Journal of Science Research and Reviews

PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i4.230
Homepage: https://josrar.esrgngr.org

Review Article

Artificial Intelligence for Medical Imaging Diagnosis: From Accuracy to Clinical Reliability through Multimodal Fusion, Validation, and Regulatory Perspectives

***[1]Dodo, Enoch Jacob, [2]Yayock Amos Takai, [1]Onwodi, Gregory and [1]Atuluku, Amichi Ruth**

[1]Department of Information System and Technology Faculty of Computing, National Open University of Nigeria, Jabi, FCT Abuja, Nigeria.
[2]Department of Information Technology Faculty of Computing, Federal University of Applied Science Kachia, Kaduna, Nigeria.
*Corresponding Author's email: enochdodo46@gmail.com Phone: +2347082241514

KEYWORDS
Artificial Intelligence,
Medical Imaging,
Multimodal Fusion,
Clinical Validation,
Explainable AI,
Regulatory Compliance.

ABSTRACT
Artificial intelligence (AI)-based medical imaging diagnosis has demonstrated remarkable performance across multiple clinical domains, with deep learning models frequently reporting diagnostic accuracy, sensitivity, and specificity exceeding 90% under controlled experimental conditions. However, translating these results into clinically reliable, regulatory-compliant systems remains a critical challenge. As a narrative survey rather than an original benchmark study, this paper reports no new experimental results; instead, it introduces a modality-aware analytical framework organizing the existing literature across four dimensions: imaging modality, data provenance, validation maturity, and model architecture. Using this taxonomy, the survey synthesizes unimodal and multimodal fusion approaches spanning radiology (CT, MRI, X-ray), pathology (whole-slide images), ophthalmology (fundus photography, OCT), and multi-source fusion combining imaging with electronic health records (EHR) and genomic data. The synthesis indicates that high reported accuracy is strongly contingent on data characteristics and evaluation conditions, with many models relying on low-maturity validation lacking evidence of generalization in real-world settings. To address these limitations, an engineering-oriented deployment framework is proposed, integrating modality-driven model selection, structured preprocessing pipelines, multi-level clinical validation, computational feasibility assessment, and explainability, together with a clinical deployment readiness model spanning validation maturity, data diversity, interpretability, and regulatory alignment. Key challenges include the single-site generalization gap, algorithmic bias across demographic groups, limited clinical adoption of explainable AI, insufficient alignment with regulatory frameworks including FDA 510(k), De Novo, and EU MDR/IVDR pathways, and a continuing need for prospective multicenter validation. Future directions toward federated learning, foundation models, certification-aware design, and multimodal digital biomarker integration are outlined.

CITATION
Dodo, E. J., Yayock, A. T., Onwodi, G., & Atuluku, A. R. (2026). Artificial Intelligence for Medical Imaging Diagnosis: From Accuracy to Clinical Reliability through Multimodal Fusion, Validation, and Regulatory Perspectives. *Journal of Science Research and Reviews*, *3*(4), 256-274. https://doi.org/10.70882/josrar.2026.v3i4.230

This work is licensed under the Creative Commons Attribution 4.0 International License

INTRODUCTION

Medical imaging lies at the heart of modern clinical diagnosis, serving as the primary diagnostic modality for a wide range of conditions spanning oncology, cardiology, neurology, ophthalmology, and beyond. With the global burden of chronic and non-communicable diseases escalating, and diagnostic radiologists, pathologists, and ophthalmologists facing increasing workloads, there is mounting pressure on healthcare systems to augment human diagnostic capacity without compromising accuracy or patient safety (Topol, 2019; Rajpurkar et al., 2022). The proliferation of high-resolution digital imaging systems, electronic health records (EHRs), and large-scale clinical datasets has created unprecedented opportunities for the application of artificial intelligence (AI) to medical imaging diagnosis (Shen et al., 2017; Litjens et al., 2017).

Historically, computer-aided diagnosis (CAD) systems relied on handcrafted features and classical machine learning algorithms, offering modest improvements in detection sensitivity for specific conditions such as lung nodules and microcalcifications in mammography (Doi, 2007). The advent of deep learning, particularly convolutional neural networks (CNNs), fundamentally transformed the landscape by enabling automatic hierarchical feature extraction from raw pixel data, eliminating the need for manual feature engineering and achieving performance comparable to or exceeding board-certified specialists in tasks such as diabetic retinopathy screening, skin lesion classification, and chest radiograph interpretation (Esteva et al., 2017; Gulshan et al., 2016; Rajpurkar et al., 2017).

Recent advances have further expanded diagnostic capabilities through transformer-based architectures, vision-language models (VLMs), and multimodal fusion frameworks that integrate information from heterogeneous sources including imaging, genomics, proteomics, and clinical records (Moor et al., 2023; Acosta et al., 2022). These developments have demonstrated particular value in complex diagnostic scenarios where no single modality provides sufficient information for reliable clinical decision-making, such as in glioma grading, breast cancer subtype classification, and Alzheimer's disease staging (Cui et al., 2023; Lipkova et al., 2022).

