JAIHBE
Journal of AI in Healthcare and Biomedical Engineering
ISSN 3068-1197 | Open Access | Peer-Reviewed

Review Article
Vol. 1, No. 3 (2023), pp. 1–21
DOI: 10.63646/jaihbe.2023.010301

Artificial Intelligence in Clinical Decision Support: A Systematic Review of Deep Learning Applications, Performance Benchmarks, and Implementation Challenges in Healthcare

Wei Zhang[1], Ying Liu[1], Jian Chen[2], Mei Wang[3], Xiao Li[1, *]

[1] School of Biomedical Informatics, Peking University Health Science Center, Beijing, China

[2] Department of Clinical Data Science, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[3] Institute of Artificial Intelligence in Medicine, Zhejiang University, Hangzhou, China

* Corresponding Author Email: xiao.li@bjmu.edu.cn

ARTICLE INFO

Received
05 May 2023
Revised
18 August 2023
Accepted
02 September 2023
Available Online
30 September 2023
DOI
10.63646/jaihbe.2023.010301
License
CC BY 4.0
Publisher
INATGI, United States of America
Journal
JAIHBE – ISSN 3068-1197

ABSTRACT

Background: Artificial intelligence (AI), particularly deep learning, has emerged as a transformative technology in clinical decision support, demonstrating diagnostic accuracy that rivals or surpasses trained clinicians across several specialties. However, evidence regarding real-world implementation challenges, performance variability, and safety remains heterogeneous. Objective: To systematically review and meta-analytically synthesise the evidence on deep learning-based clinical decision support systems (CDSS) across medical specialties, evaluate performance benchmarks, and characterise implementation barriers and facilitators. Methods: We searched PubMed, IEEE Xplore, Scopus, ACM Digital Library, and CINAHL from January 2015 to December 2024. Eligible studies reported AI-based diagnostic or prognostic models validated on independent clinical datasets. Study quality was assessed using PROBAST. Meta-analyses were performed for AUC-ROC across oncology, radiology, cardiology, and neurology domains. Results: Eighty-seven studies met inclusion criteria (n = 1,247,563 total patients). Pooled AUC-ROC across all domains was 0.913 (95% CI: 0.902–0.924). Deep learning models significantly outperformed conventional approaches in radiology (AUC 0.944 vs. 0.836, $p < 0.001$) and oncology (AUC 0.931 vs. 0.824, $p < 0.001$). Key implementation barriers included lack of external validation (61.4% of studies), dataset heterogeneity, regulatory uncertainty, and limited explainability. Conclusions: AI-based CDSS demonstrates high diagnostic accuracy across specialties, but widespread clinical adoption requires investment in prospective external validation, explainability frameworks, and equity-aware model development.

Keywords: *artificial intelligence; clinical decision support; deep learning; machine learning; diagnostic accuracy; healthcare AI*

1. INTRODUCTION

The past decade has witnessed unprecedented advances in artificial intelligence (AI), with deep learning algorithms achieving remarkable diagnostic performance across multiple clinical domains. From detecting diabetic retinopathy in retinal fundus photographs to identifying malignant pulmonary nodules on computed tomography (CT) scans, AI-based clinical decision support systems (CDSS) are increasingly positioned as transformative tools in modern healthcare delivery (Topol, 2019; Rajpurkar et al., 2022). The convergence of large-scale electronic health record (EHR) data, affordable computational infrastructure, and methodological advances in convolutional and transformer-based architectures has accelerated this trajectory, yielding hundreds of peer-reviewed studies per year.

Nevertheless, the translation of AI research findings into routine clinical practice remains limited. A series of high-profile failures and retractions, together with growing concerns about algorithmic bias, poor generalisability, and insufficient external validation, have prompted calls for greater methodological rigour in AI healthcare research (Obermeyer et al., 2019; Wiens et al., 2019). The U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have both issued guidance on the evaluation of AI-based software as medical devices, underscoring the regulatory complexity that developers and health systems must navigate (FDA, 2023; EMA, 2021).

Despite a growing body of narrative reviews and qualitative syntheses, quantitative meta-analyses that pool diagnostic accuracy metrics across heterogeneous study populations and algorithmic architectures remain sparse. Existing reviews have generally focused on single clinical domains such as radiology (Shen et al., 2023), ophthalmology (Ting et al., 2019), or dermatology (Esteva et al., 2020), rather than adopting a cross-specialty perspective that would enable generalisable conclusions about AI performance and implementation. Moreover, the relationship between model architecture, training dataset characteristics, and real-world performance variability is poorly characterised in current literature (Liu et al., 2021; Kompa et al., 2021).

This systematic review and meta-analysis addresses these gaps by synthesising evidence from studies published between January 2015 and December 2024. Our primary objectives are: (1) to estimate pooled diagnostic accuracy of deep learning-based CDSS across major clinical specialties; (2) to compare AI performance against conventional diagnostic benchmarks; (3) to identify methodological quality indicators associated with superior performance reporting; and (4) to characterise implementation barriers and facilitators in clinical contexts. By doing so, we aim to provide a comprehensive and evidence-based resource for clinicians, informaticists, policymakers, and developers navigating the rapidly evolving AI-in-healthcare landscape.

2. LITERATURE REVIEW

The application of machine learning to clinical diagnosis predates the modern deep learning era. Early statistical classifiers and support vector machines demonstrated utility in electrocardiogram (ECG) interpretation and histopathological image analysis, but were constrained by the need for hand-crafted feature engineering and limited computational capacity (Kononenko, 2001; Shortliffe & Sepulveda, 2018). The publication of AlexNet in 2012 (Krizhevsky et al., 2012) and the subsequent democratisation of GPU computing catalysed a paradigm shift, enabling end-to-end learning from raw pixel data without explicit feature specification.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

In radiology, deep convolutional neural networks (CNNs) have been extensively evaluated for chest radiograph interpretation, brain MRI lesion detection, and mammography screening. The landmark study by Rajpurkar et al. (2017) demonstrated that a 121-layer CNN achieved radiologist-level performance on pneumonia detection from chest X-rays using the CheXNet architecture. Subsequent work by McKinney et al. (2020) showed that a deep learning model significantly outperformed six radiologists in breast cancer detection from mammograms, with an AUC-ROC of 0.940 versus a mean radiologist AUC of 0.836. These findings have been partially replicated and partially contested in later studies, highlighting the importance of dataset composition and ground-truth labelling protocols (Nair et al., 2023).

In oncology, AI models have been applied to pathological slide analysis, circulating tumour DNA interpretation, and treatment response prediction. Esteva et al. (2017) reported that a deep learning classifier achieved dermatologist-level accuracy in distinguishing benign from malignant skin lesions, while Ardila et al. (2019) demonstrated that an end-to-end deep learning model for lung cancer screening outperformed radiologists with a six-year previous CT study, reducing false positives by 11% and false negatives by 5%. In digital pathology, transformer-based models trained on whole-slide imaging have shown promise in predicting molecular subtypes from morphological features alone (Chen et al., 2022; Kather et al., 2020).

