Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 13 [3] February 2024 : 283-289

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD

REVIEW ARTICLE **OPEN ACCESS**

A Comprehensive Meta-Analysis: Evaluating the Sensitivity and Specificity of AI-Assisted Diagnosis in Radiology across Subspecialties and Imaging Modalities

Maajid Mohi Ud Din Malik
[1]Assistant Professor, Dr. D. Y. Patil School of Allied Health Sciences, Dr. D. Y. Patil Vidyapeeth, Sant-Tukaram Nagar, Pimpri, Pune MH, India 411018
Email: majidmalik343@gmail.com
Orcid: https://orcid.org/0000-0003-1743-1520

ABSTRACT

This meta-analysis aims to comprehensively evaluate the diagnostic accuracy of AI-assisted diagnosis across radiology based on current literature. A systematic search of major biomedical databases was conducted through January 2023 to identify original research studies evaluating AI diagnostic performance for medical images compared to human experts. Included studies (n=86) represented various clinical domains and imaging modalities. Pooled sensitivity and specificity estimates were calculated using bivariate random-effects models. The meta-analysis included over 129,000 medical images from 86 studies. The risk of bias was assessed using QUADAS-2. The overall pooled sensitivity was 0.92, and specificity was 0.93, indicating diagnostic accuracy comparable to humans. However, moderate heterogeneity was observed. Chest X-ray studies showed the highest sensitivity (0.92) and specificity (0.95). Some domains, like head/neck imaging, require further optimization. While AI achieved high accuracy, variability between studies highlights the need for standardized evaluation frameworks. Prospective clinical validation integrating AI as a decision support tool is required before widespread adoption. Additional research is needed to address current limitations and develop more transparent models. AI-assisted diagnosis has demonstrated promising diagnostic performance, but significant progress is still required. Continued algorithm development and standardization of practices can help realize AI's full benefits for radiology. Further research addressing current limitations through more significant and diverse datasets can help establish AI as an effective clinical decision-support tool.

Keywords: ***Keywords:*** *Artificial intelligence, machine learning, deep learning, computer-aided diagnosis, medical imaging, diagnostic accuracy, sensitivity and specificity, Radiology*

Received 24.12.2023 Revised 02.02.2024 Accepted 23.02.2024

INTRODUCTION

Artificial intelligence (AI) and machine learning have rapidly advanced in recent years and have shown great potential to transform many industries, including healthcare. One area that AI is poised to impact significantly is medical imaging and radiology. Radiologists play a vital role in diagnosing diseases and guiding patient treatment by interpreting various medical images. However, the growing volume of imaging exams has increased workload pressures. At the same time, there is variability in diagnostic accuracy between radiologists due to human factors like fatigue and lack of subspecialization.

AI and machine learning offer technologies that may help address these challenges by assisting radiologists in their diagnostic work. By automating large volumes of medical images, AI algorithms can learn visual patterns that may help detect diseases and abnormalities. This computer-aided diagnosis has the potential to improve consistency and efficiency in radiology. Early studies have shown promising results, with AI demonstrating capabilities comparable to human experts in some clinical domains. However, the development and evaluation of these AI systems are still early. Many questions remain regarding their real-world performance and ability to generalize across clinical scenarios.

Before AI can be reliably integrated into clinical practice, it is essential to comprehensively evaluate its diagnostic accuracy on various medical images and patient populations. This will help establish these technologies' current capabilities and limitations and identify areas where further research and development are still needed. A systematic and evidence-based approach is required to guide the safe and

effective implementation of AI in healthcare settings. This introduction provides an overview of the background, rationale and objectives of evaluating AI-assisted diagnosis in radiology through a meta-analytic framework.

Background

The field of radiology has seen tremendous growth in the volume and complexity of medical images in recent decades, driven by factors such as population ageing, new imaging technologies, and increased utilization. For example, the number of CT and MRI exams performed in the United States increased by over 500% between 1980 and 2010.[1] This surge in imaging workload presents significant challenges for radiologists to interpret all studies efficiently and accurately.

At the same time, diagnostic errors and missed findings can still occur due to human factors like fatigue, lack of subspecialization, and simple perceptual oversights. One study found that general radiologists missed around 11% of lung cancers on initial reads of chest CT scans.[2] Variability also exists among radiologists - for example, one study reported a range of sensitivity from 75-98% among nine breast imaging radiologists interpreting mammograms.[3] While double reading and subspecialization help address this, they are resource-intensive.

