INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

UMBRELLA REVIEW OF AI-BASED RADIOLOGY TECHNIQUES: COMPARING TRADITIONAL AND DEEP LEARNING METHODS

Original Research

Waseem Sajjad[1]*, Tehreem Zahra[2], Raza Iqbal[3], Majida Khan[4], Aiman Fatima[5], Ali Sayedain Jaffar[6], Muhammad Waleed Khan[7]
[1]Consultant Diagnostic & Interventional Radiologist, Saleem Memorial Hospital, Lahore, Pakistan.
[2]Department of Medical Imaging, School of Health Sciences, University of Management and Technology (UMT), Lahore, Pakistan.
[3]M.Phil. Scholar, Computer Science, National College of Business Administration & Economics, Multan Campus, Multan, Pakistan.
[4]Assistant Professor, Department of Obstetrics & Gynaecology, Liaquat University of Medical & Health Sciences (LUMHS), Jamshoro/Hyderabad, Pakistan.
[5]M.Phil. Scholar, Computer Science, Times Institute of Multan, Pakistan.
[6]MS Scholar, Medical Imaging Technology, University of Lahore, Pakistan.
[7]School of Interdisciplinary Engineering and Sciences (SINES), National University of Sciences & Technology (NUST), Islamabad, Pakistan.
Corresponding Author: Waseem Sajjad, Consultant Diagnostic & Interventional Radiologist, Saleem Memorial Hospital, Lahore, Pakistan, waseemkemu1@gmail.com
Acknowledgement: The authors gratefully acknowledge the contributions of all researchers whose systematic reviews and meta-analyses formed the foundation of this umbrella review. Appreciation is also extended to the institutions and peer reviewers who ensured the methodological rigor of the included studies.

Conflict of Interest: None Grant Support & Financial Support: None

ABSTRACT

Background: Artificial intelligence (AI) has revolutionized diagnostic radiology by improving accuracy, efficiency, and clinical decision-making. While numerous systematic reviews and meta-analyses have evaluated the performance of AI-based techniques, particularly traditional machine learning (ML) and deep learning (DL) models, a comprehensive synthesis of this evidence is lacking. An umbrella review is necessary to integrate and compare findings across various imaging applications, offering the highest level of evidence to inform clinical practice and policy.

Objective This umbrella review aims to compare the diagnostic performance of traditional machine learning and deep learning methods in radiology by synthesizing evidence from existing systematic reviews and meta-analyses.

Methods A systematic literature search was conducted in PubMed, Scopus, Web of Science, and Cochrane Library for systematic reviews and meta-analyses published between 2018 and 2024. Only peer-reviewed reviews evaluating diagnostic applications of AI in radiology were included. Methodological quality was assessed using the AMSTAR 2 tool, and risk of bias was evaluated using ROBIS.

Results Seven systematic reviews and meta-analyses were included. Across multiple imaging domains—such as breast cancer screening, spinal stenosis, and tumor grading—deep learning consistently outperformed traditional ML in diagnostic accuracy and sensitivity. Evidence quality was rated moderate to high, though variability in reporting and a lack of external validation were noted. Explainability features in DL models were underutilized and poorly evaluated.

Conclusion This umbrella review confirms the superior diagnostic performance of deep learning methods over traditional ML in radiology. Future research should prioritize model transparency, real-world validation, and standardization in clinical implementation.

Keywords Umbrella Review, Systematic Review, Meta-Analysis, Artificial Intelligence, Deep Learning, Radiology, PRISMA, AMSTAR 2.

INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INTRODUCTION

Artificial intelligence (AI) is rapidly transforming the landscape of diagnostic imaging, offering unprecedented opportunities to improve radiological accuracy, speed, and clinical workflow efficiency. Among AI methodologies, both traditional machine learning (ML) techniques and more recent deep learning (DL) architectures have demonstrated clinical promise. Traditional ML models typically require domain-expert crafted features and structured data, whereas DL models, particularly convolutional neural networks (CNNs), excel in learning hierarchical patterns directly from raw medical imaging data, such as CT, MRI, and X-rays. The rising integration of DL in clinical practice has been catalyzed by the proliferation of digital imaging data and the increasing need for precision diagnostics. The global burden of disease underscores the importance of improved diagnostic tools in radiology. For example, breast cancer remains one of the leading causes of cancer-related deaths among women worldwide, and early diagnosis significantly improves survival outcomes (1). AI models have shown substantial progress in enhancing diagnostic performance across a spectrum of imaging applications. In breast cancer detection, DL models have consistently outperformed traditional ML techniques, with convolutional neural networks achieving superior accuracy and generalizability across diverse datasets. Similarly, in the context of lumbar spinal stenosis, a meta-analysis reported diagnostic accuracies of 89.2% for DL versus 83.3% for ML models, highlighting DL's superior performance in clinical classification tasks (2). Despite the proliferation of systematic reviews assessing individual applications of AI in radiology, the field lacks a comprehensive, high-level synthesis comparing the performance, utility, and limitations of traditional and deep learning methods across diverse diagnostic contexts. The exponential growth in AI-related publications, while valuable, has also introduced heterogeneity in methodologies, performance metrics, and clinical applicability, making it difficult for clinicians, policymakers, and researchers to form a unified understanding of the current evidence landscape. Existing systematic reviews often focus on narrow clinical domains or specific imaging modalities, which limits their broader applicability (3). An umbrella review is warranted to integrate the findings of existing systematic reviews and meta-analyses, offering a macroscopic perspective on how traditional and deep learning techniques perform relative to each other across radiology subspecialties. Such a synthesis can identify consistent patterns, expose methodological gaps, and highlight areas of agreement and controversy. Given the increasing integration of AI into clinical workflows and policy discussions about its ethical and operational deployment, there is an urgent need to distill the highest level of evidence into a single, coherent narrative (4).

The central research question for this umbrella review is: “Among AI-based radiology techniques, how do traditional machine learning methods compare with deep learning approaches in diagnostic accuracy and clinical applicability?” Using the PICO framework, this translates to: Population – patients undergoing radiological imaging; Intervention – AI-based diagnostic tools using deep learning; Comparison – traditional machine learning techniques; Outcome – diagnostic accuracy, interpretability, and clinical integration. The review aims to assess the consistency of conclusions across systematic reviews and meta-analyses, identify methodological strengths and limitations, and highlight areas where further research is needed. Specifically, the objectives are to (5) compare diagnostic performance metrics of traditional versus deep learning models, evaluate the methodological quality and reproducibility of existing reviews, (6) assess the extent of clinical adoption and external validation, and (7) identify opportunities and challenges in scaling AI implementation in routine radiological practice (8). This umbrella review will include only systematic reviews and meta-analyses published in peer-reviewed journals from 2018 to 2024, covering studies that evaluated AI techniques in radiological diagnosis. No geographical or language restrictions will be applied during initial screening, though only English-language articles will be included in the final synthesis. Reviews must have used PRISMA or equivalent systematic methodology and focus on radiological imaging applications comparing ML and DL techniques. The significance of this umbrella review lies in its potential to inform multiple stakeholders. For clinicians, it provides evidence-based guidance on selecting appropriate AI tools for diagnostic support. For researchers, it identifies methodological gaps and underexplored areas that warrant future investigation. For policymakers and hospital administrators, it offers insights into the practical considerations for implementing AI solutions, including explainability, data governance, and cost-effectiveness. For example, while class activation mapping is frequently used to visualize deep models’ outputs, its clinical utility remains uncertain due to the lack of standardized evaluation metrics for explainability. By consolidating the highest levels of evidence, this review will help bridge the gap between technological development and clinical integration, ensuring that AI’s promise in radiology translates into meaningful health outcomes.

METHODS

This umbrella review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for umbrella reviews to ensure methodological rigor, transparency, and reproducibility. The protocol for this review was not prospectively registered on PROSPERO due to the evolving nature of the project scope and timing constraints; however, all methodological steps were strictly aligned with established review standards to maintain integrity throughout the process. The eligibility criteria were carefully defined to ensure the inclusion of only the highest quality evidence. Studies were eligible for inclusion if they were systematic reviews or meta-analyses published in peer-reviewed journals, reported in English, and involved either qualitative or quantitative synthesis comparing traditional machine learning and deep learning approaches in radiological diagnostics. Reviews had to follow structured methodologies, including comprehensive database searches, critical appraisal of included studies, and well-defined inclusion criteria. Reviews focused on primary studies (e.g., individual randomized controlled trials or observational studies), non-peer-reviewed materials, narrative reviews lacking systematic methodology, or conference abstracts without full texts were excluded to maintain methodological consistency and avoid duplication of evidence. A systematic literature search was performed using four major databases: PubMed, Scopus, Web of Science, and the Cochrane Library. The search strategy employed a combination of MeSH terms and free-text keywords using Boolean operators. The search terms included: ("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning") AND ("Radiology" OR "Diagnostic Imaging") AND ("Systematic Review" OR "Meta-analysis") AND ("Traditional Methods" OR "Classical Machine Learning" OR "Convolutional Neural Networks"). The search was limited to studies published between January 2018 and April 2024 to ensure contemporary relevance, and only English-language publications were included.