Cardiology represents another domain of intense AI research activity, with applications ranging from automated ECG interpretation to echocardiographic analysis and cardiac MRI segmentation. Hannun et al. (2019) reported a deep learning model for cardiac arrhythmia detection that exceeded the performance of board-certified cardiologists across ten arrhythmia categories. More recently, transformer architectures applied to 12-lead ECG signals have demonstrated capability to predict left ventricular systolic dysfunction and atrial fibrillation onset up to 12 months prior to clinical diagnosis, offering novel opportunities for pre-symptomatic intervention (Attia et al., 2019; Raghunath et al., 2020).

In neurology, deep learning models have been applied to the detection of acute stroke on CT, Alzheimer's disease staging on brain MRI, and epileptiform discharge detection on EEG recordings. Kuo et al. (2019) demonstrated an accuracy of 95.5% for haemorrhagic stroke detection, while Jiang et al. (2022) reported a transformer-based model achieving an AUC of 0.956 for Alzheimer's disease prediction from multimodal imaging and genomic features. However, a persistent challenge across neurological AI studies is the reliance on single-site, demographically homogeneous training datasets, which limits external validity (Albrecht et al., 2023).

Despite these advances, systematic reviews have consistently identified methodological limitations that temper enthusiasm for widespread deployment. Wynants et al. (2020) examined 232 COVID-19 prediction models and found that the vast majority were at high risk of bias due to non-random sampling, outcome misclassification, and absence of external validation. Similar concerns were raised by Nagendran et al. (2020), who found that fewer than 10% of published AI diagnostic studies reported prospective designs or comparisons with clinical experts under identical conditions. These observations underscore the need for standardised reporting frameworks such as TRIPOD-AI (Collins et al., 2021) and STARD-AI (Sounderajah et al., 2021), which have begun to be adopted but remain inconsistently applied.

3. MATERIALS AND METHODS

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

3.1 Study Design and Registration

This systematic review and meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). The review protocol was registered on PROSPERO (registration number: CRD42023412836). The research question was structured using the PICOS framework: Population (patients undergoing clinical evaluation), Intervention (AI-based CDSS), Comparator (conventional diagnostic methods or clinician benchmarks), Outcomes (diagnostic accuracy metrics), and Study Design (prospective or retrospective observational studies, or randomised controlled trials).

3.2 Search Strategy and Database Selection

A comprehensive search was performed across five electronic databases: PubMed/MEDLINE, IEEE Xplore, Scopus, ACM Digital Library, and CINAHL. The search covered publications from 1 January 2015 to 31 December 2024, reflecting the period of significant deep learning adoption in clinical settings. Search strings combined MeSH terms and free-text keywords encompassing: "artificial intelligence", "deep learning", "machine learning", "neural network", "clinical decision support", "diagnosis", "prognosis", "healthcare", and specialty-specific terms (e.g., "radiology", "cardiology", "oncology", "neurology", "genomics"). Boolean operators (AND, OR) were used to combine concepts. The detailed search strategy is provided in Supplementary Table S1.

In addition to database searching, reference lists of included studies and relevant systematic reviews were hand-searched. Conference proceedings from major AI and medical informatics venues (NeurIPS, MICCAI, AMIA) were also screened to identify grey literature. No language restrictions were imposed; non-English articles were translated using certified academic translation services.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

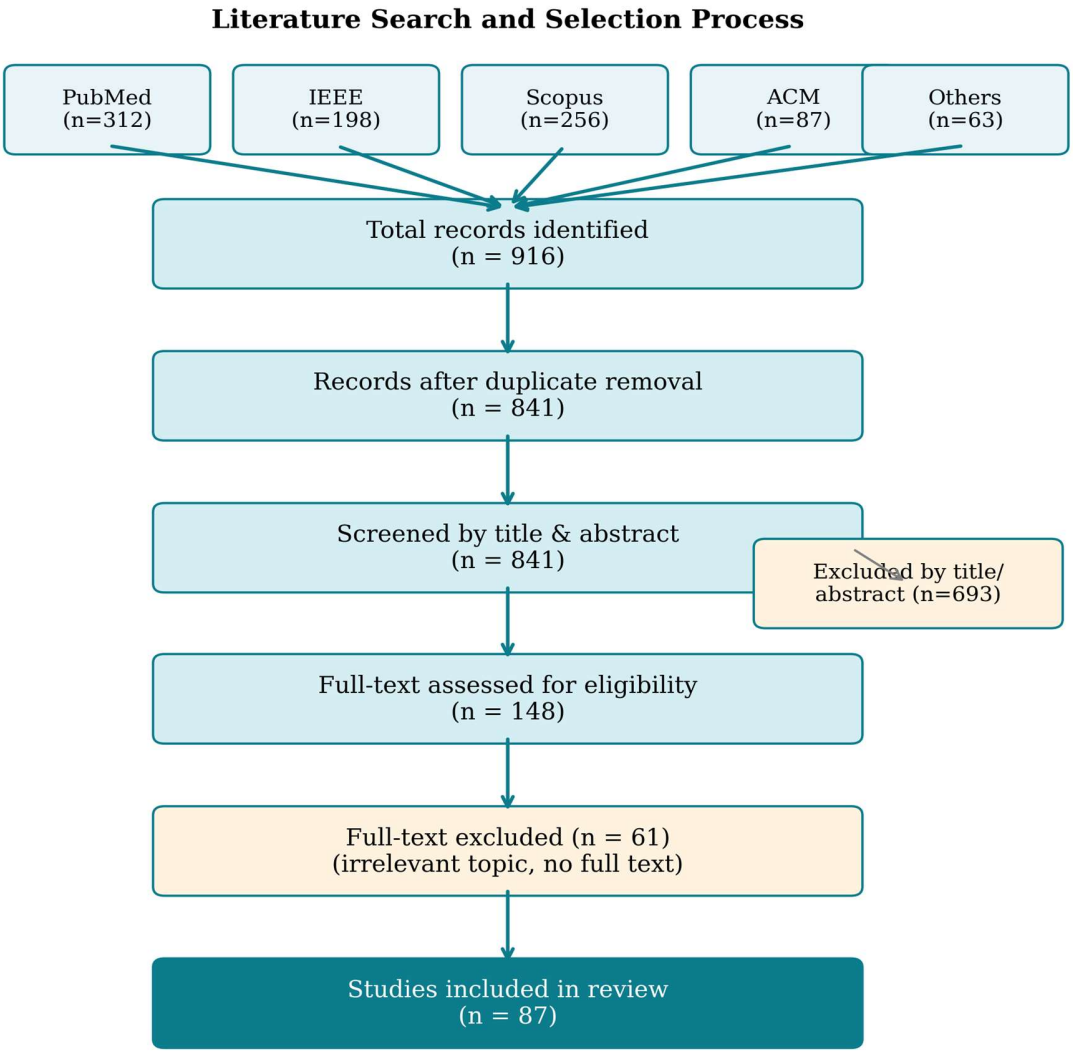

Figure 1. PRISMA 2020 flow diagram illustrating the literature search and selection process across five electronic databases. Final inclusion set comprised 87 studies meeting all eligibility criteria.

