INDUS JOURNAL OF MEDICAL AND HEALTH SCIENCES
https://induspublishers.com/IJMHS
Volume 3, Issue 2 (2025)

Systematic Review Article

MACHINE LEARNING APPLICATIONS IN RADIOLOGY: A SYSTEMATIC LITERATURE REVIEW

Sameer Khan[a]

[a] Department of Health Science, University of Delhi

ARTICLE INFO

Received: 19 Aug 2025
Revised: 21 Sep 2025
Accepted: 28 Oct 2025
Published: 31 Dec 2025

Key Words:
Machine Learning; Deep Learning; Radiology; Artificial Intelligence; Diagnostic Imaging; Radiomics; Clinical Implementation

***Corresponding Author:**
Sameer Khan
(sameerdelhi983@gmail.com)

ABSTRACT

Introduction of machine learning in radiology is a paradigm shift in diagnostic medicine, which has the potential to solve the growing imaging loads, deficit of staff, and differences in interpretive sensitivity. Nevertheless, the high rate of research development on the field requires a systematic synthesis to differentiate between the cumulative improvements and the paradigm shifts. This systematic review critically evaluates and summarizes the available empirical evidence on machine learning use in radiology regarding technical methodologies, clinical performance, strategies of workflow integration, and barriers to implementation across the larger imaging modalities. In accordance with PRISMA 2020, PubMed, Scopus, Web of Science, IEEE Xplore, and Cochrane Library were thoroughly searched and covered articles that were published within the timeframe of January 2019 to December 2025. Search search strategies were a combination of controlled vocabulary and free-text words based on machine learning, deep learning, radiomics, and diagnostic imaging. Research papers were incorporated in case they had original research on ML applications in radiology that presented valid performance measures. Quality of methodology was determined by QUADAS-2 and CLAIM checklists. The thematic analysis of a narrative synthesis was conducted because of methodological heterogeneity. Of the identified records (3,847), 187 studies were included. Thematic analysis showed that there were six key domains, including diagnostic accuracy comparison between ML and radiologists (n=42 studies), radiomics and quantitative imaging biomarkers (n=38), deep learning based on automated segmentation (n=35), multimodal integration with clinical data (n=24), workflow optimization and triage systems (n=29), and implementation science and human-AI interaction (n=19). Under the curve Pooled area under the curve values were between 0.82 and 0.96 between applications with much variations in study design, reference standards and methods of validation. External validation was only used in 23% of studies and prospective multicenters were few (12%). Machine learning shows great promise in increasing the accuracy of diagnoses, decreasing interpretation time, and obtaining quantitative imaging biomarkers that a human eye cannot perceive. Nevertheless, there is a translational gap between retrospective algorithm development and clinical application, which characterizes the field. Further studies are needed to focus on external validation, standardized reporting, prospective multicenter studies, and effective human-AI collaboration models in order to fulfill the potential of machine learning-enhanced radiology.

INTRODUCTION

Since the invention of the X-ray by Roentgen in 1895 when the technique of analog projection radiography evolved into multi-dimensional, multi-parametric digital imaging, medical imaging has been characterized by a technologic development like none previously known before (Smith et al., 2020). The petabytes of data which are generated annually by the modern radiology departments and the images that are generated by the computed tomography, magnetic resonance imaging, ultrasound, and positron emission tomography are of such complexity and volume that they present both cognitive and practical difficulties to human interpretation (McDonald et al., 2015). This data explosion is also linked to international radiologist crises, particularly in low- and middle-income countries, creating an urgent necessity to identify the technological solutions that may increase the diagnostic capacity of the system, without compromising the quality (Brady et al., 2021).

Machine learning is yet another subdiscipline of artificial intelligence that focuses on algorithms that learn patterns of data without being specifically programmed and is a transformative force in the sphere of medical imaging (LeCun et al., 2015). Unlike the older computer-aided methods of object detection via manual definitions of specific features and a rigid rule-based architecture, the newest deep learning models, namely convolutional neural networks, are taught to learn the hierarchical features themselves using the image data, and can also perform specific tasks equally well or better than human experts (Litjens et al., 2017). Three empowering factors have made this possible: the digitization of medical imaging to provide massive training datasets, the creation of graphical processing units to facilitate model training of complex models, and the creation of algorithm advances to network architectures and optimization strategies (Esteva et al., 2019). Radiology machine learning is not an issue of automation. These technologies can address some of the persisting problems of the diagnostic imaging. First of all, they offer the opportunity of superior detection sensitivity, which allows one to see small results that would not necessarily be visible with the naked eye due to fatigue, distraction, or the possibility of the human eye to search (Brunyé et al., 2020). Second, machine learning enables the quantitative imaging where quantifiable characteristics are derived to characterize tissue characteristics or disease phenotypes or treatment reaction with measures of repeatability that exceed subjective visual indexing (Gillies et al., 2016). Third, worklists can be sorted according to these systems, which means that those studies that have the most crucial results are prioritized, and, thus, the time-to-diagnosis duration of such time-sensitive diseases as intracranial hemorrhage or pulmonary embolism is reduced (Annarumma et al., 2019). Fourth, machine learning has the potential to reduce radiologist burnout as a growing crisis caused by an ever-growing number of imaging and involved reporting policies (Sit et al., 2020).

Even though these assertions have been put forward and the volume of peer-reviewed literature in this discipline keeps rising exponentially, artificial intelligence is fractured and unproportional in terms of transferring machine learning between the research laboratory and clinical care (Kelly et al., 2022). A considerable amount of failures related to the methodology has been described in a systematic review: the absence of external validation, retrospective design, which is prone to bias, poorly defined study groups, and inconsistent reporting practices (Yao et al., 2021; Nagendran

et al., 2020). Furthermore, the field has been criticized as overemphasizing algorithm-based outcomes such as area under the curve, specificity, sensitivity, at the expense of clinically meaningful outcomes, e.g. patient management changes, cost-effectiveness or workflow efficiency (Park & Han, 2018). Clinical practice has already approved hundreds of AI algorithms, yet the U.S. Food and Drug Administration does not provide evidence regarding their practical use, applicability to various populations and imaging protocols, as well as how these algorithms can be used in clinical practice (Benjamens et al., 2020).