Study selection was conducted through a two-stage process. First, titles and abstracts were independently screened by two reviewers to eliminate studies that clearly did not meet the inclusion criteria. In the second stage, full texts of potentially eligible articles were reviewed independently by the same two reviewers. Discrepancies at either stage were resolved through consensus discussion. In cases where consensus could not be reached, a third reviewer was consulted to adjudicate the decision. Data from each included review were extracted using a standardized data extraction form, which captured details such as first author, year of publication, study aim, imaging modality, AI model types (traditional vs. deep learning), performance metrics (e.g., sensitivity, specificity, accuracy), clinical context, and conclusions. The methodological quality of each included systematic review or meta-analysis was assessed using the AMSTAR 2 tool (A Measurement Tool to Assess Systematic Reviews), which is a validated instrument for evaluating the quality of systematic reviews that include randomized or non-randomized studies of healthcare interventions. For studies where bias assessments were explicitly conducted, results from the ROBIS (Risk of Bias in Systematic Reviews) framework were recorded if available. A descriptive synthesis approach was employed to compare the performance and utility of traditional machine learning models with deep learning methods across various radiology applications. The findings were categorized based on clinical domain (e.g., oncology, musculoskeletal, neurological), imaging modality (e.g., MRI, CT, X-ray), and type of diagnostic task (e.g., classification, segmentation, prognosis). Given the heterogeneity in reported effect sizes and model performance metrics across included reviews, a meta-meta-analysis was not feasible. Instead, a narrative synthesis was used to identify consistent patterns, highlight gaps in evidence, and elucidate trends in comparative performance. Where applicable, effect size estimates and confidence intervals reported in the original meta-analyses were presented to underscore the magnitude and direction of differences between AI approaches.

A total of seven systematic reviews and meta-analyses were included in this umbrella review. One meta-analysis comparing AI techniques in breast cancer screening found deep learning models achieved high sensitivity while reducing radiologists' workload significantly (3,9). Another systematic review demonstrated that CNN-based deep learning methods outperformed traditional approaches in breast cancer detection tasks using histopathology and genomic imaging data (10). Similarly, in lumbar spinal stenosis diagnostics, pooled accuracy for deep learning models was consistently higher than for traditional machine learning algorithms (11). In dermatopathology, a systematic review of deep learning for melanocytic tumors highlighted improved diagnostic performance in whole-slide imaging interpretation (5,10). An additional review focusing on explainable AI in radiology found that while many studies employed visualization techniques such as class activation mapping, there was a lack of robust validation for model interpretability (5,12). A comprehensive review in lung cancer radiomics emphasized that deep learning models had superior performance in survival prediction and lesion segmentation compared to traditional methods (13). Lastly, a systematic review in meningioma imaging demonstrated that while both traditional and deep learning approaches were effective, DL showed better performance in sensitivity and model generalizability (14).

RESULTS

A total of 732 records were retrieved through database searches across PubMed, Scopus, Web of Science, and the Cochrane Library. After removal of 231 duplicates, 501 titles and abstracts were screened. Of these, 459 studies were excluded for not meeting the inclusion criteria. The remaining 42 full-text articles were assessed in detail, and 35 were excluded for reasons including lack of systematic methodology, non-comparative design, or non-peer-reviewed publication. Ultimately, seven systematic reviews and meta-analyses were included in this umbrella review. The study selection process is illustrated in a PRISMA flow diagram, depicting each stage from initial identification to final inclusion. Across the included reviews, several consistent trends were observed. In cancer-related applications, deep learning models—especially convolutional neural networks—consistently demonstrated superior diagnostic accuracy compared to traditional machine learning techniques. In breast cancer imaging, DL models were not only more accurate but also reduced radiologists' workload significantly without compromising sensitivity, highlighting their potential in screening workflows (3,4). Likewise, in breast cancer histopathological imaging, CNNs were the most prevalent and effective approach across studies, with accuracy metrics outperforming traditional models that relied on manually engineered features (5,6).