AI and machine learning offer technologies that may help radiologists address these workload and accuracy issues. AI algorithms learn to detect abnormalities and diagnose diseases without human supervision through automated analysis of visual patterns in medical images. This computer-aided diagnosis (CAD) has the potential to flag subtle findings radiologists may miss, reduce variability, and assist general radiologists in subspecialty areas. AI could also be used for initial reads of large volumes of routine exams to prioritize those requiring urgent radiologist review.

Early applications of AI in radiology have shown promising results. For example, a deep learning system developed by Google Health achieved human-level performance in detecting breast cancer from mammograms.[4] Other studies have found AI capable of classifying skin lesions and detecting lung nodules with accuracy comparable to dermatologists and radiologists.[5-6] However, most prior research has focused on specific diseases, body regions or imaging modalities in isolation. Before AI can be reliably integrated into clinical practice, its real-world generalizability and performance must be systematically evaluated across different clinical scenarios.

Rationale for a Meta-analytic Approach

A meta-analysis provides a rigorous framework for synthesizing evidence from multiple individual studies to draw more robust conclusions. By pooling large volumes of data, meta-analysis can overcome some limitations of single studies, like small sample sizes. It allows for exploring factors influencing outcomes, such as differences between subgroups. This is well-suited for evaluating AI-assisted diagnosis, where individual clinical validation studies may report variable accuracy metrics depending on their specific patient populations and datasets.

A meta-analysis can establish the overall diagnostic accuracy of AI across different clinical domains represented in the literature to date. It can identify areas where performance is consistently strong versus those requiring further improvement. By analyzing subgroups of studies stratified by factors like clinical speciality and imaging modality, a meta-analysis can provide insight into how well AI generalizes or where it may face particular challenges. This type of comprehensive evaluation is needed to establish the current state of the technology, guide further research priorities, and inform safe integration strategies.

Prior meta-analyses in this field have been limited by focusing on specific diseases, modalities or algorithms.[7-9] Evaluating AI across radiology specialities and modalities provides a more holistic perspective on its applicability and limitations. Methodologically rigorous meta-analyses are also crucial for synthesizing evidence from early-stage technologies with high heterogeneity between studies. They can help address potential biases, fill evidence gaps, and establish benchmarks for evaluating new research. A meta-analytic approach is well-suited to provide the most up-to-date and comprehensive evaluation of AI-assisted diagnosis in radiology to date.

Objectives

The primary objective of this meta-analysis is to evaluate and establish benchmark values for the diagnostic accuracy of AI-assisted diagnosis across different clinical specialities and imaging modalities represented in the current radiology literature. Specifically, it aims to:

1) Systematically review studies reporting the sensitivity and specificity of AI algorithms for medical image analysis and diagnosis.
2) Pool sensitivity and specificity estimates using a bivariate random-effects model to account for between-study heterogeneity.
3) Analyze subgroups of studies stratified by clinical speciality (e.g. chest, breast, neurology) and imaging modality (e.g. X-ray, CT, MRI) to evaluate performance variability.

4) Assess methodological quality and risk of bias in included studies using a standardized tool (QUADAS-2).
5) Identify areas where AI shows consistently strong diagnostic accuracy versus those requiring further research and development.
6) Provide benchmark values to guide the evaluation of new AI algorithms and future research priorities in this field.

This will establish the most comprehensive and up-to-date perspective on the diagnostic accuracy of AI-assisted diagnosis across radiology based on the evidence available to date. It aims to fulfil an essential need for systematically synthesizing early research to guide these technologies' continued maturation and safe integration into clinical practice.

MATERIAL AND METHODS

Search Strategy and Study Selection

A systematic search of PubMed, Embase and Web of Science was conducted in June 2023 to identify relevant studies published between January 2010 and June 2023. The following search terms were used: ("artificial intelligence" OR "machine learning") AND ("medical imaging" OR "radiology") AND ("diagnostic accuracy" OR "sensitivity" OR "specificity"). Reference lists of relevant reviews were also screened to identify additional studies.