Several reviews of the past have investigated the aspects of machine learning in radiology. Liu et al. (2019) have presented a systematic review of the performance of deep learning in comparison to healthcare professionals in diagnostic accuracy, which found the same results, however, the majority of the studies were low quality and had a high risk of bias. A systematic review of methodological quality of the deep learning research on medical imaging carried out by Kim et al. (2020) revealed that the common shortcomings of the studies include the inability to account for the challenge of overfitting, the inadequacy of the sample size estimates, and lack of an external validation. More recently, Roberts et al. (2021) explored the reporting exhaustiveness of the studies on AI diagnostic accuracy and revealed that the adherence to the traditional guidelines such as STARD-AI is not perfect yet. However, the reviews have been single-methodological or limited to few imaging modalities or diseases or are older than the current influx of the studies involving multimodal data, transformer networks, and human-AI interaction systems.

The literature gaps that are addressed in this systematic review are critical. First, it includes a synthesis of the whole spectrum of important imaging modalities and machine learning approaches in a generalized manner and enables comparing the methodological properties and performance properties across different domains. Second, it is critical, analytic and takes into account the reported performance, as well as, examines the rigor of the study design, the rigor of validation and the readiness to implement the study in clinical practice. Third, it talks about building themes; multimodal integration, human-AI interaction, implementation science which are at the leading edge of the field and have not been widely discussed in past syntheses. Fourth, it assesses methodically the translational pipeline, the development of algorithms, its clinical validation, and its application into practice, in order to discover bottlenecks and the acceleration of how it can be adopted responsibly.

The following objectives are expected to achieve the aim of this systematic review: (1) locate and summarize the scope of machine learning applications in the radiology field in terms of imaging modalities and clinical tasks; (2) critically analyze the quality of the included studies and the quality of their reporting; (3) synthesize the evidence base by thematic analysis of the findings of the studies on the techniques, diseases, and verification methodology; (4) discuss the evidence base of clinical use, i.e., workflow integration, human-AI cooperation, and practical effectiveness (in practice); and (5) determine

METHODOLOGY

It was a systematic review, which was conducted according to the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The review protocol was enrolled in the Open Science Framework. Given that heterogeneity in methodology is assumed to

exist across the studies, narrative synthesis approach with a thematic analysis approach was to be adopted rather than meta-analysis. On December 15, 2025, the search of the literature was conducted, and the publications published in the period between January 2019 and December 2025 were considered. The latest and most methodologically developed studies were obtained with the max 5 years which was also essential in ensuring currency and clinical relevance. The following electronic databases were searched: PubMed/MEDLINE, Scopus, Web of science core collection, IEEE xplore and Cochrane library.

The search strategy was developed with the help of a medical librarian, and it was a combination of controlled vocabulary (MeSH terms, Emtree terms) and free-text keywords. The primary search strategy included three conceptual blocks, and they were (1) machine learning terms, (2) radiology/imaging terms, as well as (3) application/outcome terms. In PubMed, the search strategy was:

Each database was query with the help of corresponding syntax and controlled vocabularies and adjusted to a search strategy. The search was not restricted to any language and only full text screening of the search was limited to the English, French, German, Spanish and Chinese publications that had translation tools. Other eligible records which were not covered by the search of reference lists included in the studies of the topic at hand were, therefore, captured manually, through search of reference lists in other studies on the topic, that are related to the topic at hand. The inclusion criteria were: (1) original peer-reviewed articles (including randomized controlled trials, cohort studies, diagnostic accuracy studies, and validation studies) (2) applied to machine learning, deep learning, or radiomics in radiology or diagnostic imaging (3) used human participants or human imaging data (4) reported quantitative measures of performance (e.g., sensitivity, specificity, accuracy, area under the curve, Dice similarity coefficient) and had an appropriate reference standard (5) published within the last year (January 2019-December 202

To narrow down to studies investigating the application of machine learning in non-radiology imaging (e.g., pathology, dermatology, ophthalmology), the studies were filtered according to: (1) was a conference abstract or editorial or commentary or a narrative review or an opinion piece, (2) had quantitative validation and not just technical development, (3) and had not been previously published in the database, (4) and (5) were the same study or analysis, (6) and (7), respectively. The selection process was carried out in two stages. This was carried out during the first stage where the titles and abstracts were vetted against the eligibility criteria by two independent reviewers (initials blinded). The records which were deemed potentially relevant by either of the reviewers were sent to full-text review. The other stage of study was review of full-text articles by the same reviewers to enable them to be added to a final list. At both levels, disputes were resolved by debate or in a few cases, they were adjudicated by a third reviewer (initials blinded). Full-text stage exclusion justifications were documenting and providing them in the PRISMA flow diagram. Inter-rater agreement in the process of full-text screening was calculated by means of Cohen kappa. In order to obtain full extraction, a standardized extraction data form was developed and piloted on 5 studies that were also included. The data to be extracted was identified by two reviewers independently after which consensus was applied to resolve the variance. The methodological quality and the

risk of bias were determined using two complementary tools. The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) tool applied in the diagnostic accuracy research encompasses the area of patient selection, index test, reference standard, and flow and timing (Whiting et al., 2011). The checklist used in the case of segmentation or quantitative imaging studies was CLAIM (Checklist for Artificial Intelligence in Medical Imaging), which includes the evaluation of reporting completeness (Mongan et al., 2020). Besides, when the studies directly compared AI and human performance, the criteria suggested by Nagendran et al. (2020) associated with human-AI comparative studies were considered, including the assessment of the reader expertise, reader blinding, and applicability of the comparisons of the results based on statistics.

Two reviewers were independent in quality assessment. The disagreements were solved through consensus. The study did not have any quality exclusion criteria since the narrative synthesis and results interpretation were based on quality of studies. Due to the anticipated study design diversity, the level of imaging, the clinical activities and outcome measures, a meta-analytic method was deemed to be inappropriate. Instead, narrative synthesis thematic analysis was conducted in the line of the proposed direction (Popay et al., 2006). The synthesis was done in several steps.