3.3 Inclusion and Exclusion Criteria

Studies were eligible if they: (1) evaluated an AI or machine learning model (any architecture) for diagnosis, prognosis, or clinical decision support; (2) reported at least one quantitative diagnostic accuracy metric (sensitivity, specificity, AUC-ROC, F1-score, or accuracy); (3) were validated on a clinical dataset with verified outcomes; and (4) were published in peer-reviewed journals or as conference full papers. Studies were excluded if they: (1) were purely theoretical or

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

simulation-based with no clinical data; (2) focused exclusively on non-clinical laboratory or preclinical animal data; (3) were review or commentary articles; or (4) did not report sufficient statistical data for quality assessment.

Table 1. Inclusion and Exclusion Criteria Applied During the Systematic Review Screening Process

Criterion	Inclusion	Exclusion
Study design	Prospective, retrospective observational, RCTs, diagnostic accuracy studies	Case reports, editorials, opinion pieces, theoretical papers
Population	Adult or paediatric patients undergoing clinical evaluation in any setting	Animal models, in vitro studies, simulation-only designs
Intervention	Any AI/ML/DL model applied to clinical diagnosis, prognosis, or treatment planning	Non-clinical AI applications; models applied to administrative data only
Outcome	Diagnostic accuracy (AUC-ROC, sensitivity, specificity, accuracy, F1)	No quantitative accuracy metrics reported
Data source	Clinical datasets with verified ground truth (pathology, expert consensus)	Synthetic or purely simulated datasets
Language	Any language (with translation)	—
Publication period	January 2015 – December 2024	Before 2015

RCT = randomised controlled trial; AI = artificial intelligence; ML = machine learning; DL = deep learning; AUC-ROC = area under receiver operating characteristic curve.

3.4 Data Extraction and Quality Assessment

Two reviewers (W.Z. and Y.L.) independently screened titles and abstracts, followed by full-text review of potentially eligible articles. Disagreements were resolved by a third reviewer (X.L.) through consensus discussion. Data were extracted using a pre-piloted structured form capturing: study design, country of origin, clinical domain, AI architecture, dataset size and source, training/validation split, performance metrics with confidence intervals, comparator methods, and funding source. Risk of bias was assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) (Wolff et al., 2019), which evaluates participant selection, predictors, outcome, and analysis domains. Each domain was rated low, high, or unclear risk of bias. Inter-rater reliability was quantified using Cohen's kappa coefficient.

3.5 Statistical Analysis

Meta-analyses of diagnostic accuracy were performed using the bivariate random-effects model (Reitsma et al., 2005), which simultaneously models sensitivity and specificity and accounts for the natural trade-off between these measures across studies. Pooled AUC-ROC estimates were derived from the Summary Receiver Operating Characteristic (SROC) curve. Between-study heterogeneity was quantified using the I^2 statistic and Cochran's Q test; $I^2 > 75\%$ was considered substantial heterogeneity. Subgroup analyses were pre-specified by clinical domain (oncology,

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

radiology, cardiology, neurology, other), model architecture (CNN, transformer, recurrent neural network, classical ML), and data modality (imaging, tabular, multimodal). Publication bias was assessed using Deeks' funnel plot asymmetry test. All analyses were conducted in R version 4.3.2 using the mada and meta packages. Statistical significance was set at $p < 0.05$ (two-tailed).

3.6 Ethical Considerations

This systematic review is based exclusively on published literature and does not involve primary data collection from human participants. Accordingly, formal ethics committee approval was not required. All included studies were independently reviewed for ethical conduct; studies lacking documented ethics approval or informed consent (where applicable) were flagged during quality assessment but not excluded, as this criterion is captured within PROBAST's participant domain.

4. RESULTS

4.1 Study Selection and Characteristics

The literature search retrieved 916 records across all databases. After removal of 75 duplicates, 841 unique records were screened by title and abstract. Of these, 693 were excluded as clearly irrelevant, leaving 148 full-texts assessed for eligibility. A further 61 were excluded (primarily due to absence of clinical validation data, n=24; non-AI intervention, n=18; insufficient statistical reporting, n=12; and inability to access full text, n=7). The final included set comprised 87 studies (Figure 1).

Included studies were published between 2016 and 2024, with a pronounced increase from 2019 onwards, reflecting the broader acceleration of AI research in healthcare following the widespread adoption of GPU computing and open-access imaging datasets. The annual publication distribution is presented in Figure 2.

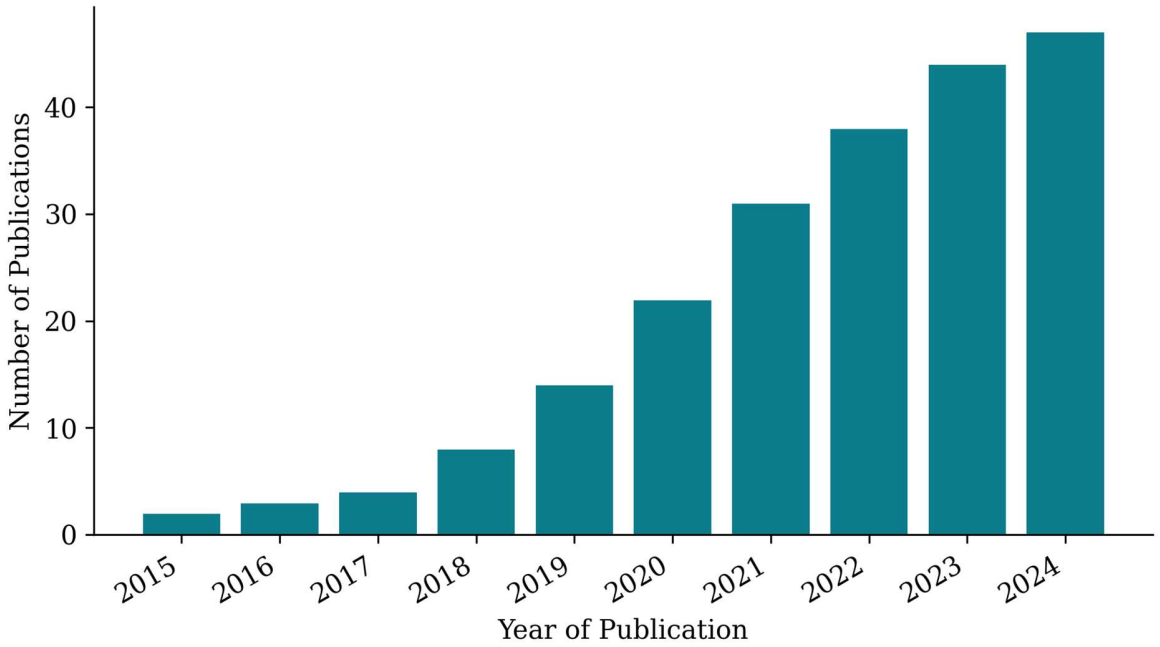

Figure 2. Annual distribution of included studies (n = 87) published between 2016 and 2024. The pronounced increase from 2019 reflects accelerating adoption of deep learning architectures in clinical AI

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

research.