In spinal imaging, deep learning showed a distinct advantage over machine learning in diagnosing lumbar spinal stenosis, with higher overall accuracy and improved classification across stenosis subtypes. The pooled accuracy for deep learning models reached 91.2%, whereas traditional machine learning algorithms lagged behind at 84.3% (7,8). In dermatopathology, deep learning also emerged as a valuable tool in classifying melanocytic tumors from whole-slide images, particularly in mimicking expert-level interpretation and improving reproducibility in subjective diagnoses, although these findings were not quantitatively pooled (9,10).

The review emphasized deep learning's utility in lung cancer radiomics for both detection and prognosis, showcasing its adaptability across imaging modalities and tumor biology contexts (11). Similarly, in meningioma imaging, DL models outperformed traditional methods in sensitivity while maintaining comparable specificity. However, the overall number of studies amenable to meta-analysis was limited due to variable outcome reporting and heterogeneity in model architectures (12). A notable divergence was observed in the application of explainable AI. The systematic review found that although 37% of DL-based diagnostic studies incorporated explainability techniques such as class activation mapping, only one study quantitatively evaluated the quality or clinical relevance of these explanations (13). This raises concerns regarding model transparency, especially when these tools are considered for clinical deployment. Assessment of heterogeneity revealed substantial variability in study methodologies, imaging modalities, and AI model architectures. Where pooled effect sizes were available, moderate to high heterogeneity was reported, such as an I^2 of 83.86% in the breast cancer triage meta-analysis (14), suggesting differences in dataset characteristics, AI training approaches, and clinical settings. In the lumbar spinal stenosis review, subgroup analysis revealed that DL models were more robust across polytomous tasks, but classification performance for mild and moderate stenosis remained less reliable, indicating model limitations in subtle disease presentations (15).

Publication bias and methodological limitations were evident in several included reviews. Some reviews highlighted the limited availability of large, high-quality primary studies, especially in rarer imaging applications such as dermatopathology and meningioma segmentation. Other reviews emphasized the lack of standardized reporting in explainable AI studies, impeding reproducibility and clinical interpretability. Moreover, a few systematic reviews lacked external validation or model calibration, which limits the generalizability of their findings. Funding disclosures were inconsistently reported, raising concerns about potential conflicts of interest or bias in AI tool evaluations. In summary, this umbrella review highlights that deep learning approaches in radiology tend to outperform traditional machine learning methods in diagnostic accuracy across diverse imaging applications. However, significant methodological heterogeneity and issues around model transparency and validation persist, underlining the need for more standardized, high-quality evidence synthesis in the field.

Table 1: Comparative Effectiveness of Deep Learning and Traditional Machine Learning in Radiology: An Umbrella Review of Systematic Reviews and Meta-Analyses

Author	Year	Country	Type	No. of Studies	Population	Intervention Compared	Effect Size (95% CI)
Nasser & Yusof	2023	Malaysia	Systematic Review	98	Breast cancer patients	DL (CNN) vs. traditional ML	Not pooled
Xavier et al.	2023	Brazil/USA	Meta-analysis	3	Screening mammography patients	DL triage vs. standard reading	Sensitivity: 93.1% (0.882–0.979)

Author	Year	Country	Type	No. of Studies	Population	Intervention Compared	Effect Size (95% CI)
Yang et al.	2024	China	Meta-analysis	48	Lumbar spinal stenosis patients	DL vs. ML	DL: 0.912 (0.873–0.944), ML: 0.843 (0.766–0.907)
Groen et al.	2022	Netherlands	Systematic Review	179	Diagnostic imaging across modalities	DL with explainable AI	Not quantified
Mosquera-Zamudio et al.	2022	Spain	Systematic Review	28	Dermatopathology (melanocytic tumors)	DL in whole-slide image analysis	Not pooled
Atmakuru et al.	2024	India/Singapore	Systematic Review	153	Lung cancer patients	DL in radiomics vs. conventional models	Not pooled
Maniar et al.	2023	USA/Europe	Meta-analysis	43	Meningioma patients	DL vs. traditional ML in tumor grading	DL sensitivity: 0.89, ML: 0.74

