head CT scans: A systematic review." *International Journal of Medical Informatics*, vol. 189, 2024, pp. 105523. <https://doi.org/10.1016/j.ijmedinf.2024.105523>
- [29] McInnes M.D.F. et al. "Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement." *JAMA*, vol. 319, no. 4, 2018, pp. 388-396. <https://doi.org/10.1001/jama.2017.19163>
- [30] Çorbacıoğlu Ş.K. and Aksel G. "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value." *Turkish Journal of Emergency Medicine*, vol. 23, no. 4, 2023, pp. 195-198. https://doi.org/10.4103/tjem.tjem_182_23
- [31] Whiting P.F. et al. "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies." *Annals of Internal Medicine*, vol. 155, no. 8, 2011, pp. 529-536. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>
- [32] Landis J.R. and Koch G.G. "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33, no. 1, 1977, pp. 159-174. <https://doi.org/10.2307/2529310>
- [33] Reitsma J.B. et al. "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews." *Journal of Clinical Epidemiology*, vol. 58, no. 10, 2005, pp. 982-990. <https://doi.org/10.1016/j.jclinepi.2005.02.022>
- [34] Harbord R.M. et al. "A unification of models for meta-analysis of diagnostic accuracy studies." *Biostatistics*, vol. 8, no. 2, 2007, pp. 239-251. <https://doi.org/10.1093/biostatistics/kxl004>
- [35] Deeks J.J. et al. "The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed." *Journal of Clinical Epidemiology*, vol. 58, no. 9, 2005, pp. 882-893. <https://doi.org/10.1016/j.jclinepi.2005.01.016>
- [36] Schilcher J. et al. "Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures." *Computers in Biology and Medicine*, vol. 168, 2024, pp. 107704. <https://doi.org/10.1016/j.compbiomed.2023.107704>
- [37] Wang W. et al. "Study on the classification of benign and malignant breast lesions using a multi-sequence breast MRI fusion radiomics and deep learning model." *European Journal of Radiology Open*, vol. 13, 2024, pp. 100607. <https://doi.org/10.1016/j.ejro.2024.100607>
- [38] Eliwa E.H.I. and Abd El-Hafeez T. "A novel YOLOv11-based framework with dynamic cross-scale context aggregation for high-performance blood cell classification." *Egyptian Informatics Journal*, vol. 32, 2025, pp. 100851. <https://doi.org/10.1016/j.eij.2025.100851>
- [39] Shi Y. et al. "Artificial intelligence-assisted detection of nasopharyngeal carcinoma on endoscopic images: A national, multicentre, model development and validation study." *The Lancet Digital Health*, vol. 7, no. 6, 2025, pp. 100869. <https://doi.org/10.1016/j.landig.2025.03.001>
- [40] Yang Y. et al. "Machine learning outperforms deep learning in adhesive capsulitis diagnosis: A clinical-radiomics model bridging PD-T2 MRI and multimodal data fusion." *European Journal of Radiology*, vol. 193, 2025, pp. 112470. <https://doi.org/10.1016/j.ejrad.2025.112470>
- [41] Valesano A.L. et al. "Weakly supervised deep learning-based detection of serous tubal intraepithelial carcinoma in fallopian tubes." *Journal of Pathology Informatics*, vol. 19, 2025, pp. 100522. <https://doi.org/10.1016/j.jpi.2025.100522>
- [42] Liu H. et al. "Deep learning based on automated breast volume scanner images for the diagnosis of breast lesions: A multicenter diagnostic study." *International Journal of Medical Sciences*, vol. 22, no. 15, 2025, pp. 3924-3937. <https://doi.org/10.7150/ijms.118430>
- [43] Lee H. et al. "Comparative diagnostic performance of a multimodal large language model versus a dedicated electrocardiogram AI in detecting myocardial infarction from electrocardiogram images: Comparative study." *JMIR AI*, vol. 4, 2025, p. e75910. <https://doi.org/10.2196/75910>
- [44] Xu N. et al. "Deep learning algorithms for identifying developmental dysplasia of the hip based on sonographic images: A retrospective, prospective, multicenter study in China." *eClinicalMedicine*, vol. 89, 2025, pp. 103552. <https://doi.org/10.1016/j.eclinm.2025.103552>
- [45] Wang J. et al. "Deep learning models for thyroid nodules diagnosis of fine-needle aspiration biopsy: A retrospective, prospective, multicentre study in China." *The Lancet Digital Health*, vol. 6, no. 7, 2024, pp. e458-e469. [https://doi.org/10.1016/S2589-7500\(24\)00085-2](https://doi.org/10.1016/S2589-7500(24)00085-2)

- [46] Liu X. et al. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis." *The Lancet Digital Health*, vol. 1, no. 6, 2019, pp. e271-e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)
- [47] Yu A.C. et al. "External validation of deep learning algorithms for radiologic diagnosis: A systematic review." *Radiology: Artificial Intelligence*, vol. 4, no. 3, 2022, pp. e210064. <https://doi.org/10.1148/ryai.210064>