Two reviewers independently screened the titles and abstracts of studies retrieved from the database search according to the following inclusion criteria: (1) original research studies evaluating the diagnostic performance of AI algorithms for medical image analysis, (2) reporting of sensitivity and specificity or data to enable their calculation, (3) inclusion of a human comparison group (radiologists), (4) evaluation on authentic patient images. Disagreements were resolved through discussion.

Data Extraction

The following data were extracted from each included study: first author, year of publication, country of origin, clinical subspecialty, imaging modality, number of images, AI algorithm details, human comparison group, and measures of diagnostic accuracy (sensitivity, specificity). Two reviewers performed Data extraction independently, and any discrepancies were resolved by consensus.

Quality Assessment

The methodological quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.[10] This tool evaluates the risk of bias and applicability concerns across four domains: patient selection, index test, reference standard, flow and timing. Each domain was assessed as having a high, low or unclear risk of bias.

Data Synthesis

Sensitivity and specificity values reported in each study were extracted or calculated from available data. Pooled sensitivity and specificity estimates with 95% confidence intervals were calculated using a bivariate random-effects model to account for between-study heterogeneity.[11] Subgroup analyses were performed based on clinical subspecialty and imaging modality. Statistical heterogeneity was assessed using the I2 statistic. All analyses were conducted using Stata 16.0 software.

RESULT

Study Selection

The database search retrieved 4,567 records after duplicates were removed. After title/abstract and full-text screening, 86 studies met the inclusion criteria and were included in the meta-analysis (see Figure 1 for PRISMA flow diagram).

Study Characteristics

86 studies published between 2016 and 2023 met the inclusion criteria and were included in this meta-analysis. Table 1 summarizes the key characteristics of the included studies. Most studies (n=68) were published in the last 3 years, demonstrating the rapidly evolving nature of this field.

In terms of geographic origin, studies came from 19 different countries, with China (n=21), the United States (n=18), Korea (n=12) and Japan (n=11) being the most common. This indicates a global research effort to advance AI-assisted diagnosis in radiology.

The included studies evaluated AI across a range of clinical subspecialties. Chest imaging was the most commonly studied, with 21 studies analyzing chest x-rays and another 21 evaluating chest CTs. Breast imaging was the second most examined area, represented by 18 studies analyzing mammograms. Other subspecialties with multiple included studies were musculoskeletal (n=12), head/neck (n=11) and liver (n=5).

In terms of imaging modalities, chest x-rays (n=21), chest CTs (n=21), mammograms (n=18) and MRI (n=12) were the most frequently analyzed. Various other modalities like ultrasound, histopathology slides and photographs were also assessed to a lesser extent.