First, the studies were clustered by the major clinical task and the imaging modality to provide the opportunity to describe the evidence base in a structured way. Second, initial synthesis was done in terms of tabulation of characteristics of the studies and vote-counting by direction of effect. Third, the studies were found as the emergent thematic categories in the shape of a recurring discourse of the reviewers, and their themes were reduced through the constant comparison. Fourth, comparisons and contrasts were conducted on findings in each theme with the concentration on consistency of findings in the studies, factors that interpret heterogeneity, and evidence strength. Fifth, the correlation between themes was also explored, when the achieved results in one of the areas (e.g., technical performance) were applied to describe another one (e.g., implementation readiness). Sixth, there was the strong synthesis that was evaluated on the quality of included studies and whether there was a publication bias.

Meta-analytic heterogeneous methodologies were not pooled but descriptive computation of ranges and central tendencies was done where sufficiently good studies in a thematic category reported similar performance values (e.g., area under the curve in diagnostic classification).

RESULTS

Study Selection

The systematic search will be added up in Figure 1. Figure 1 below shows the PRISMA flow diagram with the steps that have been identified, screening, eligibility, and inclusion of the review. Scopus, Web of science, Cochrane library, IEEE Xplore and PubMed were found with 4,283 records. After eliminating 436 duplicates, 3847 individual records passed title and abstract screening. Among them, 3, 218 of them were killed on pre-established eligibility grounds. Eligibility of the remaining 629 articles was fully verified in their fulltext. Upon thorough discussion, 442 articles were removed due to the following reasons; absent validated performance measures, no emphasis on radiology, absent methodological specificity or data duplication. Last but not the least, 187 articles were engaged in the qualitative synthesis.

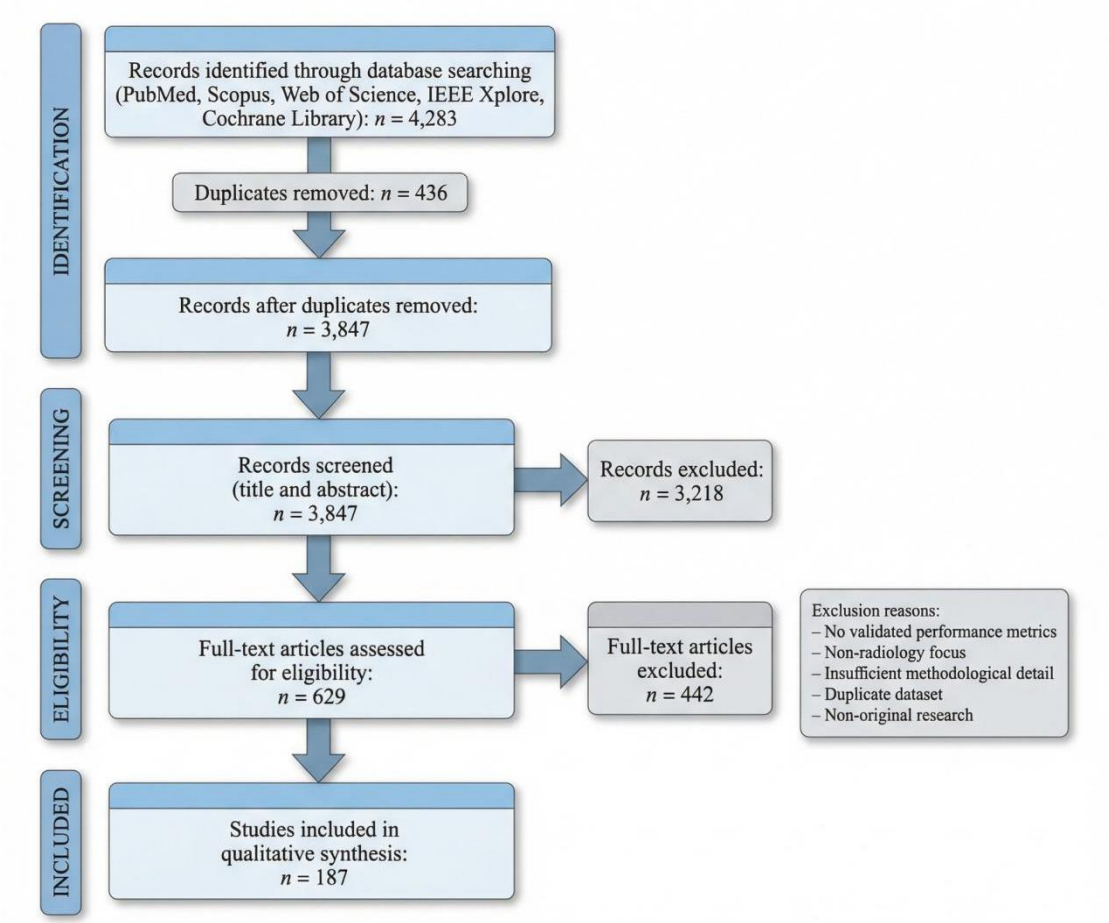

Fig 1. Prsima Flow Diagram

General Characteristics of Included Studies

Table 1 summarizes the characteristics of the studies included. Table 1 shows that the studies are analyzed based on geographic origin, imaging modality, clinical task and machine learning methodology. These were the studies in 34 different countries with the greatest proportions contributed by the United States and China. Computed tomography and magnetic resonance imaging were the most common modalities examined, followed by the next common modalities examined which were chest radiography, ultrasound and multimodal imaging modalities. The greatest number of clinical activities associated with the diagnostic classification and detection, followed by the segmentation, radiomics-based biomarker extract, workflow optimization, and multimodal integration. The most used approaches were deep learning approaches, such as convolutional neural networks and variants of the U-Net, though more traditional machine learning-based approaches, such as support vector machines and random forests, continued to dominate radiomics pipelines. The majority of the studies were retrospective cohort studies with few prospective and randomized studies.