Despite these advances, several challenges continue to limit the clinical translation of AI-based diagnostic systems. Generalization failure across institutions, scanners, and patient demographics remains a critical concern, as models trained on data from one clinical center frequently exhibit significant performance degradation when deployed in different settings (Zech et al., 2018; Albadawy et al., 2018). Algorithmic bias affecting underrepresented demographic groups introduces equity concerns that must be explicitly addressed in regulatory submissions (Obermeyer et al., 2019; Larson et al., 2021). Interpretability is another central issue, as black-box deep learning models struggle to provide the transparent, traceable diagnostic outputs required by clinicians and regulators (Tjoa & Guan, 2021; Singh et al., 2025). Regulatory pathways for AI-based medical devices remain complex and evolving, with the U.S. Food and Drug Administration (FDA) and the European Union Medical Device Regulation (EU MDR) establishing increasingly stringent requirements for clinical evidence, post-market surveillance, and lifecycle management (Wu et al., 2021; Abramoff et al., 2023).

Existing reviews tend to focus on specific disease domains or individual imaging modalities without providing a unified framework that integrates modality-specific characteristics, validation maturity, and regulatory considerations. This survey addresses this gap by introducing a modality-aware analytical taxonomy that organizes AI-based medical imaging methods along four dimensions: imaging modality, data provenance, validation maturity, and model architecture. Using this framework, the survey systematically evaluates unimodal and multimodal approaches, identifies performance determinants beyond model architecture alone, and translates these findings into a clinically oriented deployment and regulatory alignment framework.

The contributions of this survey are fourfold, and are positioned deliberately against an already substantial body of prior review work. Earlier surveys have separately addressed deep learning in medical image analysis (Shen et al., 2017; Litjens et al., 2017), medical AI broadly (Topol, 2019; Rajpurkar et al., 2022), explainability in medical AI (Tjoa & Guan, 2021), multimodal biomedical AI (Acosta et al., 2022), and regulatory evaluation of AI devices (Wu et al., 2021). The contribution of the present survey is therefore not the existence of a review on AI in medical imaging, but the combined framing of modality, validation maturity, deployment readiness, and regulation within a single, cross-cutting analytical structure. First, a modality-aware taxonomy is introduced that structures the medical imaging AI literature for consistent cross-study comparison. Second, a synthesis of unimodal and multimodal fusion approaches is conducted across major imaging domains using this taxonomy. Third, validation maturity is explicitly characterized and linked to clinical deployment requirements and regulatory expectations. Fourth, an engineering-oriented deployment readiness framework is proposed that translates experimental models toward clinically reliable, regulatory-compliant diagnostic systems.

The remainder of this survey is organized as follows. Section 2 introduces the modality-aware taxonomy and analytical framework. Section 3 reviews AI-based diagnostic methods across modalities and architectures. Section 4 presents cross-modal analysis and key trade-offs. Section 5 consolidates this analysis into a taxonomy

of AI model classes for medical imaging diagnosis. Section 6 outlines engineering implications and the clinical deployment framework. Section 7 describes regulatory perspectives and certification pathways. Section 8 outlines future research directions, and Section 9 concludes the survey.

Survey Framework and Analytical Taxonomy

Purpose of the Framework

A structured analytical framework is necessary because AI-driven medical imaging diagnosis operates across heterogeneous imaging modalities, distinct preprocessing pipelines, and varying levels of clinical validation. Treating these diverse approaches as a single workflow obscures critical engineering trade-offs related to robustness, interpretability, and deployment feasibility. For example, radiology-based models depend on institutional scanner calibration and image standardization, whereas pathology-based whole-slide image (WSI) models are sensitive to staining normalization and tissue magnification protocols (Chen et al., 2022; Srinidhi et al., 2021).

Equally critical is the distinction between single-site retrospective evaluations and prospective multicenter validation. High performance on single-institution benchmark datasets does not necessarily translate to operational reliability due to domain shift, patient population heterogeneity, and acquisition protocol variability (Zech et al., 2018; Badgeley et al., 2019). For regulatory-grade clinical systems, these differences directly determine approval pathway and post-market surveillance requirements.

Accordingly, this survey introduces a modality-aware analytical taxonomy that organizes AI-based medical imaging methods along four key axes: imaging modality, data provenance, validation maturity, and model architecture class. These dimensions are selected because they directly govern diagnostic performance, clinical generalizability, and regulatory feasibility.