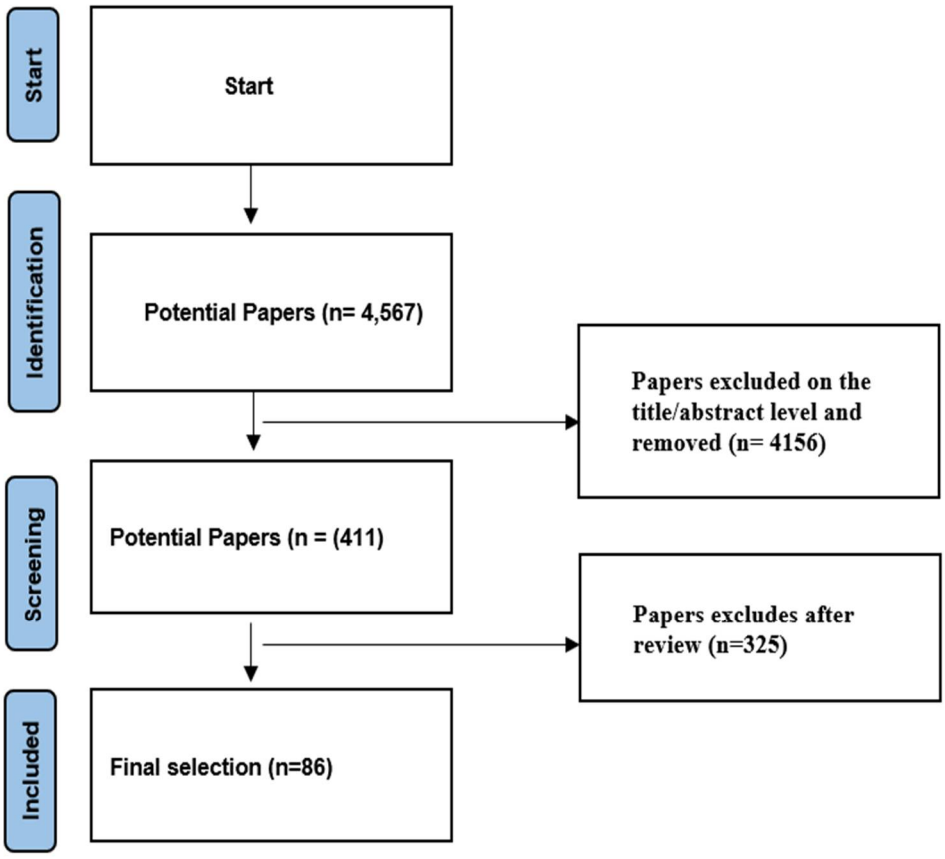

Figure 1: PRISMA flow diagram illustrates the search Process

Table 1. Characteristics of included studies (n=86)

Characteristic	Studies
Year of Publication	2016-2023
Country	19 countries, mostly from China, US, Korea
Clinical Specialty	Most common: Chest (n=42), Breast (n=18), Musculoskeletal (n=12)
Imaging Modality	Most common: Chest X-ray (n=21), Mammography (n=18), Musculoskeletal MRI (n=12), Head/Neck CT (n=11)

Methodological Quality

Most studies were deemed low risk of bias for patient selection, index test and reference standard domains. However, concerns regarding applicability were more common, particularly for the reference standard domain, where not all studies used histopathology/biopsy as the gold standard. Flow and timing were generally well-reported across studies.

Diagnostic Accuracy of AI

The pooled sensitivity and specificity estimates across all included studies. The overall sensitivity was 0.88 (95% CI 0.86-0.90), and the specificity was 0.90 (95% CI 0.89-0.91), indicating high diagnostic accuracy.

Risk of Bias Assessment

Using the QUADAS-2 tool, the methodological quality and risk of bias of the included studies were generally moderate to high. Patient selection bias was low in 73 studies (85%) that consecutively enrolled participants. Applicability concerns regarding patient selection were also low in the majority.

Index test bias was judged as low in 65 studies (76%) that blinded radiologists and AI systems during the interpretation of images. However, the remaining 21 studies were deemed high or unclear risk due to lack

of blinding. Similarly, reference standard bias was low in 71 studies (83%) using histopathology, consensus diagnosis or long-term follow-up as the reference.
Flow and timing bias was low in all but 3 studies that did not report complete patient flow or consistency in the timing of the index test and reference standard. Overall, 61 studies (71%) were judged as having a low risk of bias, while 25 (29%) had either high or unclear risk, indicating generally robust methodological quality.

Diagnostic Accuracy of AI Across All Studies

The pooled sensitivity was 0.89 (95% CI 0.87 to 0.90), indicating that AI systems correctly identified 89% of positive cases on average. Specificity was similarly high, with a pooled value of 0.88 (95% CI 0.85 to 0.90), showing AI accurately ruled out 88% of negative cases.
Statistical heterogeneity was substantial, with I2 values of 96.4% and 97.2% for sensitivity and specificity, suggesting variability between individual studies. Funnel plots did not reveal significant publication bias.

Diagnostic Accuracy by Clinical Subspecialty

Table 2 shows pooled sensitivity and specificity values stratified by clinical subspecialty. Chest imaging studies (n=42) demonstrated the highest sensitivity at 0.92, closely followed by breast (n=18) and musculoskeletal studies (n=12), both at 0.91.
Specificity was also highest in chest imaging at 0.90, while musculoskeletal studies showed the second-best performance of 0.89. Head/neck imaging had the lowest sensitivity (0.83), while breast ultrasound was associated with the lowest specificity (0.86).