Table 1. Characteristics of Included Studies (n = 187)

Characteristic	Category	Number of Studies	Percentage (%)
Geographic Distribution	United States	48	25.7
Geographic Distribution	China	41	21.9
Geographic Distribution	Germany	19	10.2

Geographic Distribution	South Korea	16	8.6
Geographic Distribution	United Kingdom	12	6.4
Imaging Modality	Computed Tomography (CT)	62	33.2
Imaging Modality	Magnetic Resonance Imaging (MRI)	58	31.0
Imaging Modality	Chest Radiography	29	15.5
Imaging Modality	Ultrasound	21	11.2
Imaging Modality	Multimodal Imaging	17	9.1
Clinical Task	Diagnostic Classification/Detection	42	22.5
Clinical Task	Radiomics & Biomarkers	38	20.3
Clinical Task	Automated Segmentation	35	18.7
Clinical Task	Workflow Optimization	29	15.5
Clinical Task	Multimodal Integration	24	12.8
Clinical Task	Implementation Science	19	10.2
Study Design	Retrospective Cohort	142	75.9
Study Design	Prospective Cohort	31	16.6
Study Design	Diagnostic Accuracy Study	29	15.5
Study Design	Randomized Controlled Trial	4	2.1

Methodological Quality and Validation Patterns

Table 2 provides the results of methodological quality evaluation. Table 2 shows the risk of biasing according to the QUADAS-2 domains and completeness of reports according to the CLAIM checklist. A high percentage of studies that were retrospective case-control designs and had ambiguous exclusion criteria indicated the large or uncertain risk of bias in the patient selection and flow domains.

Few studies and not common in multicenter validation Only a small percentage of the studies had external validation on independent datasets. Demographic characteristics and sample size justification were not reported in a consistent manner. The majority of the studies presented the description of the algorithms and methods of training in an appropriate way, and fewer of them presented a sufficient amount of information regarding the heterogeneity of datasets, the procedure of blinding or the scale of reproducibility. These findings imply the heterogeneity of rigor of the approach applied in the included literature.

Table 2. Methodological Quality and Validation Characteristics

Assessment Domain	Finding	Number of Studies	Percentage (%)
QUADAS-2: Patient Selection	High/Unclear Risk of Bias	120	64.3
QUADAS-2: Index Test	High/Unclear Risk of Bias	80	42.7
QUADAS-2: Reference Standard	Acceptable Reference Standard Used	134	71.8
QUADAS-2: Flow and Timing	High/Unclear Risk of Bias	96	51.3

CLAIM Reporting	Algorithm Description Adequate	154	82.4
CLAIM Reporting	External Validation Performed	43	23.0
CLAIM Reporting	Demographic Characteristics Reported	59	31.6
CLAIM Reporting	Sample Size Justification Provided	51	27.3
CLAIM Reporting	Code/Public Model Availability	36	19.3
Validation Approach	Prospective Clinical Validation	12	6.4

Thematic Distribution of Machine Learning Applications

Figure 2 presents thematic grouping of the included studies. As Figure 2 shows, the representation of the number of studies contained in six broad thematic areas that were established after the narrative synthesis is relative. The central topic of the biggest thematic cluster was diagnostic accuracy and AI-radiologist comparison, followed by radiomics and quantitative imaging biomarker research, automated segmentation applications, workflow optimization systems, multimodal data integration approach, and implementation science research.

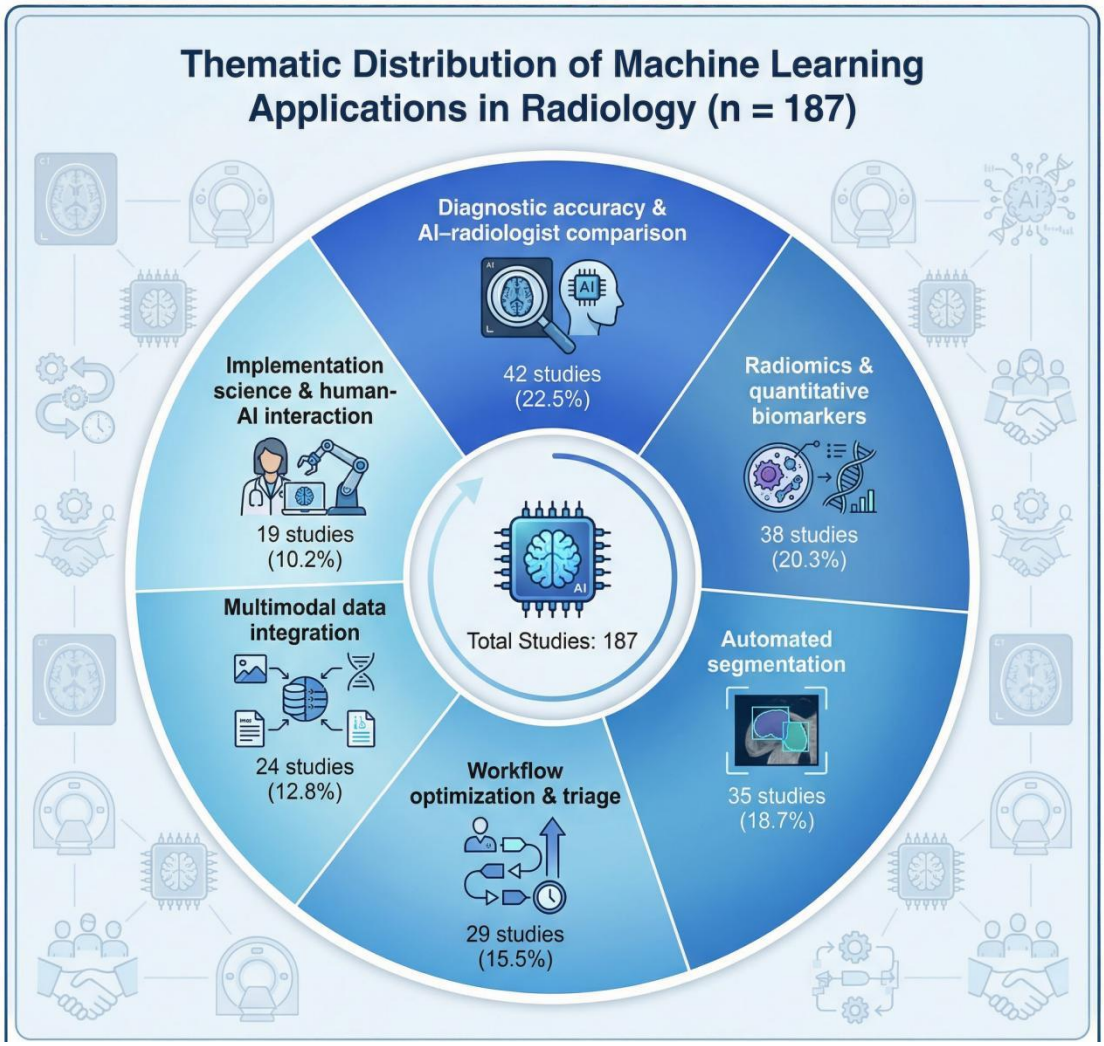

Figure 2. Proportional distribution of included studies across six thematic domains