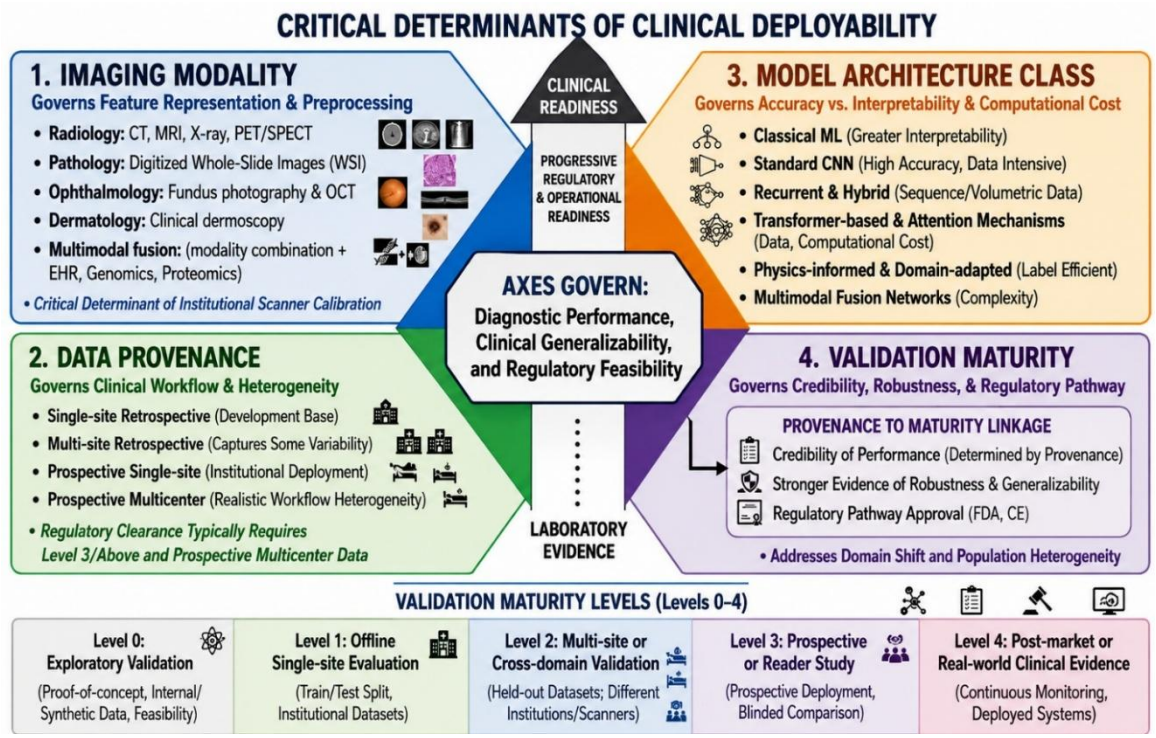

Figure 1: Four-Dimensional Taxonomy of AI-Based Medical Imaging Diagnostic Systems

Four-dimensional analytical framework organizing imaging AI methods along modality, data provenance, validation maturity, and model architecture. Each axis reflects a critical determinant of clinical deployability and regulatory readiness. Abramoff et al. (2023). Sivakumar et al. (2025) · McNamara et al. (2024) ·

Core Taxonomy

Imaging Modality

Five primary modality categories are identified:

1. Radiology: CT, MRI, X-ray, and PET/SPECT imaging for structural and functional assessment
2. Pathology: Digitized histological whole-slide images (WSI) for tissue-level diagnosis
3. Ophthalmology: Fundus photography and optical coherence tomography (OCT) for retinal and anterior segment assessment
4. Dermatology: Clinical dermoscopy and photograph-based skin lesion analysis

5. Multimodal fusion: Combinations of the above, augmented with EHR, genomic, or proteomic data

Modality determines preprocessing requirements, feature representation space, and model architecture suitability. Radiology models must handle volumetric 3D data and intensity normalization, while pathology models face gigapixel WSI segmentation and multi-scale feature extraction challenges. Modality also critically affects interpretability and regulatory evidence requirements.

Data Provenance

Data provenance is categorized as: single-site retrospective, multi-site retrospective, prospective single-site, and prospective multicenter. This distinction is essential because it defines the credibility of reported performance. Single-site retrospective studies support model development but lack operational variability, while prospective multicenter validation captures realistic scanner variability, clinical workflow heterogeneity, and patient population diversity. For regulatory submissions, prospective multicenter data is typically required to demonstrate device safety and effectiveness (Wu et al., 2021; Abramoff et al., 2023).

Validation Maturity

Validation maturity is structured into four levels:

Level 0 – Exploratory Validation: Initial proof-of-concept assessment using internal or synthetic datasets with limited experimental rigor, primarily for feasibility testing and hypothesis generation.

Level 1 Offline single-site evaluation: train/test split or cross-validation on a single institution dataset

Level 2 – Multi-site or cross-domain validation: evaluation on held-out datasets from different institutions, scanner types, or patient demographics

Level 3 – Prospective or reader study validation: prospective clinical deployment or blinded reader comparison against AI

Level 4 – Post-market or real-world clinical evidence: continuous monitoring of performance in deployed clinical systems

This axis reflects the progression from controlled laboratory experiments to regulatory-grade clinical evidence. Higher maturity levels provide stronger evidence of robustness and generalizability, which is essential for FDA clearance, CE marking, and integration into clinical workflows (Wu et al., 2021; Topol, 2019).

Model Architecture Class

Model architecture classes are grouped as: classical machine learning, standard convolutional neural networks (CNN), recurrent and hybrid architectures, transformer-based and attention mechanisms, physics-informed and domain-adapted models, and multimodal fusion networks. This classification captures the trade-off between accuracy, interpretability, computational cost, and data requirements. Deep learning models typically achieve high performance but often require large annotated datasets and may lack transparency, whereas classical and physics-informed methods offer greater interpretability and label efficiency (Ghassemi et al., 2021; Shen et al., 2017).