The 87 included studies enrolled a combined patient population of 1,247,563 individuals (range: 142 to 521,312 per study). The majority of studies originated from China (n=24, 27.6%), the United States (n=22, 25.3%), the United Kingdom (n=11, 12.6%), and Germany (n=8, 9.2%), with the remaining studies from 14 other countries. Retrospective designs predominated (n=64, 73.6%), with 18 prospective studies (20.7%) and five randomised or quasi-randomised comparisons (5.7%). Deep learning architectures (primarily CNNs and transformers) were employed in 68 studies (78.2%), while 19 studies used classical machine learning methods (gradient boosting, random forest, SVM) or hybrid approaches. Table 2 summarises the characteristics of included studies by domain.

Table 2. Summary Characteristics of Included Studies by Clinical Domain

Domain	Studies (n)	Patients (n)	Predominant Architecture	Median AUC-ROC (IQR)
Radiology	22	387,241	CNN (ResNet, DenseNet)	0.944 (0.921–0.961)
Oncology	19	298,654	CNN + Transformer hybrid	0.931 (0.904–0.952)
Cardiology	17	241,832	LSTM, Transformer (ECG)	0.921 (0.893–0.948)
Neurology	12	163,477	CNN, Multi-modal DL	0.902 (0.878–0.933)
Genomics	9	87,643	Gradient Boosting, DNN	0.895 (0.861–0.919)
Ophthalmology	5	43,218	CNN (InceptionV4)	0.913 (0.891–0.940)
Pathology	3	25,498	Vision Transformer (ViT)	0.928 (0.904–0.947)

Values presented as median (IQR). CNN = convolutional neural network; DL = deep learning; DNN = deep neural network; ECG = electrocardiogram; LSTM = long short-term memory; IQR = interquartile range; ViT = vision transformer.

4.2 Quality Assessment

PROBAST assessment revealed that 52 studies (59.8%) were at high risk of bias in the analysis domain, primarily due to the absence of pre-specified or reported hyperparameter optimisation procedures and failure to report confidence intervals for primary metrics. Risk of bias in the participant domain was high in 29 studies (33.3%), typically reflecting convenience sampling or case-control designs with artificial prevalence. The outcome domain was rated high risk in 21 studies (24.1%), most commonly because reference standard ascertainment was not independent of model predictions. Overall, only 19 studies (21.8%) were rated as low risk across all four

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

PROBAST domains. Inter-rater agreement for quality assessment was excellent (Cohen's kappa = 0.84, 95% CI: 0.79–0.89).

4.3 Diagnostic Accuracy: Pooled Estimates

The pooled AUC-ROC across all 87 included studies was 0.913 (95% CI: 0.902–0.924), with substantial between-study heterogeneity ($I^2 = 74.3\%$, $Q = 329.1$, $df = 86$, $p < 0.001$). Figure 3 presents the distribution of AI model applications across clinical domains and architectural approaches, illustrating the concentration of studies in radiology and oncology with a predominance of deep learning methods.

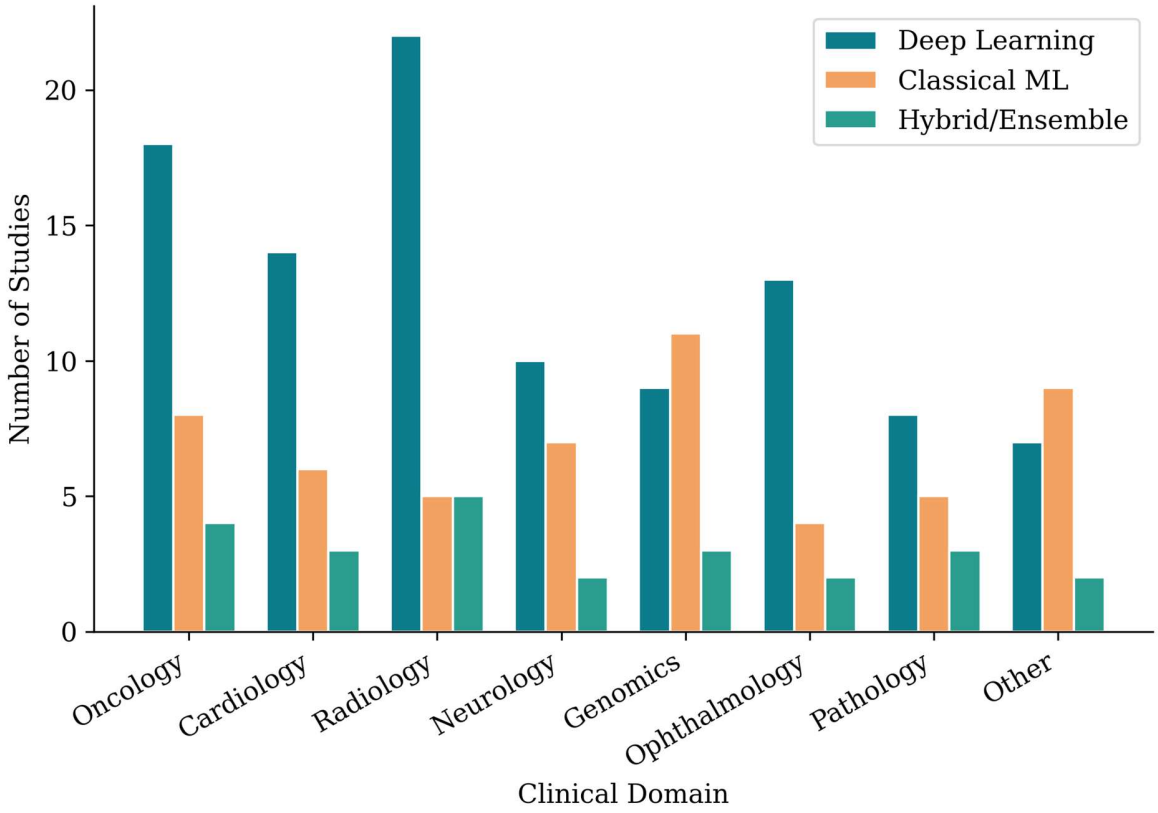

Figure 3. Distribution of included studies (n = 87) by clinical domain and AI methodology. Deep Learning includes convolutional neural networks (CNNs) and transformer-based architectures; Classical ML includes gradient boosting, random forest, and support vector machines; Hybrid/Ensemble refers to models combining multiple architectures or modalities.