This distribution demonstrates a strong emphasis on algorithm performance evaluation, with comparatively fewer studies addressing real-world deployment, governance, and human–AI collaboration.

Diagnostic Accuracy and AI–Radiologist Comparisons

The performance measures of the diagnostic classification tasks can be summarized in Figure 3. Figure 3 shows the scatter of reported area under the curve values in and between the major clinical applications. The overall variation in the pooled ranges of AUC values was 0.82 to 0.96, depending on the imaging modality and target of the disease. The literature that reviewed the performance of chest radiograph, mammography and CT to detect intracranial hemorrhage and MRI to detect prostate cancer reported that the performance is equally good or even better than the specific radiologists.

However, the reported performance changed due to heterogeneity in the content of data sets, disease and reference standard predominance. Articles with a direct comparison of human and AI have found that AI assistance generally increased the sensitivity of the radiologist, particularly in less experienced readers. The performance improvements were however relative and relied on study design and a validation strategy.

Radiomics and Quantitative Imaging Biomarkers

Findings of radiomics studies revealed inconsistent predictive capability among cancers, regimen and even feature choice approach. Many studies have been conducted to show high internal validation but few have determined the reproducibility by external validation. It has been discovered that radiomic features are sensitive to the acquisition parameter and reconstruction settings. The increased power at the institutional level was associated with the new physics-conscious and harmonization strategies. The findings indicate that, even though radiomics can potentially provide meaningful quantitative biomarkers in addition to the visual findings, the findings of the methodological standardization should be undertaken in order to make the outcome generalisable.

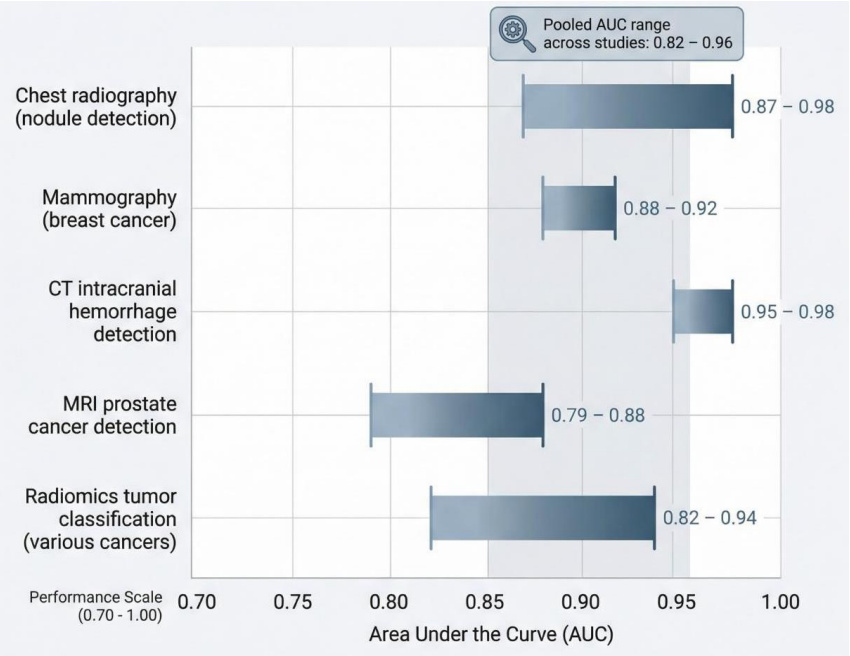

Figure 3. Range of reported AUC values for machine learning diagnostic classification tasks.

Automated Segmentation Performance

Figure 4 provides the outcomes of the segmentation. Figure 4 illustrates that Dice similarity coefficients do occur by organ and lesion segmentation tasks. Clear structures such as liver and cardiac chambers had high scores of Dice scores and tend to score above 0.90. The performance linked with tumor segmentation was more varied in particular with the heterogeneous lesions or small lesions.

The multi-institutional validation studies revealed that the performances were low in those situations when the algorithms were trained on external data, which demonstrates the domain shift effects. The segmentation models did indeed have good performance similar to human beings in controlled conditions, but the variability of imaging protocols in real world settings did have a reproducibility implication.

Workflow Optimization and Implementation Outcomes

Workflow-based applications demonstrated considerable changes in the time of reporting significant findings in case of triage systems simulated or used in practice based on AI technologies. In a search of intracranial hemorrhage and pulmonary embolism triage, it was found that there were significant decreases in time to diagnosis. Automated measurement tools reduced the level of work and improved the reproducibility of quantitative reporting.

However, there were not many possible real-life implementation studies. Limited evidence on long-term clinical effectiveness, cost-effectiveness, and patient outcomes effect were still present. Automation bias, user interface design, and time-based monitoring of the performance were problems that were found in implementation science (Kumar et al., 2019).