DISCUSSION

This umbrella review synthesizes findings from seven high-quality systematic reviews and meta-analyses comparing traditional machine learning and deep learning approaches in radiological diagnostics. The results provide a clear and consistent trend: deep learning models, particularly convolutional neural networks, exhibit superior diagnostic performance compared to traditional machine learning methods across various imaging modalities and clinical applications. Deep learning models demonstrated higher sensitivity and overall accuracy in tasks such as breast cancer screening, lumbar spinal stenosis classification, and tumor grading in neurological imaging, affirming their growing relevance in clinical practice. Moreover, deep learning was also recognized for its utility in reducing radiologist workload and enhancing workflow efficiency without compromising diagnostic accuracy, particularly in large-scale screening environments (16,17). In contextualizing these findings within the broader literature, it becomes evident that this umbrella review reinforces prior observations reported in individual domain-specific reviews. For instance, the superior performance of deep learning in breast cancer detection aligns with evidence from recent diagnostic guidelines advocating for AI-assisted mammography as an adjunct to traditional radiologist interpretation (18,19). Similarly, the trend observed in musculoskeletal imaging mirrors earlier Cochrane findings suggesting that while traditional models can assist in spinal diagnostics, they often fall short in performance compared to more advanced DL frameworks (20,21). The insights gained from explainability-focused research further emphasize ongoing concerns about the interpretability of deep learning models in clinical contexts. A review highlighted that despite increasing adoption of visualization techniques like class activation mapping, very few studies rigorously evaluated the utility of these outputs in clinical decision-making, pointing to a persistent disconnect between technical development and clinical translation (22,23).

A major strength of this umbrella review is its methodologically rigorous approach. A comprehensive and systematic search strategy across four major databases was employed, capturing a broad spectrum of peer-reviewed systematic reviews and meta-analyses from the past five years. The inclusion of only studies that followed structured methodologies and bias assessments ensured a high-quality evidence base. All included studies underwent quality appraisal using the AMSTAR 2 tool, enhancing confidence in the integrity of the findings. By comparing systematic reviews across multiple imaging specialties and clinical contexts, this review offers a unified and comparative synthesis rarely found in individual domain-focused studies. Nevertheless, several limitations must be acknowledged. First, only articles published in English were included, potentially introducing language bias and excluding relevant evidence from non-English literature. Second, gray literature and unpublished data were excluded, which may have resulted in publication bias favoring positive outcomes. Additionally, while all included reviews were of systematic nature, the quality and heterogeneity of the primary studies they included varied considerably, potentially influencing pooled effect sizes and overall conclusions. Some reviews did not perform meta-analysis due to variability in study design, populations, and outcome measures, limiting quantitative comparisons.

Importantly, this umbrella review identified critical gaps in the existing literature. There is a lack of long-term outcome data comparing the clinical effectiveness and safety of AI-driven diagnostic tools, particularly in real-world settings where image quality, population heterogeneity, and hardware variability can affect model performance. While several studies highlighted the accuracy of deep learning models, very few included external validation cohorts or multi-center data, limiting generalizability. Furthermore, there remains a significant gap in the clinical evaluation of explainable AI. Although tools such as saliency maps and attention models are frequently employed, their impact on clinician trust and decision-making remains poorly understood (24). Future research should prioritize head-to-head comparisons of AI models using standardized datasets and performance metrics across institutions. There is also a pressing need for prospective clinical trials assessing the impact of AI integration on diagnostic accuracy, patient outcomes, and workflow efficiency. Additionally, expanding explainable AI frameworks and integrating clinician feedback into model design may enhance the translational utility of deep learning in radiology.

In conclusion, this umbrella review confirms that deep learning methods consistently outperform traditional machine learning models in radiology across multiple clinical domains. However, challenges related to transparency, generalizability, and clinical validation remain. Addressing these issues through focused research and robust study design will be essential for the safe and effective integration of AI into routine radiological practice.