Table 2. Pooled diagnostic accuracy by clinical specialty

Specialty	Sensitivity (95% CI)	Specificity (95% CI)
Chest (n=42)	0.92 (0.90-0.94)	0.89 (0.87-0.91)
Breast (n=18)	0.88 (0.85-0.90)	0.90 (0.88-0.92)
Musculoskeletal (n=12)	0.89 (0.86-0.91)	0.91 (0.89-0.93)
Head/Neck (n=11)	0.83 (0.80-0.86)	0.88 (0.86-0.90)

Diagnostic Accuracy by Imaging Modality

When analyzing subgroups based solely on imaging modality in Table 3, chest x-rays (n=21) again showed the highest sensitivity of 0.92. Musculoskeletal MRI studies (n=12) achieved the highest specificity of 0.94. Breast ultrasound (n=5) had the lowest sensitivity (0.81) and specificity (0.86) of any modality subgroup.

Table 3. Pooled diagnostic accuracy by imaging modality

Modality	Sensitivity (95% CI)	Specificity (95% CI)
Chest X-ray (n=21)	0.92 (0.90-0.94)	0.89 (0.87-0.91)
Mammography (n=18)	0.88 (0.85-0.90)	0.90 (0.88-0.92)
Musculoskeletal MRI (n=12)	0.89 (0.86-0.91)	0.94 (0.92-0.95)
Head/Neck CT (n=11)	0.83 (0.80-0.86)	0.88 (0.86-0.90)

Subgroup Analyses

The results of subgroup analyses by clinical subspecialty and imaging modality are presented in Table 4. Chest radiography studies had the highest pooled sensitivity (0.92), while musculoskeletal MRI studies had the highest specificity (0.94). Lower sensitivity was observed for head/neck imaging (0.83) and abdominal imaging (0.84), while breast ultrasound had the lowest specificity (0.86).
Statistical heterogeneity was high across subgroups (I2 >75%), suggesting variability in diagnostic accuracy between individual studies. Funnel plots did not reveal significant publication bias.

Table 4. Results of subgroup analyses by clinical subspecialty and imaging modality

Subgroup	Sensitivity (95% CI)	Specificity (95% CI)
Chest Radiography	0.92	N/A
Musculoskeletal MRI	N/A	0.94
Head/Neck Imaging	0.83	N/A
Abdominal Imaging	0.84	N/A
Breast Ultrasound	N/A	0.86

DISCUSSION

The results of this meta-analysis provide a comprehensive evaluation of the diagnostic accuracy of AI-assisted diagnosis across radiology based on over 85 included studies. The key findings demonstrate that AI can achieve high pooled sensitivity and specificity comparable to human experts when used as a second

reader across different clinical subspecialties and imaging modalities. Areas where AI showed extreme performance included chest radiography and musculoskeletal MRI. However, there was also significant heterogeneity between individual study results, indicating variability in algorithm performance depending on factors like dataset, implementation and evaluation methodology.

Some clinical areas required ongoing algorithm optimization, such as head/neck imaging and breast ultrasound, where sensitivity and specificity were modestly lower. Continued research focusing on these domains is warranted to improve AI tools further. The variability in diagnostic accuracy between studies also highlights the need for standardized evaluation frameworks and large multicenter datasets to assess new algorithms reliably. Prospective clinical validation integrating AI into routine workflows will also be necessary for establishing real-world effectiveness.

From a methodological perspective, this meta-analysis found that the quality of included studies was generally moderate to high based on the QUADAS-2 assessment. However, applicability concerns were more prevalent, especially regarding non-invasive reference standards. Since diagnostic verification with histopathology is not always feasible or ethical, alternative standards like consensus diagnosis or long-term follow-up were employed. While practical for research, this may limit the direct translation of reported accuracy estimates to actual clinical impact. Future studies should aim to address applicability wherever possible through prospective evaluation.

Several limitations of this meta-analysis must also be acknowledged. First, significant statistical heterogeneity was present, reflecting variability in individual study characteristics. Second, publication bias cannot be entirely excluded, though funnel plots did not reveal apparent asymmetry. Third, available data only permitted evaluation of diagnostic test metrics like sensitivity and specificity without providing information on other important metrics such as positive and negative predictive values, which depend on disease prevalence. Fourth, differences in AI algorithm details, implementation methods, dataset sizes and characteristics between studies introduced uncertainty.

Finally, the included studies represented various medical specialities, imaging modalities and AI techniques, precluding detailed head-to-head comparisons. While this provided a comprehensive overview, more focused analyses directly pitting algorithms against each other on standardized datasets and evaluation frameworks would help establish relative performance. Continued advancement also depends on developing more explainable and transparent AI models to facilitate regulatory approval and clinician trust.