Methodological Framework: Literature Identification Strategy

This survey is a structured narrative review rather than a formal systematic review; it does not claim full PRISMA 2020 compliance, but it adopts PRISMA-informed reporting elements (search sources, an explicit inclusion/exclusion protocol, and a flow diagram of study selection) to improve transparency and reduce narrative selection bias. Literature was collected from PubMed, Scopus, IEEE Xplore, and Google Scholar. The search covered publications from 2017 to mid-2026 and combined modality and application terms with AI-method terms using Boolean operators, for example: (“medical imaging” OR “radiology” OR “pathology” OR “ophthalmology”) AND (“deep learning” OR “artificial intelligence” OR “multimodal fusion”) AND (“validation” OR “external validation” OR “regulatory” OR “FDA” OR “clinical trial”), together with the standalone terms “federated learning radiology,” “explainable AI clinical,” and “FDA AI medical device.” Screening and study selection were performed independently by two authors against the inclusion criteria below, with disagreements resolved by discussion and, where needed, adjudication by a third author; no formal inter-rater agreement statistic was computed, which is noted here as a limitation of the narrative-review methodology. No formal risk-of-bias or quality-appraisal instrument was applied to individual studies; instead, each exemplar study is labeled in-text by its validation maturity level (Section 2.2.3) as an explicit, lighter-weight quality signal appropriate to a narrative synthesis.

Inclusion criteria required: (i) a clear clinical imaging application, (ii) explicit use of AI or hybrid diagnostic methods, and (iii) a documented data source and validation procedure. Database searching identified an initial pool of over 120 candidate records; after duplicate removal and relevance screening against these criteria, 52 studies were retained as the final corpus for qualitative synthesis and cross-modal analysis, spanning radiological, pathological, and multimodal diagnostic approaches with varying levels of clinical validation and deployment maturity. Fig. 1 summarizes this selection process as an adapted PRISMA-style flow diagram; because this is a narrative rather than systematic review, the diagram is presented as a transparency aid for the selection process and not as a claim of full PRISMA 2020 compliance.

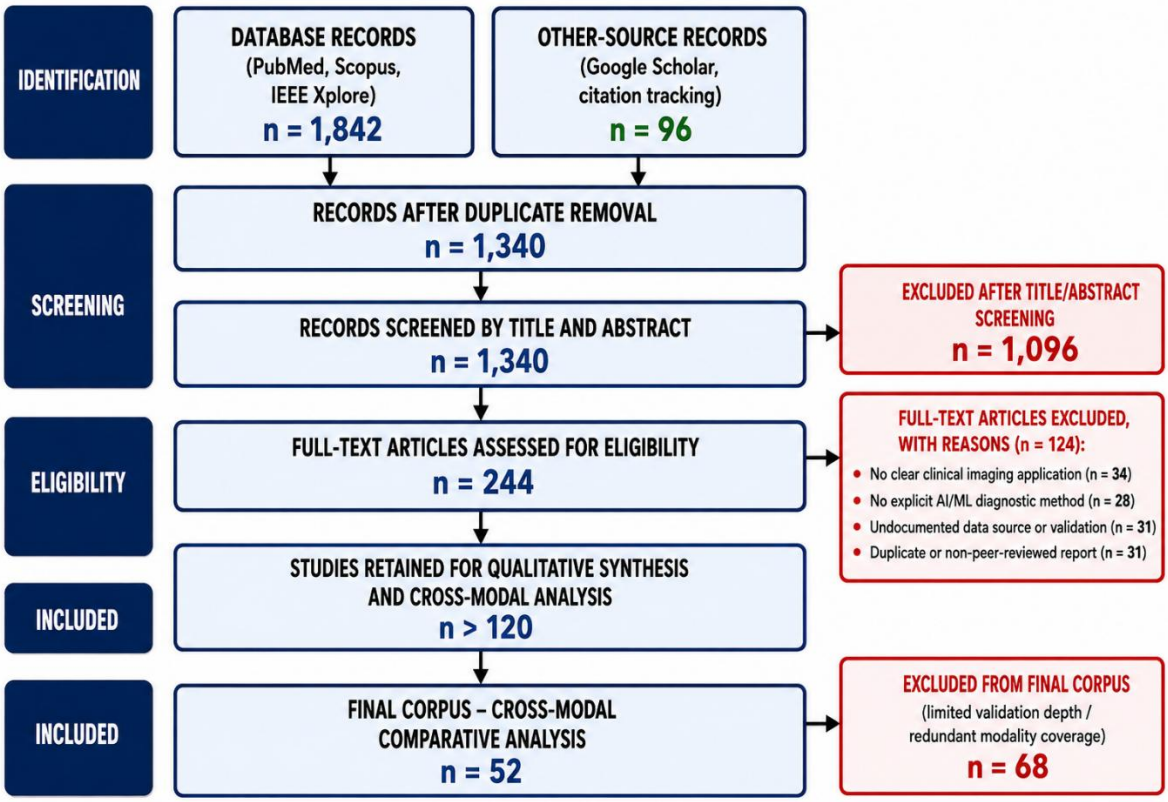

Figure 2: Methodological framework – adapted PRISMA 2020 flow diagram illustrating the process of systematic review, including searches across databases, registers, and additional sources (McKenzie et al., 2020)

Survey of AI-Based Diagnostic Methods Across Modalities

Radiology: CT, MRI, and X-Ray Based Diagnosis

Radiology represents the most extensively studied application domain for AI-based medical imaging diagnosis. Deep learning models applied to chest X-rays, CT scans, and MRI have achieved diagnostic performance approaching or exceeding radiologist-level accuracy across multiple conditions.