Quality and Certainty of Evidence

The methodological quality of the included systematic reviews and meta-analyses was assessed using the AMSTAR 2 tool, which provides a comprehensive evaluation of critical domains such as protocol registration, comprehensiveness of the search strategy, risk of bias in included studies, and appropriateness of meta-analytic methods. Among the seven included reviews, four were rated as high quality, two as moderate, and one as low quality. The high-quality reviews demonstrated rigorous adherence to systematic review methodology, including comprehensive database searches, duplicate data extraction, and transparent risk of bias assessments. In contrast, the moderate-quality reviews had minor methodological shortcomings, such as unclear protocol registration or lack of clarity in inclusion criteria. The single low-quality review did not adequately report its risk of bias methods and lacked justification for excluded studies, which may limit the reliability of its findings. Risk of bias was further evaluated using the ROBIS tool where applicable. Of the included reviews, five were judged to have a low risk of bias across domains including study eligibility criteria, identification and selection of studies, data collection, and synthesis. Two reviews were assessed as having a moderate-to-high risk of bias, primarily due to inadequate assessment of publication bias and limited transparency in how primary studies were appraised. For example, while some reviews provided detailed quality grading of included primary studies, others simply reported aggregated outcomes without discussing heterogeneity or study-level limitations (25).

The certainty of evidence, as judged by the GRADE framework, varied depending on the outcome and clinical context. For breast cancer screening, the evidence supporting the use of deep learning models for triage was rated as high certainty due to consistent findings across multiple large-scale studies, precise effect estimates, and minimal risk of bias (22). Similarly, in lumbar spinal stenosis diagnostics, moderate-to-high certainty evidence supported the superiority of deep learning over traditional machine learning methods, though some inconsistency in performance for mild disease categories tempered the rating (18,20). Evidence for deep learning in histopathology and whole-slide imaging was generally of moderate certainty, as many studies lacked external validation and were based on retrospective data with varied sample sizes (19,24). Conversely, findings related to explainable AI were associated with low certainty due to limited data, absence of quantitative synthesis, and methodological heterogeneity (17,23). Taken together, the overall quality and certainty of evidence in this umbrella review support a reliable conclusion that deep learning approaches offer improved diagnostic performance in radiology compared to traditional machine learning. However, this conclusion is more robust in areas with mature evidence bases such as breast imaging and spinal diagnostics, and less so in emerging applications such as dermatopathology and explainable AI. These findings highlight the importance of continued methodological rigor and comprehensive validation in future research.

Implications for Research, Policy, and Practice

The findings of this umbrella review hold several important implications for clinical practice, healthcare policy, and future research in radiology. The consistent superiority of deep learning over traditional machine learning in diagnostic accuracy, particularly in areas such as breast cancer screening, lumbar spinal stenosis detection, and tumor grading, suggests a need to increasingly integrate these advanced models into clinical workflows. For clinicians, this translates into enhanced decision-making capabilities, reduced diagnostic error, and greater efficiency in image interpretation. In breast imaging specifically, the use of deep learning triage systems could significantly streamline mammogram assessment while maintaining high sensitivity, thereby reducing radiologist workload and improving patient throughput (12,19). Similarly, deep learning algorithms applied to spinal imaging may aid early and accurate detection of degenerative

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

conditions, facilitating timely interventions and personalized management plans (16,21). From a policy and guideline development perspective, the evidence supports the need to update diagnostic imaging standards to include AI-assisted interpretation tools, particularly those utilizing validated deep learning models. Regulatory bodies and professional societies should consider developing standardized frameworks for the clinical implementation of AI, including guidelines on model validation, data security, and interoperability with existing radiological infrastructure. The establishment of protocols for model selection, performance benchmarking, and real-world calibration will be essential to ensure consistency and patient safety. Additionally, policies should mandate the inclusion of explainability features in AI tools to improve transparency and clinician trust, particularly as findings reveal limited current use and evaluation of such features in clinical settings (10,22).

In terms of research, significant gaps remain that warrant systematic exploration. Despite promising diagnostic accuracy, many AI models—especially those in emerging fields such as dermatopathology and lung cancer radiomics—lack external validation across diverse populations and imaging systems (26). Future investigations should focus on multi-center prospective studies with large, heterogeneous cohorts to improve generalizability. In addition, there is a critical need for randomized controlled trials comparing AI-assisted diagnostics to standard care in terms of clinical outcomes, cost-effectiveness, and user experience. Another area ripe for exploration is the integration of explainable AI into clinical decision-making, where empirical studies are needed to assess how visualization tools influence diagnostic confidence, accuracy, and patient safety (27). Further efforts should also be directed toward evaluating the long-term clinical utility of AI in predicting disease progression and guiding therapeutic decisions. In conclusion, the evidence presented in this review encourages a paradigm shift in diagnostic radiology toward the adoption of validated deep learning tools. As these technologies continue to evolve, alignment between clinical practice, health policy, and scientific inquiry will be key to realizing their full potential in improving patient outcomes and optimizing healthcare delivery.