- [48] Chen X. et al. "An enhanced AlexNet-based model for femoral bone tumor classification and diagnosis using magnetic resonance imaging." *Journal of Bone Oncology*, vol. 48, 2024, pp. 100626. <https://doi.org/10.1016/j.jbo.2024.100626>
- [49] Chen J. et al. "EE-Explorer: A multimodal artificial intelligence system for eye emergency triage and primary diagnosis." *American Journal of Ophthalmology*, vol. 252, 2023, pp. 253-264. <https://doi.org/10.1016/j.ajo.2023.04.007>
- [50] Hendrawan J. et al. "Hybrid deep learning model for acute lymphoblastic leukemia (ALL) detection using pre-trained ResNet-50 and vision transformer architecture." *Procedia Computer Science*, vol. 269, 2025, pp. 933-942. <https://doi.org/10.1016/j.procs.2025.09.036>
- [51] Ansari M.M. et al. "SVMVGGNet-16: A novel machine and deep learning based approaches for lung cancer detection using combined SVM and VGGNet-16." *Current Medical Imaging*, vol. 21, 2025, pp. e15734056348824. <https://doi.org/10.2174/0115734056348824241224100809>
- [52] Wang Z. et al. "Spinal magnetic resonance image segmentation based on U-net." *Journal of Radiation Research and Applied Sciences*, vol. 16, no. 3, 2023, pp. 100627. <https://doi.org/10.1016/j.jrras.2023.100627>
- [53] Cai D. et al. "Automatic segmentation of knee CT images of tibial plateau fractures based on three-dimensional U-Net: Assisting junior physicians with Schatzker classification." *European Journal of Radiology*, vol. 178, 2024, pp. 111605. <https://doi.org/10.1016/j.ejrad.2024.111605>
- [54] Liu J. et al. "YOLOv12 algorithm-aided detection and classification of lateral malleolar avulsion fracture and subfibular ossicle based on CT images: Multicenter study." *JMIR Medical Informatics*, vol. 13, 2025, pp. e79064. <https://doi.org/10.2196/79064>
- [55] Dehbozorgi P. et al. "A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities." *Computers in Biology and Medicine*, vol. 187, 2025, pp. 109768. <https://doi.org/10.1016/j.compbiomed.2025.109768>
- [56] Yan F. et al. "A multi-task deep learning pipeline integrating vessel segmentation and radiomics for multiclass retinal disease classification." *Photodiagnosis and Photodynamic Therapy*, vol. 56, 2025, pp. 105209. <https://doi.org/10.1016/j.pdpdt.2025.105209>
- [57] Choudhury A.R. et al. "Deep learning based automated lung cancer detection from CT scan leveraging transfer learning." *Procedia Computer Science*, vol. 258, 2025, pp. 2748-2759. <https://doi.org/10.1016/j.procs.2025.04.535>
- [58] Shatnawi M.Q. et al. "Deep learning-based approach to diagnose lung cancer using CT-scan images." *Intelligence-Based Medicine*, vol. 11, 2025, pp. 100188. <https://doi.org/10.1016/j.ibmed.2024.100188>
- [59] Ayogu I.I. et al. "Investigation of ensembles of deep learning models for improved chronic kidney diseases detection in CT scan images." *Franklin Open*, vol. 11, 2025, pp. 100298. <https://doi.org/10.1016/j.fraope.2025.100298>
- [60] Chen Y. et al. "Deep learning-based computer-aided diagnostic system for lumbar degenerative diseases classification using MRI." *Biomedical Signal Processing and Control*, vol. 109, 2025, pp. 108002. <https://doi.org/10.1016/j.bspc.2025.108002>
- [61] Sánchez-Moreno L. et al. "Ensemble-based convolutional neural networks for brain tumor classification in MRI: Enhancing accuracy and interpretability using explainable AI." *Computers in Biology and Medicine*, vol. 195, 2025, pp. 110555. <https://doi.org/10.1016/j.compbiomed.2025.110555>
- [62] Altal O.F. et al. "Hybrid attention-enhanced MobileNetV2 with particle swarm optimization for endometrial cancer classification in CT images." *Informatics in Medicine Unlocked*, vol. 57, 2025, pp. 101662. <https://doi.org/10.1016/j.imu.2025.101662>
- [63] Bo Z. et al. "Application of artificial intelligence radiomics in the diagnosis, treatment, and prognosis of hepatocellular carcinoma." *Computers in Biology and Medicine*, vol. 173, 2024, pp. 108337. <https://doi.org/10.1016/j.compbiomed.2024.108337>
- [64] Prabhas K.S. et al. "A deep learning framework for brain tumor detection using CNNs and transfer learning on MRI scans." *Systems and Soft Computing*, vol. 7, 2025, pp. 200389. <https://doi.org/10.1016/j.sasc.2025.200389>
- [65] Ramakrishna M.T. et al. "Leveraging EfficientNetB3 in a deep learning framework for high-accuracy MRI tumor classification." *Computers, Materials & Continua*, vol. 81, no. 1, 2024, pp. 867-883. <https://doi.org/10.32604/cmc.2024.053563>