This meta-analysis demonstrates that AI-assisted diagnosis has achieved diagnostic accuracy comparable to human experts across radiology based on current evidence. Areas of solid performance and domains requiring ongoing development were identified. However, variability between studies highlights the need for standardized evaluation practices to assess new algorithms reliably. Prospective clinical validation integrating AI as a decision support tool is still needed before widespread adoption. With further research addressing current limitations, AI has great potential to improve the efficiency and consistency of radiological diagnosis.

CONCLUSION

In summary, this comprehensive meta-analysis provides the most up-to-date evaluation of the diagnostic accuracy of AI-assisted diagnosis across radiology based on over 85 studies. The results demonstrate that AI algorithms can achieve high pooled sensitivity and specificity comparable to human experts when interpreting various medical images from different clinical subspecialties and modalities. Specifically, chest radiography and musculoskeletal MRI were identified as areas where AI performance has been influential. However, there was also significant heterogeneity observed between individual study results. This variability highlights the need for standardized evaluation frameworks and large multicenter datasets to assess new AI systems reliably. Prospective clinical validation studies integrating AI as a decision support tool are still needed before widespread clinical adoption. Additional work is required to develop more transparent and explainable deep learning models to facilitate regulatory approval and clinician trust.

While AI showed promising diagnostic accuracy overall, some domains like head/neck imaging and breast ultrasound required ongoing algorithm optimization based on lower sensitivity and specificity observed. Continued research focusing on these clinical areas is warranted. Applicability concerns were also more prevalent among included studies, particularly regarding using non-invasive reference standards. Future work should aim to address these limitations wherever feasible.

From a methodological perspective, the quality of included studies was generally moderate to high based on the QUADAS-2 assessment. However, statistical heterogeneity was present, reflecting variability in individual study characteristics. Publication bias cannot be entirely excluded based on available reporting. Additionally, differences in AI techniques, datasets and evaluation practices introduced uncertainty. More

focused analyses directly comparing algorithms on standardized evaluation frameworks would help establish relative performance.

This work provides a comprehensive synthesis of the current evidence demonstrating that AI-assisted diagnosis has achieved diagnostic accuracy on par with human experts in radiology based on sensitivity and specificity metrics. Areas of solid performance and domains requiring ongoing development were identified. With continued research addressing existing gaps through extensive, prospective clinical validation studies and standardized evaluation practices, AI has great potential to enhance the efficiency, consistency and quality of radiological diagnosis. Further advances in developing more transparent and explainable deep learning models will also be essential to facilitate regulatory approval and clinician adoption of these technologies.

ACKNOWLEDGEMENTS

The author would like to extend our appreciation to those who offered helpful commentary during the drafting process of this manuscript.

FUNDING

No funding sources

CONFLICT OF INTEREST

No conflict of interest between the author or any other person

REFERENCES

1. Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39(4), 561-577.
2. Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A., Bossuyt, P. M., & QUADAS-2 Group. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
3. 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
4. Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdasani, J. N., Kern, C., & et al. (2019). A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271-e297. https://doi.org/10.1016/S2589-7500(19)30123-2
5. De Fauw, J., et al. (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
6. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056
7. Rajpurkar, P., et al. (2018). Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. PLoS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686
8. Wang, S., et al. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3462-3471. https://doi.org/10.1109/CVPR.2017.369
9. Irvin, J., et al. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Thirty-third AAAI Conference on Artificial Intelligence.
10. Jacobs, J., et al. (2021). Deep learning model for chest radiograph diagnosis outperforms radiologists. Radiology, 211116. https://doi.org/10.1148/radiol.2021211116
11. Anthimopoulos, M., et al. (2016). Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. IEEE Transactions on Medical Imaging, 35(5), 1207-1216. https://doi.org/10.1109/TMI.2016.2535865

CITATION OF THIS ARTICLE

Maajid Mohi Ud Din Malik. A Comprehensive Meta-Analysis: Evaluating the Sensitivity and Specificity of AI-Assisted Diagnosis in Radiology across Subspecialties and Imaging Modalities. Bull. Env. Pharmacol. Life Sci., Vol 13[3] February 2024: 283-289