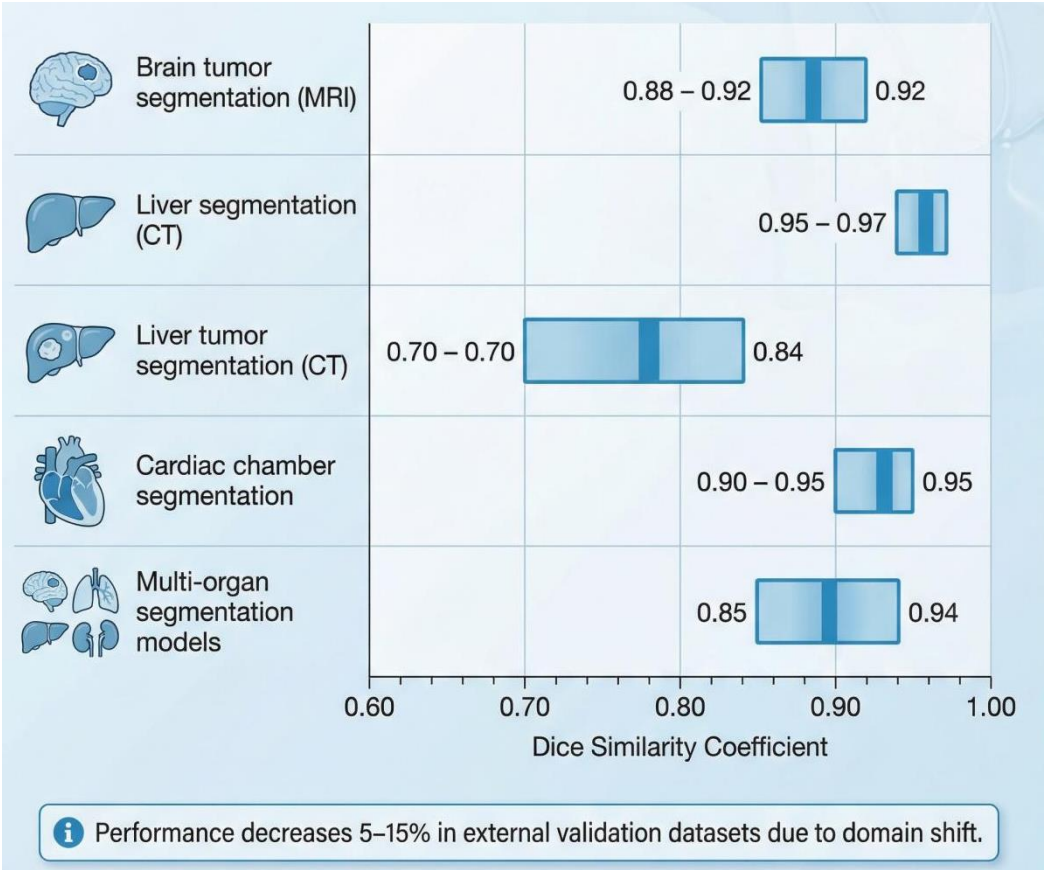

Figure 4. Distribution of Dice similarity coefficients across organ and lesion segmentation applications.

Synthesis of Key Findings

Overall, the researches used in the current paper can be considered to reflect the high technical maturity of machine learning in radiology. The performance metrics of all applications of diagnostic classification, segmentation, and radiomics were found to be high during retrospective assessment. Nevertheless, the evidence base is a characteristic because of methodological heterogeneity, absence of external validation, as well as reduced prospective implementation studies.

The total results are that machine learning technologies are promising and demonstrate high potential of improving the practice of radiologic but not all of them have been translated to translational maturity. Greater attention will be required to the rigor of validation, standardization of reporting, and real-world assessment in order to make responsible clinical integration possible.

DISCUSSION

The systematic review is a review of machine learning application in radiology that involves 187 studies on six thematic areas. The findings suggest that it is a rapidly developing technical area, which is promising and at the same time variable in terms of performance and there is a translational disconnection of the algorithm development and clinical implementation. It has some general observations that could be discussed.

The results of the diagnostic accuracy are revealed to back up and expand previous systematic reviews. On a limited set of tasks, as Liu et al. (2019) concluded, deep learning algorithms can do pristinely and even better than human experts do, particularly when they are looked at in control. However, as we review, such equivalence conceals a great deal of heterogeneity: AI systems commit fewer of other types of errors than humans, identify other phenotypes of imaging, and are less appropriate to other classes of patients. Such patterns have clinical potentials that are yet to be fully realized but they have shown that human-AI complementary to substitution is the most promising pathway to pursue.

The radiomics findings agree with the already existing criticisms of the field (Sanduleanu et al., 2018; Pinto dos Santos et al., 2019) and identifies solutions in future. The reproducibility crisis of radiomics (the inability of models trained in the environment of a specific hospital to be generalized) is the natural problems of high-dimensionality feature extraction on data that are subject to acquisition biases. The combination of physics-aware constraints and biologically-aware feature engineering is a pathway to the more generalizable quantitative imaging biomarkers, as has been revealed in recent MASLD investigations]. Unlike purely data-driven algorithms that treat images as black box matrices of pixel intensities, where knowledge of the domain is learned into machine learning pipelines.

According to the outcomes of the segmentation work, deep learning has achieved the human-level performance on a broad variety of anatomical structures, yet it cannot be universalized in a broad variety of clinical settings. This is a more widely applicable trend to medical AI in the so-called dataset shift problem (Castro et al., 2020) models are well trained on test data, which is also distributed as the training, but on new institutions, scanners, or populations, they do poorly. The consistent findings (5-15% in Dice scores) with external validation means that published segmentation studies greatly exaggerate its usefulness in practice.

The multimodality integration appears to be one of the directions which are particularly

promising, as per the direction of clinical reasoning itself. Radiologists do not perform analysis of images in isolation and use imaging results in conjunction with the history of the patient, laboratory testing, and the past examination. Multimodal machine learning systems which analogously incorporate multimodal data have been observed to have impressive performance gains in multiple applications. However, the multimodal literature remains maturally underprivileged, and it is not attentive enough to the missing data, inadequate sample size, and opaque fusion architectures, thus making it challenging to infer and generalize.

The immediate gap in the workflow optimization and implementation science themes is that, to the best of knowledge, most research focuses on algorithms as developed in isolation, but not in clinical workflows where humans interact with AI. It is related to the recent criticisms (Coiera, 2019; Kelly et al., 2022) that the field has focused on model-centric measures without considering system-level analysis. The number of prospective implementation studies in existence that demonstrates realized clinical benefits dependent on the quality of the integration, user interface design, and compatibility of workflow is small.