- [66] Talukder M.A. "An improved XAI-based DenseNet model for breast cancer detection using reconstruction and fine-tuning." *Results in Engineering*, vol. 26, 2025, pp. 104802. <https://doi.org/10.1016/j.rineng.2025.104802>
- [67] Chitta S. et al. "Hybrid deep learning model for enhanced breast cancer diagnosis using histopathological images." *Procedia Computer Science*, vol. 260, 2025, pp. 245-251. <https://doi.org/10.1016/j.procs.2025.03.199>
- [68] Junia R.C. and K. S. "Deep learning-based automatic segmentation of COVID-19 in chest X-ray images using ensemble neural net sentinel algorithm." *Measurement: Sensors*, vol. 33, 2024, pp. 101117. <https://doi.org/10.1016/j.measen.2024.101117>
- [69] Li H. and Cheng T. "Multicenter and multimodal ultrasound-based radiomics and transformer-driven end-to-end deep learning for breast cancer molecular subtype classification." *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 3, 2025, pp. 101656. <https://doi.org/10.1016/j.jrras.2025.101656>
- [70] Sharifi Y. et al. "Using deep learning for thyroid nodule risk stratification from ultrasound images." *WFUMB Ultrasound Open*, vol. 3, no. 1, 2025, pp. 100082. <https://doi.org/10.1016/j.wfumbo.2025.100082>
- [71] Harikumar A. et al. "Explainable AI in deep learning based classification of fetal ultrasound image planes." *Procedia Computer Science*, vol. 233, 2024, pp. 1023-1033. <https://doi.org/10.1016/j.procs.2024.03.291>
- [72] Podda A.S. et al. "A deep learning strategy for the 3D segmentation of colorectal tumors from ultrasound imaging." *Image and Vision Computing*, vol. 162, 2025, pp. 105668. <https://doi.org/10.1016/j.imavis.2025.105668>
- [73] Yilmaz M.T. et al. "A comparative study of advanced deep learning architectures for breast cancer classification on ultrasound and histological images." *Results in Engineering*, vol. 28, 2025, pp. 107600. <https://doi.org/10.1016/j.rineng.2025.107600>

- [74] Kılıç Ş. "HybridVisionNet: An advanced hybrid deep learning framework for automated multi-class ocular disease diagnosis using fundus imaging." *Ain Shams Engineering Journal*, vol. 16, 2025, pp. 103594. <https://doi.org/10.1016/j.asej.2025.103594>
- [75] Inamullah *et al.* "Deciphering the impact of diversity in CNN-based ensembles on overcoming data imbalance and scarcity in medical datasets: A case study on diabetic retinopathy." *Informatics in Medicine Unlocked*, vol. 49, 2024, pp. 101557. <https://doi.org/10.1016/j.imu.2024.101557>
- [76] Anvesh K. *et al.* "A novel approach deep learning framework for automatic detection of diseases in retinal fundus images." *Computer Modeling in Engineering & Sciences*, vol. 143, no. 2, 2025, pp. 1485-1517. <https://doi.org/10.32604/cmescs.2025.063239>
- [77] Duan X. *et al.* "Is the diagnostic model based on convolutional neural network superior to pediatric radiologists in the ultrasonic diagnosis of biliary atresia?" *Frontiers in Medicine*, vol. 10, 2024, pp. 1308338. <https://doi.org/10.3389/fmed.2023.1308338>
- [78] Yi S. and Chen Z. "MIDC: Medical image dataset cleaning framework based on deep learning." *Heliyon*, vol. 10, no. 19, 2024, pp. e38910. <https://doi.org/10.1016/j.heliyon.2024.e38910>
- [79] Jeong J. *et al.* "Reducing the workload of medical diagnosis through artificial intelligence: A narrative review." *Medicine*, vol. 104, no. 6, 2025, pp. e41470. <https://doi.org/10.1097/MD.00000000000041470>
- [80] Hennrich J. *et al.* "Enabling physicians to make an informed adoption decision on artificial intelligence applications in medical imaging diagnostics: Qualitative study." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e63668. <https://doi.org/10.2196/63668>
- [81] Naili Y.T. *et al.* "Regulatory challenges in AI-based diagnostics: Legal implications of AI use in medical diagnostics." *BIO Web of Conferences*, vol. 152, 2025, pp. 01034. <https://doi.org/10.1051/bioconf/202515201034>
- [82] Jhansi G. *et al.* "Regulatory considerations for AI/ML medical devices: Bridging the gap between innovation and patient safety." *International Journal of Pharmaceutical Sciences Review and Research*, vol. 84, no. 12, 2024, pp. 1-10. <https://doi.org/10.47583/ijpsrr.2024.v84i12.001>
- [83] Fu L. *et al.* "The applications and advances of artificial intelligence in drug regulation: A global perspective." *Acta Pharmaceutica Sinica B*, vol. 15, no. 1, 2025, pp. 1-14. <https://doi.org/10.1016/j.apsb.2024.11.006>
- [84] Miró Catalina Q. *et al.* "Developing an artificial intelligence model for reading chest X-rays: Protocol for a prospective validation study." *JMIR Research Protocols*, vol. 11, no. 11, 2022, pp. e39536. <https://doi.org/10.2196/39536>