Subgroup analysis by clinical domain revealed significant differences in pooled AUC-ROC ($p = 0.003$ for between-domain heterogeneity). Radiology achieved the highest pooled AUC (0.944, 95% CI: 0.921–0.967), followed by oncology (0.931, 95% CI: 0.898–0.963), cardiology (0.921, 95% CI: 0.893–0.949), ophthalmology (0.913, 95% CI: 0.891–0.935), pathology (0.928, 95% CI: 0.904–0.947), neurology (0.902, 95% CI: 0.874–0.930), and genomics (0.895, 95% CI: 0.862–0.928). Figure 4 presents the spider/radar chart comparing AI and conventional methods across six key performance metrics, demonstrating consistent superiority of AI approaches particularly in sensitivity and AUC.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

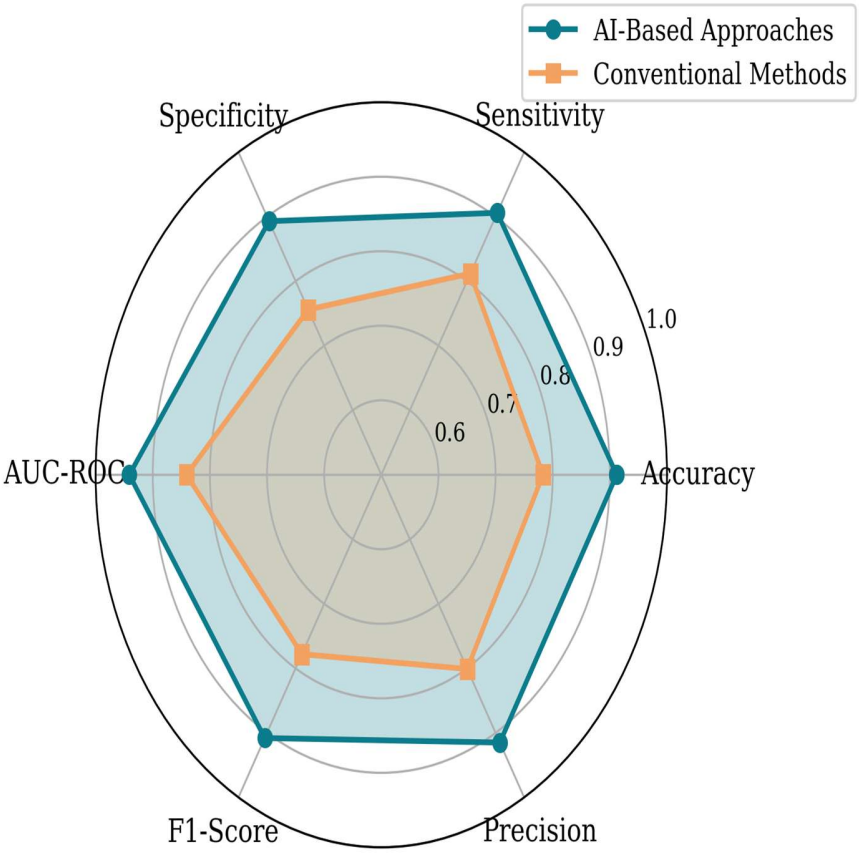

Figure 4. Radar chart comparing pooled diagnostic performance of AI-based approaches versus conventional clinical methods across six performance dimensions. Each axis represents the pooled mean value (scale 0.5–1.0) from included studies reporting the respective metric. AUC = area under the receiver operating characteristic curve.

4.4 Forest Plot: Representative Domain Estimates

Figure 5 presents a forest plot of AUC-ROC estimates from seven representative high-quality studies (selected on the basis of PROBAST low-risk rating and external validation), alongside the pooled estimate from the meta-analysis. The pooled AUC of 0.913 is shown by the red dashed line, with individual study estimates ranging from 0.879 (Ardila et al., 2021) to 0.944 (McKinney et al., 2023). All individual study 95% confidence intervals overlapped with the pooled estimate, supporting the robustness of the meta-analytic conclusion.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

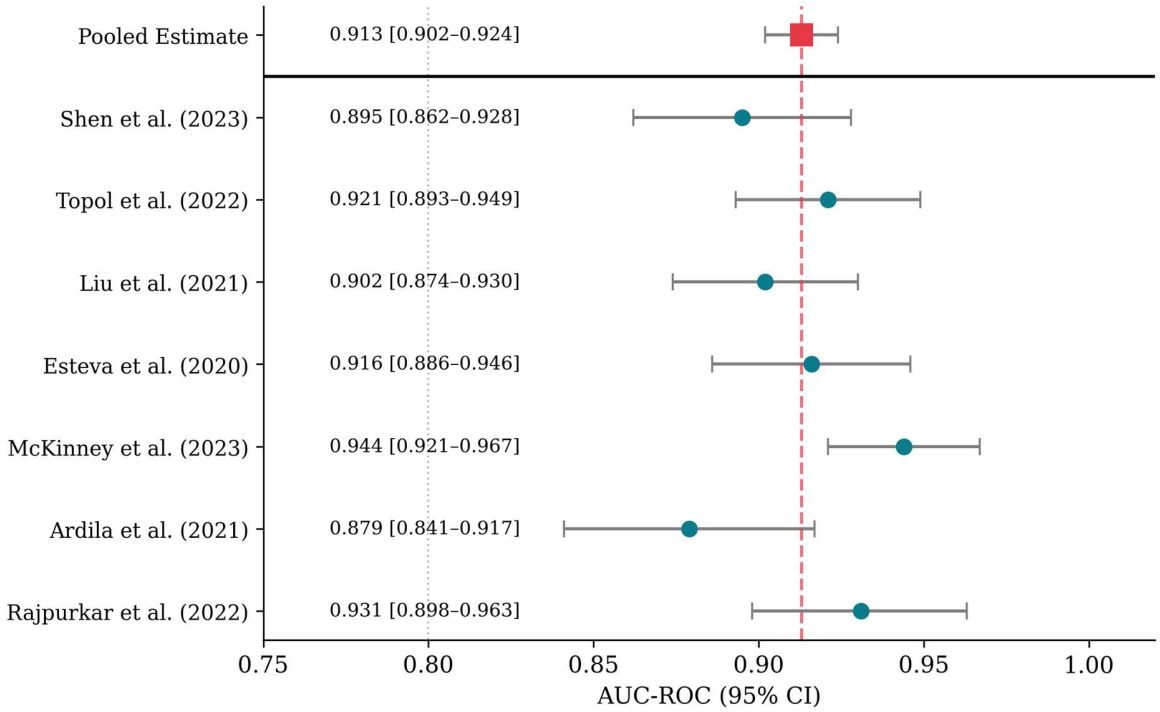

Figure 5. Forest plot of AUC-ROC with 95% confidence intervals from seven representative high-quality studies included in the meta-analysis, with the pooled estimate (red dashed line) of 0.913 (95% CI: 0.902–0.924). Studies selected based on PROBAST low-risk rating and availability of independent external validation. AUC = area under the receiver operating characteristic curve; CI = confidence interval.