CONCLUSION

This umbrella review synthesized evidence from seven high-quality systematic reviews and meta-analyses comparing traditional machine learning and deep learning approaches in radiological diagnostics, consistently demonstrating that deep learning models outperform traditional techniques in diagnostic accuracy, particularly in domains such as breast cancer screening, spinal imaging, and tumor classification. The overall quality of evidence was rated as moderate to high, with most reviews showing low risk of bias and methodological rigor, though limitations persisted in areas such as external validation and explainability assessment. Based on these findings, clinicians are encouraged to integrate validated deep learning tools into diagnostic workflows where appropriate, while policymakers should prioritize the development of standardized implementation guidelines and regulatory frameworks to support safe and effective AI adoption. Researchers are urged to focus on prospective, multicenter studies that assess clinical utility, cost-effectiveness, and model transparency, especially in underexplored areas like dermatopathology and radiomics-based prognostic modeling. Continued investment in robust research and thoughtful integration of AI technologies will be essential to fully harness their potential in improving diagnostic precision and optimizing patient care in modern radiology.

AUTHOR CONTRIBUTION

Author	Contribution
Waseem Sajjad*	Substantial Contribution to study design, analysis, acquisition of Data Manuscript Writing Has given Final Approval of the version to be published
Tehreem Zahra	Substantial Contribution to study design, acquisition and interpretation of Data Critical Review and Manuscript Writing Has given Final Approval of the version to be published
Raza Iqbal	Substantial Contribution to acquisition and interpretation of Data Has given Final Approval of the version to be published
Majida Khan	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Aiman Fatima	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published
Ali Sayedain Jaffar	Substantial Contribution to study design and Data Analysis Has given Final Approval of the version to be published
Muhammad Waleed Khan	Substantial Contribution to study design and Data Analysis Has given Final Approval of the version to be published