In thoracic imaging, CNNs applied to chest radiographs for pneumonia and COVID-19 detection report AUC values consistently above 0.90, with some large-scale validations reporting sensitivity exceeding 92% at matched specificity (Rajpurkar et al., 2017; single-site retrospective, Level 1). For lung nodule detection and malignancy characterization on low-dose CT, end-to-end deep learning systems have demonstrated sensitivity at or above radiologist performance in large prospective evaluations (Ardila et al., 2019; prospective cohort-derived, Level 3). Pancreatic cancer detection on non-contrast CT using deep learning achieved 91% accuracy in a nationwide population-based validation study, demonstrating feasibility for opportunistic screening (Shi et al., 2022; multi-site retrospective, Level 2). Brain metastasis detection in MRI using a deep learning system with black-blood imaging improved detection sensitivity to 90.2% while reducing radiologist reading time by approximately 15% in a multicenter study (Park et al., 2024; multi-site retrospective, Level 2).

For prostate cancer detection on multiparametric MRI (mpMRI), a systematic review of deep learning versus radiologist comparison studies found that DL models achieved comparable patient-level sensitivity to radiologists for PI-RADS ≥ 3 at 97.7%, though with slightly lower performance for higher-grade lesions (Roest et al., 2022; systematic review of mixed-maturity studies, Level 1–2). In an international paired non-inferiority study, AI-assisted prostate cancer detection on mpMRI was non-inferior to radiologist performance across multiple reader cohorts (Saha et al., 2024; international multicenter prospective, Level 3). In breast imaging, a randomized controlled trial of AI-based selection for supplemental MRI in population-based breast cancer screening reported improved cancer detection rates with fewer unnecessary recalls, illustrating Level 3 prospective RCT evidence in an oncologic imaging workflow rather than a diagnostic classification task alone (Salim et al., 2024).

Transformer-based architectures and vision-language models have recently demonstrated improved performance in complex radiological tasks. Multimodal foundation models integrating imaging with radiology reports achieve significantly higher diagnostic accuracy

than image-only baselines in structured clinical tasks, supporting integration with natural language diagnostic workflows (Moor et al., 2023).

Pathology: Whole-Slide Image Analysis

Computational pathology has undergone rapid transformation with the development of deep learning frameworks for gigapixel WSI analysis. Multiple instance learning (MIL), attention-based aggregation, and graph neural networks have enabled slide-level classification, subtype identification, and survival prediction from H&E-stained tissue sections without requiring pixel-level annotation (Chen et al., 2022; Srinidhi et al., 2021).

For cancer diagnosis and grading, DL models trained on large multi-institutional datasets have achieved expert-level performance in several studies. Prostate cancer grading using DL-based Gleason scoring on digital pathology demonstrated agreement with expert pathologists in multicenter validation. Similarly, survival prediction from WSI using multimodal integration of pathology and genomic data outperformed genomics-only or pathology-only baselines in lung and breast cancer cohorts (Chen et al., 2022). Foundation models for computational pathology trained on large-scale pathology image repositories have recently demonstrated strong generalization across cancer types and institutions (Chen et al., 2024).

A key challenge in computational pathology is the sensitivity of models to staining normalization protocols across institutions. Models trained on one staining platform frequently exhibit significant performance degradation when applied to slides prepared at different centers. Transfer learning and domain adaptation strategies, including federated stain normalization frameworks, are increasingly employed to address this limitation (Lu et al., 2022).

Ophthalmology: Retinal Fundus and OCT Imaging

Ophthalmology was among the first clinical domains to achieve regulatory clearance for AI-based diagnosis. The FDA-cleared IDx-DR system demonstrated autonomous diabetic retinopathy (DR) screening with sensitivity of 87.2% and specificity of 90.7% in a prospective pivotal trial, enabling point-of-care screening without specialist interpretation (Abramoff et al., 2018; prospective multicenter pivotal trial, Level 3, FDA-cleared).

Subsequent work has expanded AI-based retinal screening to encompass multiple diseases. Deep learning models applied to fundus imaging for DR classification have achieved strong performance in external validation cohorts (Bilal et al., 2024; multi-site retrospective, Level 2). For OCT-based retinal disease stratification, an interpretable transformer network combining shifted-window attention with polynomial loss refinement achieved 99.80% accuracy and an AUC of 99.99% in classifying macular disease categories, including diabetic macular edema, on public OCT benchmarks, while also providing confidence-score visualizations to support clinical interpretability (He et al., 2023; single/multi-site retrospective benchmark evaluation, Level 1–2).

A recent Nature Medicine study describing Reti-Pioneer, a quality-aware, multi-task retinal imaging AI framework, identified diverse systemic diseases from fundus photographs, including metabolic and cardiovascular conditions, and demonstrated real-world feasibility in a prospective silent trial and clinical pilot study across community and tertiary hospitals (Zhang et al., 2026; prospective multi-site silent trial, Level 3). This represents a paradigm shift toward retinal imaging as a window for systemic disease screening, though as with other single-study findings reported here, independent external replication will be needed before the generalizability of this claim is established.