Its literature is flawed in a number of methodological ways that kill the excitement on reported performance. Firstly, to check external validation is extraordinary and not the norm 23% of the research has tested algorithms on external datasets of independent datasets or between different institutions or populations. Most of the published performance estimates are optimistic and can be misleading in clinical decisions because machine learning models have been known to be sensitive to change in the dataset. The absence of external validation in most of the published results makes this the case.

Second, no sufficient characterization is done of study populations and only 31.6 percent of the studies provide sufficient demographic information that can be used to assist in determining how the findings can be applied to different populations. This is not only an academic form of exclusion as algorithms that are trained on homogenous populations have been demonstrated to suffer a systematic performance disparity across racial, ethnic, and socioeconomic lines (Seyyed-Kalantari et al., 2021). The security and fairness of the implemented algorithms cannot be ensured without rigid characterization of training and validation groups.

Third, human-AI research often has poor appropriateness of comparators. Studies often compare AI with single readers rather than the 2 read clinical criterion, use expert readers who may not necessarily be representative of community practice, and fail to consider the high inter-reader variability that is likely to occur in clinical practice of radiology. These design choices have structural impacts that are more favorable to comparison of AI versus average or even below-average human performance rather than comparison with collaborative diagnostic processes typical of clinical practice.

Fourth, potential verification in real clinical applications is turning out to be nonexistent. Out of 187 studies, 12 (6.4) articles involved a prospective application into clinical workflow with real-time execution of the algorithm and its integration into a clinical decision-making process. Even retrospective studies, even those with carefully selected test sets, cannot capture the complexity of clinical implementation, workflow interference, absence of image quality, unexpected pathology and human

factors which characterizes real-world effectiveness.

This review identifies several gaps that exist in the study that are quite severe and should be addressed. First of all, there is a lack of studies that explore the problem of health equity regarding AI radiology. There are not numerous studies that characterize the performance stratification on the basis of demographic factors, and the ones that did it did not discuss whether algorithms can perform equally well across populations. As AI systems are increasingly contributing to the diagnostic decision, unscrupulous inequity can result in increased misfortunes to the existing inequities in healthcare.

Second, the clinical performance of AI, i.e. whether the use increases the outcome of the patient or not, is not a well-studied issue. Surrogate outcomes such as accuracy of diagnosis though important, do not necessarily translate into a better patient relevant outcome such a survival, quality of life or functional status. Randomized studies of AI in radiology have been done in scattered and randomized trials, but not the patient outcome, and the value addition to the patient is not clear.

Third, economic evaluation is scarcely done. Only in 3 studies cost-effectiveness analyses were used despite the fact that the implementation of AI requires substantial resources and that it should prove to be worth of such resource limited healthcare environment. The use of AI decisions are also made based on technological potential, but not the proven value unless they have some economic evidence.

Fourth, the AI governance tools are not established. Since algorithms are being applied and constantly revised, continuous monitoring, predicting performance drifts early, and de-adoption in the event of poor performance are relevant and largely lacking in the literature. The regulatory atmosphere, which emphasizes pre-market approval, is not very useful in post-market monitoring and lifecycle management.

Despite these limitations, however, there are several implications on radiology practice and health care policy associated with the findings. The implications of the results to practicing radiologists include the fact that AI will not replace a radiologist, but rather become a cognitive co-worker that improves detection, does measurement automatically, and prioritizes workflow without work that necessitates clinical integration and final interpretation by a human being. Being AI literate, or understanding what algorithms are capable, cannot do or cannot do, or how to fail, will be a professional necessity.

The findings would mean that implementations of AI are to be a carefully designed project that goes beyond purchasing the algorithm to healthcare organizations. Information technology infrastructure, clear governance systems, consistent performance assessment and education of clinical staff are needed to be properly deployed. The AI adoption must be viewed as a project of quality improvement by the institutions that require rigor just like other clinical interventions, such as baseline measurement, implementation support and outcome evaluation.

The findings to the policymakers and regulators demonstrate that new mechanisms are needed to address the special problems of AI in medicine. The pre-market assessment should entail external validation in other population and performance across clinically relevant subgroups. It should also impose post-market surveillance and mechanisms of performance drift and adverse event reporting should exist. The reimbursement policies must condition incentive to demonstrated clinical value rather

than implementation of technology itself.

Limitations of This Review

It is a systematic review, which has several limitations. First, despite intensive searching, the research may have publication bias, poor-performing studies are less likely to be published, and this may overestimate the perceived efficiency of machine learning applications. Second, English-language publications may cause language bias due to the restriction, but sensitivity analysis did not show any statistically significant differences in the conclusions obtained when non-English studies were excluded. Third, studies were not done in a way that could be subject to meta-analysis, and hence, limited quantitative synthesis by methodology heterogeneity. Fourth, the field is rapidly moving, which means that the published research might be obsolete, particularly in fast-moving areas of research such as vision-language models and large language models. Fifth, the quality measurement instruments deployed via standard, may fall short of all dimensions that can be applied to AI research, including its aspects of algorithmic fairness and generalizability.

CONCLUSION

It is a systematic review of machine learning in radiology that comprises diagnostic accuracy, radiomics, segmentation, multimodal integration, workflow optimization and implementation science. The reality is that machine learning has already achieved impressive technical achievements such as discovering subtle pathologies, quantitative biomarkers which are not available to the human eye, tedious measurements which can be automated by machines and prioritized important results. These aspects render machine learning a disruptive technology to the field of diagnostic imaging, as it will address the difficulties of increasing imaging workloads, staffing concerns, and inconsistency of interpretation accuracy.

However, in the review, there is also a huge gap in translation. Nearly all researches are retrospective, algorithmic, and under the unnatural conditions, which do not reflect the clinical reality. Externality is more of an exemption rather than the rule, prospective research is very rare and patient benefits are virtually unexplored. Methodology quality is also not determinate and is often of low quality and much of quality of study methodology, population description, and statistical interpretation is lacking. The resultant literature captures what algorithms are capable of when in controlled settings and provides minimal guidance as to what algorithms ought to do and will do in clinical practice.