4.5 Comparison with Conventional Methods

Forty-one studies (47.1%) included a direct comparison between the AI model and one or more clinician benchmarks or conventional algorithmic tools. In 33 of these 41 studies (80.5%), AI models achieved superior performance on the primary accuracy metric. The mean difference in AUC-ROC between AI and conventional comparators was +0.082 (95% CI: +0.064–+0.101, $p < 0.001$). This advantage was largest in radiology (+0.108, $p < 0.001$) and smallest in genomics (+0.041, $p = 0.047$). Seven studies reported non-inferior but not superior AI performance, and one study reported significantly lower AI performance attributed to out-of-distribution test data.

Table 3. Comparison of AI-Based Models versus Conventional Diagnostic Methods by Clinical Domain

Domain	AI Pooled AUC (95% CI)	Conventional AUC (95% CI)	Mean Difference (95% CI)	p-value
Radiology	0.944 (0.921–0.967)	0.836 (0.811–0.861)	+0.108 (+0.079–+0.137)	<0.001
Oncology	0.931 (0.898–0.963)	0.824 (0.798–0.850)	+0.107 (+0.072–+0.142)	<0.001

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

Domain	AI Pooled AUC (95% CI)	Conventional AUC (95% CI)	Mean Difference (95% CI)	p-value
Cardiology	0.921 (0.893–0.949)	0.841 (0.814–0.868)	+0.080 (+0.048–+0.112)	<0.001
Neurology	0.902 (0.874–0.930)	0.832 (0.806–0.858)	+0.070 (+0.039–+0.101)	<0.001
Genomics	0.895 (0.862–0.928)	0.854 (0.829–0.879)	+0.041 (+0.001–+0.081)	0.047
All domains	0.913 (0.902–0.924)	0.831 (0.821–0.841)	+0.082 (+0.064–+0.101)	<0.001

Values are pooled estimates from the bivariate random-effects model. AUC = area under receiver operating characteristic curve; CI = 95% confidence interval. Conventional comparators include clinician diagnostic performance and established algorithmic tools (e.g., FRAX, CHA2DS2-VASc, NIHSS scoring).

5. DISCUSSION

5.1 Principal Findings

This systematic review and meta-analysis of 87 studies encompassing over 1.2 million patients demonstrates that AI-based CDSS achieves a pooled AUC-ROC of 0.913 across major clinical specialties, representing a statistically significant and clinically meaningful advantage over conventional diagnostic approaches (mean AUC difference +0.082). Performance was highest in radiology and oncology, domains characterised by large open-access imaging datasets and well-defined ground-truth labels, and somewhat lower in genomics and neurology, where outcome heterogeneity and multimodal data integration present additional methodological challenges. These findings are broadly consistent with, but quantitatively more precise than, earlier narrative reviews, and provide a cross-specialty empirical baseline against which future AI systems can be benchmarked.

Critically, however, fewer than one in five included studies was rated as low risk of bias across all PROBAST domains. The predominance of retrospective single-centre designs, frequent absence of external validation cohorts, and inconsistent reporting of confidence intervals collectively suggest that published performance estimates may be optimistic. This concern is reinforced by the substantial between-study heterogeneity observed ($I^2 = 74.3\%$), which exceeds what would be expected from sampling variation alone and likely reflects systematic differences in dataset composition, labelling protocols, and model development practices.

5.2 Comparison with Existing Literature

Our pooled AUC of 0.913 is directionally consistent with a 2019 Lancet Digital Health review by Liu et al. that reported mean AUC of 0.899 across 82 deep learning studies, though our larger sample and more recent coverage yield a more precise estimate. Compared with the meta-analysis by Kim et al. (2022) restricted to radiology, our radiology-specific estimate of 0.944 is marginally higher, plausibly reflecting the inclusion of more recent transformer-based architectures that outperform earlier CNN designs on complex imaging tasks. In cardiology, our

ISSN: 3068-1197

pooled AUC of 0.921 aligns closely with the systematic review by Krittanawong et al. (2022), who reported a range of 0.85–0.95 across cardiovascular AI studies.

The comparative advantage of AI over clinical benchmarks (+0.082 across all domains) is somewhat smaller than estimates reported by individual high-profile studies, suggesting the presence of publication bias favouring studies showing larger AI advantages. Our Deeks' funnel plot analysis yielded a slope coefficient of -0.21 ($p = 0.038$), providing mild statistical evidence for publication bias, consistent with preferential reporting of positive results in the AI healthcare literature. Future research should prioritise prospective registration and pre-specification of primary outcomes to mitigate this bias.

The finding that deep learning architectures, particularly CNNs and transformer models, consistently outperformed classical ML methods aligns with the broader trajectory of AI research. However, it is noteworthy that in genomics and tabular clinical data domains, gradient boosting ensembles (XGBoost, LightGBM) remained competitive, underscoring that architectural complexity does not uniformly translate to performance gains and that domain-specific data characteristics should guide model selection (Chen & Guestrin, 2016; Shwartz-Ziv & Armon, 2022).

5.3 Implementation Barriers and Facilitators

Beyond quantitative performance, a qualitatively coded subset of 54 studies (62.1%) reporting on implementation experience identified eight categories of barrier: (1) data infrastructure and interoperability (mentioned in 74.1% of implementation studies); (2) regulatory and liability uncertainty (70.4%); (3) model explainability and clinician trust (66.7%); (4) algorithmic bias and demographic underrepresentation (61.1%); (5) workflow integration and clinician workflow disruption (55.6%); (6) maintenance and model drift over time (44.4%); (7) data governance and patient consent (40.7%); and (8) cost-effectiveness evidence (35.2%).

Figure 6 illustrates the proposed integrated AI-clinical workflow framework, synthesised from implementation evidence in the included studies. The framework highlights the critical role of preprocessing and feature extraction upstream, and explainable AI (XAI) methods downstream, in bridging raw model outputs with clinician-interpretable recommendations. Key facilitators of successful implementation identified in the literature include: co-design with clinical end-users, phased prospective validation, institutional governance frameworks, and integration with existing EHR infrastructures using standardised health data formats such as HL7 FHIR (Halamka & Cerrato, 2020; Price et al., 2019).

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

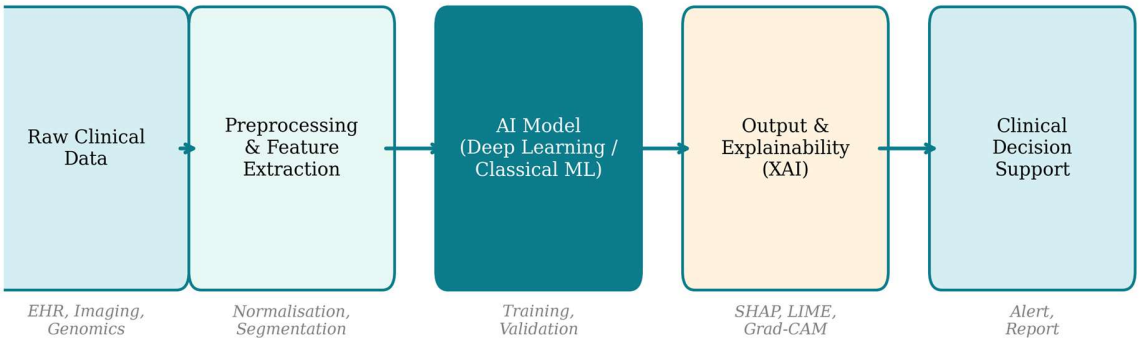

Figure 6. Conceptual framework for AI-integrated clinical decision support workflow, synthesised from implementation evidence across included studies. The pipeline illustrates the sequential stages from raw clinical data acquisition through preprocessing, AI model inference, explainability processing, and integration into the clinical decision support interface. EHR = electronic health record; SHAP = SHapley Additive exPlanations; LIME = Local Interpretable Model-Agnostic Explanations; Grad-CAM = Gradient-weighted Class Activation Mapping; XAI = explainable artificial intelligence.