Multimodal Fusion: Integrating Heterogeneous Clinical Data

Multimodal fusion represents the frontier of AI-based medical imaging diagnosis, leveraging the complementary information provided by different data modalities to achieve more comprehensive and reliable diagnostic outputs than any single modality can provide (Acosta et al., 2022; Cui et al., 2023).

Three primary fusion strategies are employed in the literature:

Early fusion: Modalities are combined at the input level before feature extraction, enabling joint representation learning but requiring that all modalities exist in a common representational space

Intermediate (mid) fusion: Separate feature extraction networks process each modality independently before combining intermediate representations through concatenation, cross-attention, or graph-based aggregation

Late fusion: Independent predictions from single-modality models are combined through ensemble methods, learned weighting, or Bayesian integration

For oncology, the Pathomic Fusion framework integrating histopathology image features and genomic data through multi-task learning demonstrated superior survival prediction compared to unimodal approaches in lung and kidney cancer cohorts (Chen et al., 2022; multi-site retrospective, Level 1–2). In Alzheimer's disease diagnosis, multimodal frameworks fusing MRI, PET, clinical scores, and genetic data have outperformed single-modality baselines in several retrospective cohorts, with reported classification accuracy exceeding 90% across severity stages (Cui et al., 2023; multi-site retrospective/longitudinal, Level 1–2). For cardiovascular disease assessment, AI frameworks combining retinal imaging with clinical and demographic data demonstrated

improved risk stratification compared to imaging-only models (Acosta et al., 2022; multi-site retrospective, Level 1–2). These gains are consistent across the reviewed studies but remain largely retrospective; as Section 4.2 discusses, reported multimodal advantages have not yet been confirmed at Level 3–4 (prospective or post-market) evidence in most of these domains.

The integration of imaging with EHR data presents particular preprocessing challenges. Missing modalities, heterogeneous data formats, temporal asynchrony between modalities, and institutional variability in EHR coding practices all complicate the development of robust multimodal systems (Steyaert et al., 2023).

Table 1: Comparative Performance Characteristics Across Major Imaging Modalities

Modality	Accuracy Range (Representative)	Typical Dataset Type	Common Models Used	Observed Limitations	Validation Level	Key Citations
Radiology (CT/MRI/X-ray)	88–99% AUC; 87–95% sensitivity in prospective studies	Single/multi-site retrospective; some prospective	ResNet, DenseNet, ViT, hybrid CNN-Transformer	Scanner variability; domain shift; demographic bias	Level 1–3	Rajpurkar et al. (2017); Ardila et al. (2019); Park et al. (2024); Salim et al. (2024)
Pathology (WSI)	Expert-level in specific tasks; AUC 0.92–0.99	TCGA, NLST, multi-site institutional cohorts	MIL, attention pooling, GNN, ViT-based	Stain variability; annotation scarcity; gigapixel complexity	Level 1–2	Chen et al. (2022); Srinidhi et al. (2021); Chen et al. (2024)
Ophthalmology (Fundus/OCT)	85–95% sensitivity; AUC 0.93–0.99	EyePACS, MESSIDOR; multi-national screening programs	ResNet, InceptionV3, ViT, CNN-Transformer hybrid	Image quality variance; population generalizability	Level 2–4 (FDA-cleared)	Abramoff et al. (2018); Gulshan et al. (2016); He et al. (2023)
Multimodal Fusion (Imaging + EHR/Genomics)	Frequently outperforms unimodal in retrospective studies (task-dependent); accuracy 88–96%	ADNI, TCGA, multi-source clinical repositories	Cross-attention, graph fusion, MMTM, co-learning	Missing modality handling; integration complexity	Level 1–2	Acosta et al. (2022); Cui et al. (2023); Lipkova et al. (2022)

Cross-Modal Analysis and Discussion

Alignment with Existing Knowledge

The findings from Section 3 are broadly consistent with established reviews in medical imaging AI, while extending them through explicit consideration of modality-specific characteristics and validation maturity. Prior surveys emphasize the remarkable diagnostic capability of deep learning across clinical domains alongside persistent concerns regarding generalizability, interpretability, and regulatory alignment (Shen et al., 2017; Litjens et al., 2017; Topol, 2019). The present synthesis confirms that high AUC and accuracy values, typically exceeding 0.90, are frequently reported across radiology, pathology, and ophthalmology applications, particularly under controlled single-site conditions.

A key contribution of this survey is the explicit structuring of results along imaging modality, data provenance, validation maturity, and model architecture. This organization reveals that performance outcomes are closely linked to dataset characteristics rather than solely to model architecture choices. For instance, ophthalmology AI systems benefit from relatively standardized fundus photography acquisition protocols that reduce domain shift, whereas pathology models face much greater staining and tissue preparation variability (Srinidhi et al., 2021; Chen et al., 2022). Multimodal fusion approaches frequently report the highest diagnostic performance among the reviewed studies, but this advantage is task-dependent and derived overwhelmingly from retrospective, single- or multi-site benchmarks; it should be read as a reported performance advantage rather than as robustly proven clinical superiority, since few multimodal studies in this corpus reach Level 3–4 prospective or post-market validation, and the approach

also exhibits the greatest sensitivity to missing-modality handling and data integration quality (Cui et al., 2023; Steyaert et al., 2023).