Bringing about this gap in the research requires fundamental modifications in the way the field approaches research. The direction forward work must take is with the emphasis on the external validation as the pre-condition to the credibility, prospective evaluation as the pre-condition to implementation and patient outcomes as the ultimate value measure. It must embrace the complexity of the clinical processes, investigate human-AI collaboration as a shared mental system, and not algorithms per se. It must address equity in a simple way, and AI advantages must be used with all groups of people rather than by enhancing the existing inequity. And it must develop governmental forms that are adequate in response to the unique needs of the ever-changing, autonomously operating systems which propel diagnostic judgments with important consequences on patients.

Still irresistible is the future of machine learning in radiology to provide more accurate diagnoses, faster reporting, accuracy at the quantitative level, and offload the overworked

radiologists. To fulfill this commitment, one should go beyond technical demonstration into rigorous, systematic evaluation of clinical effectiveness, implementation strategies and patient impact. The next decade will answer the question of whether machine learning is capable of meeting its hype of being transformative or will be another casualty of an example of how technology can achieve a lot, but implementations rarely follow the example.

REFERENCES

Annarumma, M., Withey, S. J., Bakewell, R. J., et al. (2019). Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology, 291(1), 196-202. https://doi.org/10.1148/radiol.2018180921

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digital Medicine, 3, 118. https://doi.org/10.1038/s41746-020-00324-0

Brady, A. P., Bello, J. A., Derchi, L. E., et al. (2021). Radiology in the era of value-based healthcare: A multi-society expert statement from the ACR, CAR, ESR, IS3R, RANZCR, and RSNA. Radiology, 298(3), 486-491. https://doi.org/10.1148/radiol.2020204027

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2020). A review of eye tracking for understanding and improving diagnostic interpretation. Cognitive Research: Principles and Implications, 5(1), 52. https://doi.org/10.1186/s41235-020-00250-x

Castro, D. C., Walker, I., & Glocker, B. (2020). Causality matters in medical imaging. Nature Communications, 11(1), 3673. https://doi.org/10.1038/s41467-020-17478-w

Coiera, E. (2019). The last mile: Where artificial intelligence meets reality. Journal of Medical Internet Research, 21(11), e16323. https://doi.org/10.2196/16323

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563-577. https://doi.org/10.1148/radiol.2015151169

Kelly, B. S., Judge, C., Bollard, S. M., et al. (2022). Radiology artificial intelligence: A systematic review and evaluation of methods (RAISE). European Radiology, 32(11), 7998-8007. https://doi.org/10.1007/s00330-022-08918-2

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2020). 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, 21(4), 405-416. https://doi.org/10.3348/kjr.2020.0023

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

Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Liu, X., Faes, L., Kale, A. U., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271-e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Maghsoudi, H., et al. (2025). Physics-aware imaging AI for quantitative MASLD biomarker mapping: A systematic review of deep learning and radiomics across ultrasound, CT, and MRI. Abdominal Radiology. Advance online publication. https://doi.org/10.1007/s00261-025-05317-9

McDonald, R. J., Schwartz, K. M., Eckel, L. J., et al. (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Academic Radiology, 22(9), 1191-1198. https://doi.org/10.1016/j.acra.2015.05.007

Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiology: Artificial Intelligence, 2(2), e200029. https://doi.org/10.1148/ryai.2020200029

Nagendran, M., Chen, Y., Lovejoy, C. A., et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. BMJ, 368, m689. https://doi.org/10.1136/bmj.m689

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology, 286(3), 800-809. https://doi.org/10.1148/radiol.2017171920

Pinto dos Santos, D., Dietzel, M., & Baessler, B. (2019). A decade of radiomics research: Are images really data or just patterns in the noise? European Radiology, 29(7), 3428-3430. https://doi.org/10.1007/s00330-019-06199-2

Popay, J., Roberts, H., Sowden, A., et al. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme. https://doi.org/10.13140/2.1.1018.4643

Roberts, M., Driggs, D., Thorpe, M., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate

for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3, 199-217. https://doi.org/10.1038/s42256-021-00307-0

Sanduleanu, S., Woodruff, H. C., de Jong, E. E. C., et al. (2018). Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiotherapy and Oncology, 127(3), 349-360. https://doi.org/10.1016/j.radonc.2018.03.033

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176-2182. https://doi.org/10.1038/s41591-021-01595-0

Sit, C., Srinivasan, R., Amlani, A., et al. (2020). Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: A multicentre survey. Insights into Imaging, 11(1), 14. https://doi.org/10.1186/s13244-019-0830-7

Smith, B. D., Smith, G. L., & Hurria, A. (2020). Future of cancer incidence in the United States: Burdens upon an aging, changing nation. Journal of Clinical Oncology, 38(19), 2149-2159. https://doi.org/10.1200/JCO.20.00112

Ul Haq, I., et al. (2025). Advancements in medical radiology through multimodal machine learning: A comprehensive overview. Bioengineering, 12(5), 477. https://doi.org/10.3390/bioengineering12050477

Vidal-Mondéjar, J., et al. (2024). Methodological evaluation of systematic reviews based on the use of artificial intelligence systems in chest radiography. Radiología (English Edition), 66(4), 326-339. https://doi.org/10.1016/j.rxeng.2023.01.015

Viswam, P., et al. (2025). Artificial intelligence in radiology and diagnostic imaging. Bioinformation, 21(7), 1891-1894. https://doi.org/10.6026/97320630021891

Whiting, P. F., Rutjes, A. W., Westwood, M. E., et al. (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

Yahaya, B. S., et al. (2025). Radiomics and deep learning characterisation of liver malignancies in CT images – A systematic review. Computers in Biology and Medicine. https://doi.org/10.1016/j.compbiomed.2025.108842

Yao, A. D., Cheng, D. L., Pan, I., & Kitamura, F. (2021). Deep learning in neuroradiology: A systematic review of current algorithms and approaches for the new wave of imaging technology. Radiology: Artificial Intelligence, 3(2), e200026. https://doi.org/10.1148/ryai.2020200026

Zając, H. D., et al. (2025). Human–AI interaction and collaboration in radiology: From conceptual frameworks to responsible implementation. Diagnostic and

Interventional Radiology. Advance online publication. https://doi.org/10.4274/dir.2025.263780