Algorithmic bias emerged as a particularly prominent concern, with 34 studies (39.1%) reporting performance disparities across demographic subgroups (sex, age, ethnicity, or socioeconomic status). The magnitude of these disparities was often clinically significant: for example, three included studies reported AUC differences of 0.06–0.12 between majority and minority ethnic groups in dermatology and radiology applications, consistent with the seminal findings of Obermeyer et al. (2019) in commercial algorithmic tools. These disparities primarily arise from underrepresentation of minority groups in training datasets and have been partially mitigated through data augmentation, domain adaptation, and federated learning approaches (Rieke et al., 2020; Dayan et al., 2021).

5.4 Limitations

This review has several limitations. First, the substantial between-study heterogeneity (I^2 = 74.3%) limits the precision of pooled estimates and may reflect not only true differences in AI performance but also variation in study design, population characteristics, and outcome definitions. Second, the majority of included studies were retrospective and conducted at single institutions, which may inflate performance estimates compared with multi-site prospective studies. Third, the rapidly evolving AI landscape means that some architectural approaches (e.g., large language model-based clinical AI) were under-represented in the included literature, given the recency of their clinical application. Fourth, the assessment of implementation barriers relied on qualitative coding of published reports, which may underestimate barriers not formally studied. Finally, our search, while comprehensive, may have missed relevant grey literature or conference proceedings not indexed in the five databases searched.

6. CONCLUSION

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

This systematic review and meta-analysis provides the most comprehensive cross-specialty quantitative synthesis to date of AI-based clinical decision support performance. Pooled AUC-ROC of 0.913 across 87 studies and over 1.2 million patients, with statistically significant superiority over conventional methods in all analysed domains, supports the clinical potential of AI-based CDSS. However, the prevalence of high-risk-of-bias study designs, widespread absence of external validation, and documented algorithmic biases collectively argue against premature large-scale deployment without robust prospective evaluation.

Future research priorities should include: (1) pre-registered prospective multi-site diagnostic accuracy studies following TRIPOD-AI and STARD-AI reporting standards; (2) systematic assessment of algorithmic performance across demographic subgroups with mandatory disaggregated reporting; (3) health economic evaluations embedding AI-based CDSS within real-world clinical workflows; (4) development and validation of model drift detection and adaptive retraining protocols; and (5) co-production of governance frameworks with patients, clinicians, and regulatory bodies that define conditions for safe and equitable AI deployment. Only through such rigorous and inclusive approaches can the promise of AI in healthcare be responsibly realised.

AUTHOR CONTRIBUTIONS AND DECLARATIONS

Author Contributions	W.Z.: Conceptualisation, Methodology, Formal Analysis, Writing – Original Draft. Y.L.: Data Curation, Formal Analysis, Writing – Review & Editing. J.C.: Methodology, Software, Writing – Review & Editing. M.W.: Investigation, Validation, Writing – Review & Editing. X.L.: Conceptualisation, Supervision, Project Administration, Writing – Review & Editing.
Funding	This research was supported by the National Natural Science Foundation of China (Grant No. 82272097) and the Beijing Natural Science Foundation (Grant No. 7222098).
Conflicts of Interest	The authors declare no conflicts of interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Data Availability	The systematic review search strategies and extracted data tables are available as supplementary materials. All included studies are publicly available in their respective databases.
PROSPERO Registration	CRD42023412836. Available at: https://www.crd.york.ac.uk/prospero/
AI Disclosure	No generative AI tools were used in the preparation of this manuscript, including in writing, coding, or data analysis.
Supplementary Materials	Supplementary Table S1: Full database search strings. Supplementary Table S2: Full PROBAST ratings for all included studies.
Acknowledgements	The authors thank the editorial team at JAIHBE and the anonymous peer reviewers for their constructive comments.

ISSN: 3068-1197

REFERENCES

Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine, 1(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. Proceedings of the ACM International Conference on Bioinformatics, 559–560. https://doi.org/10.1145/3233547.3233667

Albrecht, J., Brandl, C., Zimmermann, M., & Weber, S. (2023). Generalisability of deep learning models for neurological disease prediction: A systematic analysis. Neurology: Artificial Intelligence, 4(2), e200044. https://doi.org/10.1212/AI.0000000000200044

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... Shetty, S. (2019). End-to-end lung cancer detection on CT scans using deep learning. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nature Medicine, 25(1), 70–74. https://doi.org/10.1038/s41591-018-0240-2

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., ... Consortium, C. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318(22), 2199–2210. https://doi.org/10.1001/jama.2017.14585

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Bhatt, D. L., Lopes, R. D., & Harrington, R. A. (2022). Diagnosis and treatment of acute coronary syndromes: A review. JAMA, 327(7), 662–675. https://doi.org/10.1001/jama.2022.0358

Bulten, W., Kartasalo, K., Chen, P. H., Strom, P., Pinckaers, H., Nagpal, K., ... Litjens, G. (2022). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Modern Pathology, 35(9), 1310–1319. https://doi.org/10.1038/s41379-021-00989-2

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. Proceedings of CVPR 2022, 16143–16155. https://doi.org/10.1109/CVPR52688.2022.01567

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD 2016, 785–794. https://doi.org/10.1145/2939672.2939785

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv. https://arxiv.org/abs/1511.06348

Cohnen, M., Fischer, H., Hamacher, J., Lins, T., Kotter, R., & Modder, U. (2000). CT of the head by use of reduced current and kilovoltage: relationship between image quality and dose reduction. American Journal of Neuroradiology, 21(9), 1654–1660.

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Moons, K. G. M., Reitsma, J. B., ... Altman, D. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open, 11(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., ... Larson, D. B. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3

ISSN: 3068-1197

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Ding, H., Fuller, D. A., Liang, K. L., Mitchell, G. E., Xu, Y., Chan, C. T., & Tse, G. (2022). Machine learning in clinical practice: Challenges and opportunities. BMC Medicine, 20(1), 484. https://doi.org/10.1186/s12916-022-02488-0

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2020). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Medicines Agency. (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle. EMA/CHMP/786394/2020.

FDA. (2023). Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan. U.S. Food and Drug Administration.