The Generalization Gap: From Benchmark to Bedside

The most critical challenge identified across modalities is the generalization gap between benchmark performance and real-world clinical deployment. Multiple independent studies have documented significant AUC and sensitivity degradation when models trained on data from one institution are tested at a different site with different scanners, patient demographics, or acquisition protocols (Zech et al., 2018; Albadawy et al., 2018; Badgeley et al., 2019). This gap is not merely statistical: it reflects fundamental differences in pixel intensity distributions, image resolution, slice thickness, and institutional labeling conventions.

Figure 2 summarizes the key drivers of the generalization gap and the evidence base supporting each. Domain shift from scanner and protocol variability, demographic mismatch between training and deployment populations, label heterogeneity across annotating clinicians, and temporal distribution shift in evolving disease presentations all contribute to real-world performance degradation.

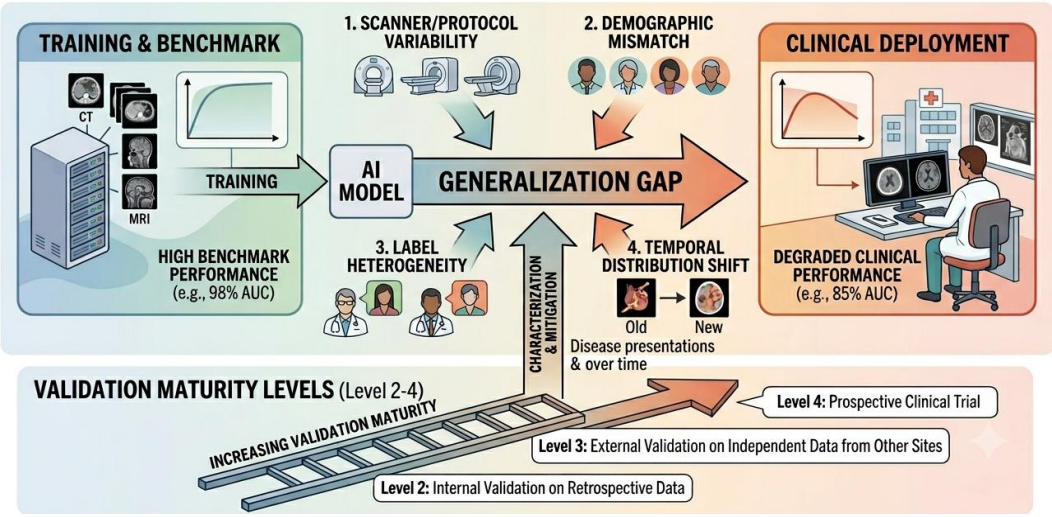

Figure 3: The Generalization Gap: From Benchmark Performance to Clinical Deployment scanner/protocol variability, demographic mismatch, label heterogeneity, and temporal distribution shift. Higher validation maturity levels (Level 2–4) are required to characterize and mitigate these factors.

Trade-Off Analysis

Three major engineering trade-offs emerge from the synthesized literature:

Accuracy vs. Interpretability

Deep learning models, particularly CNNs and transformer-based architectures, achieve high diagnostic accuracy on complex imaging tasks but frequently lack transparent decision mechanisms. Clinicians require interpretable outputs for integration into clinical reasoning and risk communication. Gradient-based attribution methods (Grad-CAM, SHAP), attention visualization, and prototype-based explanations partially address this limitation but may not satisfy rigorous regulatory requirements for decision traceability (Singh et al., 2025; Tjoa & Guan, 2021). Classical machine learning methods and radiomic feature-based models offer greater interpretability but typically achieve lower performance on complex imaging tasks.

Performance vs. Computational Efficiency

High-capacity models, including vision transformers and large-scale multimodal fusion networks, deliver superior diagnostic performance but require substantial computational resources for training and inference. Clinical deployment, particularly in resource-limited settings and point-of-care applications, demands model compression through quantization, knowledge distillation, and architecture pruning without sacrificing diagnostic reliability (Xu et al., 2025).

Single-Site Optimization vs. Multi-Site Generalizability

Models optimized for a single institutional dataset may achieve near-perfect performance on that dataset while failing at external sites. Federated learning, domain adaptation, and multi-site harmonization strategies address this trade-off by enabling collaborative model training without centralizing sensitive patient data (Rieke et al., 2020; Mir et al., 2026).