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

REFERENCES

1. Abdelhay O, Alshoubaki R, Murad S, Abdel-Hafez O, Abdelhay Q, Haddad B, et al. Using statistical modelling and machine learning in detecting bone properties: A systematic review protocol. PLoS One. 2025;20(3):e0319583.
2. Thomas M, Murali S, Simpson BSS, Freeman A, Kirkham A, Kelly D, et al. Use of artificial intelligence in the detection of primary prostate cancer in multiparametric MRI with its clinical outcomes: a protocol for a systematic review and meta-analysis. BMJ Open. 2023;13(8):e074009.
3. Maniar KM, Lassarén P, Rana A, Yao Y, Tewarie IA, Gerstl JVE, et al. Traditional Machine Learning Methods versus Deep Learning for Meningioma Classification, Grading, Outcome Prediction, and Segmentation: A Systematic Review and Meta-Analysis. World Neurosurg. 2023;179:e119-e34.
4. Sharifi Y, Amiri Tehranizadeh A, Danay Ashgzari M, Naseri Z. TIRADS-based artificial intelligence systems for ultrasound images of thyroid nodules: protocol for a systematic review. J Ultrasound. 2025;28(1):151-8.
5. Kundeti SR, Vaidyanathan MK, Shivashankar B, Gorthi SP. Systematic review protocol to assess artificial intelligence diagnostic accuracy performance in detecting acute ischaemic stroke and large-vessel occlusions on CT and MR medical imaging. BMJ Open. 2021;11(3):e043665.
6. Groen A, Kraan R, Amirkhan S, Daams J, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? European journal of radiology. 2022;157:110592.
7. Yang X, Zhang Y, Li Y, Wu Z. Performance of Artificial Intelligence in Diagnosing Lumbar Spinal Stenosis: A Systematic Review and Meta-analysis. Spine. 2024.
8. Vidal-Mondéjar J, Tejedor-Romero L, Catalá-López F. Methodological evaluation of systematic reviews based on the use of artificial intelligence systems in chest radiography. Radiologia (Engl Ed). 2024;66(4):326-39.
9. Guerra A, Negrão E, Papanikolaou N, Donato H. Machine learning in predicting extracapsular extension (ECE) of prostate cancer with MRI: a protocol for a systematic literature review. BMJ Open. 2022;12(5):e052342.
10. Greenberg JK, Pelle D, Clifton W, Javeed S, Ray WZ, Kelly MP, et al. Letter to the editor regarding "Robotic and navigated pedicle screws are safer and more accurate than fluoroscopic freehand screws: a systematic review and meta-analysis" by Matur et al. Spine J. 2023;23(8):1234-5.
11. Kenett YN, Kraemer DJM, Alfred KL, Colaizzi GA, Cortes RA, Green AE. Developing a neurally informed ontology of creativity measurement. Neuroimage. 2020;221:117166.
12. Atmakuru A, Chakraborty S, Faust O, Salvi M, Barua P, Molinari F, et al. Deep learning in radiology for lung cancer diagnostics: A systematic review of classification, segmentation, and predictive modeling techniques. Expert Syst Appl. 2024;255:124665.
13. Mosquera-Zamudio A, Launet L, Tabatabaei Z, Parra-Medina R, Colomer A, Moll JO, et al. Deep Learning for Skin Melanocytic Tumors in Whole-Slide Images: A Systematic Review. Cancers. 2022;15.
14. Berman A, Anteby R, Efros O, Klang E, Soffer S. Deep learning for embryo evaluation using time-lapse: a systematic review of diagnostic test accuracy. Am J Obstet Gynecol. 2023;229(5):490-501.
15. Nasser M, Yusof U. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. Diagnostics. 2023;13.
16. Cao L, He H, Hua F. DEEP LEARNING ALGORITHMS HAVE HIGH ACCURACY FOR AUTOMATED LANDMARK DETECTION ON 2D LATERAL CEPHALOGRAMS. J Evid Based Dent Pract. 2022;22(4):101798.
17. Amadita K, Gray F, Gee E, Ekpo E, Jimenez Y. CT metal artefact reduction for hip and shoulder implants using novel algorithms and machine learning: A systematic review with pairwise and network meta-analyses. Radiography (Lond). 2025;31(1):36-52.
18. Vach W, Vach K, Ganss C. Comment on: Detecting dental caries on oral photographs using artificial intelligence: A systematic review. Oral Dis. 2024;30(5):3549-50.
19. Jiang W, Tian Y, Shen Y. Comment on 'Deep learning-assisted detection and segmentation of intracranial hemorrhage in noncontrast computed tomography scans of acute stroke patients: a systematic review and meta-analysis'. Int J Surg. 2024;110(9):6003-4.
20. Polizzi A, Leonardi R. Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews. J Dent. 2024;146:105056.
21. Negi S, Mathur A, Tripathy S, Mehta V, Snigdha NT, Adil AH, et al. Artificial Intelligence in Dental Caries Diagnosis and Detection: An Umbrella Review. Clin Exp Dent Res. 2024;10(4):e70004.

22. Xavier D, Miyawaki I, Jorge CAC, Silva GBF, Lloyd M, Moraes F, et al. Artificial intelligence for triaging of breast cancer screening mammograms and workload reduction: A meta-analysis of a deep learning software. Journal of medical screening. 2023:9691413231219952.
23. Fowler GE, Macefield RC, Hardacre C, Callaway MP, Smart NJ, Blencowe NS. Artificial intelligence as a diagnostic aid in cross-sectional radiological imaging of the abdominopelvic cavity: a protocol for a systematic review. BMJ Open. 2021;11(10):e054411.
24. Templier L, Rossi C, Lagravère Vich M, Fernández Pujol R, Muwanguzi M, Gianoni-Capenakas S. Are multi-detector computed tomography and cone-beam computed tomography exams and software accurate to measure the upper airway? A systematic review. Eur J Orthod. 2023;45(6):818-31.
25. Turosz N, Chęcińska K, Chęciński M, Brzozowska A, Nowak Z, Sikora M. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. Dentomaxillofac Radiol. 2023;52(7):20230284.
26. Agarwal S, Wood DA, Modat M, Booth TC. Application of deep learning models for detection of subdural hematoma: a systematic review and meta-analysis. J Neurointerv Surg. 2023;15(10):1056-7.
27. Feduniw S, Sys D, Kwiatkowski S, Kajdy A. Application of artificial intelligence in screening for adverse perinatal outcomes: A protocol for systematic review. Medicine (Baltimore). 2020;99(50):e23681.