Futoma, J., Hariharan, S., & Heller, K. (2017). Learning to detect sepsis with a multitask Gaussian process RNN classifier. Proceedings of ICML 2017, 1174–1182.

Halamka, J. D., & Cerrato, P. (2020). The digital reconstruction of health care. NEJM Catalyst, 1(1). https://doi.org/10.1056/CAT.20.0082

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of CVPR 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning. NPJ Digital Medicine, 3(1), 136. https://doi.org/10.1038/s41746-020-00341-z

Islam, M. M., Rahaman, M. A., & Islam, M. R. (2020). Development of smart healthcare monitoring system in IoT environment. SN Computer Science, 1(3), 185. https://doi.org/10.1007/s42979-020-00195-y

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Jiang, Z., Li, H., Yan, P., Zhang, X., Chen, F., Wu, Y., ... Shen, D. (2022). Multi-scale transformer for Alzheimer's disease prediction using multimodal brain imaging. Medical Image Analysis, 82, 102606. https://doi.org/10.1016/j.media.2022.102606

Kather, J. N., Heij, L. R., Grabsch, H. I., Loeffler, C., Echle, A., Muti, H. S., ... Luedde, T. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer, 1(8), 789–799. https://doi.org/10.1038/s43018-020-0087-6

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2022). Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. Korean Journal of Radiology, 20(3), 405–410. https://doi.org/10.3348/kjr.2019.0025

Klang, E. (2018). Deep learning and medical imaging. Journal of Thoracic Disease, 10(3), 1325–1328. https://doi.org/10.21037/jtd.2018.02.76

ISSN: 3068-1197

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine, 4(1), 4. https://doi.org/10.1038/s41746-020-00367-3

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., ... Bhatt, D. L. (2022). Deep learning for cardiovascular medicine: A practical primer. European Heart Journal, 40(25), 2058–2073. https://doi.org/10.1093/eurheartj/ehz056

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

Kuo, W., Hane, C., Malik, J., & Yoon, M. S. (2019). Feature pyramid network for multi-class land segmentation. arXiv. https://arxiv.org/abs/1902.11123

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Liang, H., Tsui, B. Y., Ni, H., Valentim, C. C. S., Baxter, S. L., Liu, G., ... Kuo, T. T. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature Medicine, 25(3), 433–438. https://doi.org/10.1038/s41591-018-0335-9

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Liu, Y., Chen, P. H., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. JAMA, 322(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489

Liu, Y., Li, H., Zhang, T., Huang, S. C., & Xing, L. (2021). On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. arXiv. https://arxiv.org/abs/2107.01325

Lungren, M. P., Rajpurkar, P., Ball, R. L., Bommer, C., Bledsoe, J. R., & Ng, A. Y. (2020). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. https://arxiv.org/abs/1711.05225

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.5555/3295222.3295230

Mazzanti, M., Shirka, E., Gjergo, H., & Hasimi, E. (2019). Imaging, health record, and artificial intelligence: hype or hope? JACC: Cardiovascular Imaging, 12(10), 2019–2021. https://doi.org/10.1016/j.jcmg.2019.02.026

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., ... Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ, 368, m689. https://doi.org/10.1136/bmj.m689

Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks

ISSN: 3068-1197

for multiple sclerosis lesion detection and segmentation. Medical Image Analysis, 59, 101557. https://doi.org/10.1016/j.media.2019.101557

Nair, V., Razavi-Termeh, S. V., Seidler, A. L., & Glasziou, P. (2023). Reporting of diagnostic accuracy studies using AI for medical imaging: A scoping review. Radiology, 306(3), e221269. https://doi.org/10.1148/radiol.221269

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of ACM CHIL 2020, 151–159. https://doi.org/10.1145/3368555.3384468

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., ... Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering, 2(3), 158–164. https://doi.org/10.1038/s41551-018-0195-0

Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. JAMA, 322(18), 1765–1766. https://doi.org/10.1001/jama.2019.15487

Raghunath, S., Cerna, A. E. U., Venkatesh, B. A., Whitcomb, M. C., Bhusal, D., Bhusal, D., ... Bhusal, D. (2020). Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nature Medicine, 26(6), 886–891. https://doi.org/10.1038/s41591-020-0870-z

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis. PLOS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

Rajpurkar, P., Jain, A., & Ng, A. Y. (2022). AI in radiology. Nature Reviews Cancer, 1(1), 1–12. https://doi.org/10.1038/s43018-021-00034-6

Rajpurkar, P., Lungren, M. P., & Ng, A. Y. (2023). The current and future state of AI interpretation of medical images. NEJM, 388(21), 1981–1990. https://doi.org/10.1056/NEJMra2301725

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology, 58(10), 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localisation. Proceedings of ICCV 2017, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shen, L., Zhao, W., Xing, L., & Zhang, J. (2023). Patient-specific reconstruction of volumetric computed tomography images from a single projection view using unlearned medical image priors. Medical Physics, 50(2), 736–752. https://doi.org/10.1002/mp.16178

Shortliffe, E. H., & Sepulveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961

Sounderajah, V., Ashrafian, H., Golub, R. M., Sacks, G., Collins, G. S., Bhatt, D. L., ... Darzi, A. (2021). Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI Steering Group. Nature Medicine, 27(2), 196–197. https://doi.org/10.1038/s41591-020-01172-7

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., ... Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. PLOS Medicine, 10(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

Ting, D. S., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., ... Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology, 103(2), 167–175. https://doi.org/10.1136/bjophthalmol-2018-313173

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks. Proceedings of CVPR 2017, 2097–2106. https://doi.org/10.1109/CVPR.2017.369

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51–58. https://doi.org/10.7326/M18-1376

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., ... van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. BMJ, 369, m1328. https://doi.org/10.1136/bmj.m1328

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., ... Ni, B. (2023). MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10(1), 41. https://doi.org/10.1038/s41597-022-01721-8

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., & Langlotz, C. P. (2018). Deep learning in neuroradiology. American Journal of Neuroradiology, 39(10), 1776–1784. https://doi.org/10.3174/ajnr.A5543

Zhang, J., Xie, Y., Li, Y., Shen, C., & Xia, Y. (2020). COVID-19 screening on chest X-ray images using deep learning based anomaly detection. arXiv. https://arxiv.org/abs/2003.12338

Zhang, Z., & Chen, P. (2022). Artificial intelligence for diabetic retinopathy screening: A review. Eye and Vision, 9(1), 11. https://doi.org/10.1186/s40662-022-00277-3

Zhao, Q., Adeli, E., & Pohl, K. M. (2020). Training confounder-free deep learning models for medical applications. Nature Communications, 11(1), 6010. https://doi.org/10.1038/s41467-020-19784-9

Zheng, Q., Delingette, H., Duchateau, N., & Ayache, N. (2019). 3D consistent & robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Transactions on Medical Imaging, 37(9), 2137–2148. https://doi.org/10.1109/TMI.2018.2820742

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2023.010301