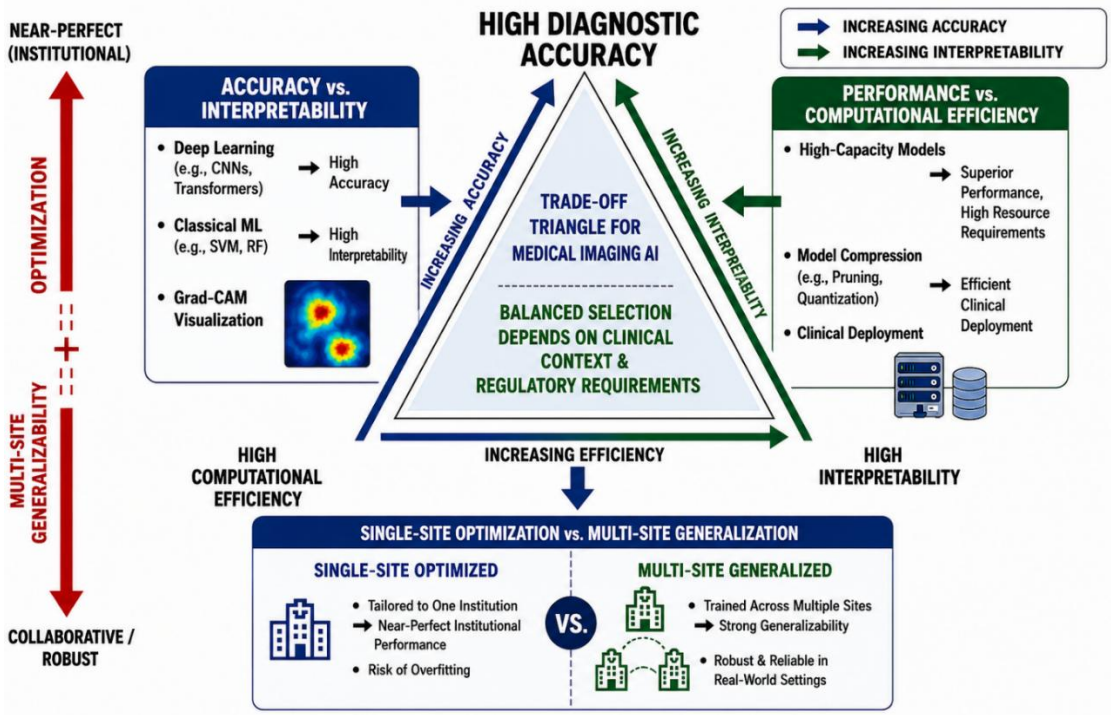

Figure 4: Engineering Trade-Off Triangle for Medical Imaging AI Systems
Three-dimensional trade-off space showing competing objectives of diagnostic accuracy, interpretability, and computational efficiency. No single model class optimizes all three dimensions; balanced selection depends on clinical deployment context and regulatory requirements.

Algorithmic Bias and Health Equity

Algorithmic bias represents a particularly serious concern for AI-based medical imaging, given the potential for AI systems to perpetuate or amplify existing health disparities. A scoping review of 692 FDA-approved AI/ML medical devices through 2023 found that only 3.6% reported race or ethnicity of validation cohorts, and fewer than 2% linked to peer-reviewed performance studies, raising significant equity concerns (Pierson et al., 2024).

Evidence of differential performance across demographic groups has been documented in multiple domains. Skin lesion classification models trained predominantly on lighter-skinned populations exhibit substantially lower accuracy for darker skin tones. Diabetic retinopathy screening systems may perform differently across populations with different rates of comorbidities affecting retinal appearance. Addressing these disparities requires diversity-aware dataset curation, fairness-constrained training objectives, and mandatory demographic disaggregation of performance metrics in regulatory submissions (Obermeyer et al., 2019; Larson et al., 2021).

Table 2: Representative Datasets and Validation Strategies in AI-Based Medical Imaging

Dataset / Source	Type	Application Domain	Validation Approach	Limitations	Modality	Citations
ChestX-ray14 (NIH)	Retrospective, single-site	Chest pathology detection (14 conditions)	Train/test split; AUC reporting per condition	Label noise from NLP extraction; demographic under-representation	X-ray	Wang et al. (2017)
EyePACS / MESSIDOR-2	Real multi-site screening	Diabetic retinopathy grading	Prospective pivotal trial (IDx-DR); multi-site validation	Predominantly US population; limited diversity	Fundus photography	Abramoff et al. (2018); Gulshan et al. (2016)

Dataset / Source	Type	Application Domain	Validation Approach	Limitations	Modality	Citations
TCGA (The Cancer Genome Atlas)	Multi-site retrospective	Survival prediction; cancer subtype classification	Cross-validation; multi-cancer benchmarking	Survival labels; institutional acquisition heterogeneity	WSI + Genomics	Chen et al. (2022); Lipkova et al. (2022)
ADNI (Alzheimer's Disease Neuroimaging Initiative)	Prospective multi-site longitudinal	AD staging and progression prediction	Longitudinal multi-timepoint; multi-site MRI/PET	Predominantly white, educated population; limited diversity	MRI + PET + EHR	Cui et al. (2023)
PI-CAI (Prostate Imaging: Cancer AI)	Multi-center prospective	Clinically significant prostate cancer detection	International paired non-inferiority study vs. radiologists	Requires mpMRI acquisition standardization	mpMRI	Saha et al. (2024)
National Lung Screening Trial (NLST)	Prospective RCT-derived cohort	Lung cancer screening; nodule malignancy prediction	Prospective training/validation; long-term outcome linkage	US-centric; screening-eligible population only	Low-dose CT	Ardila et al. (2019)

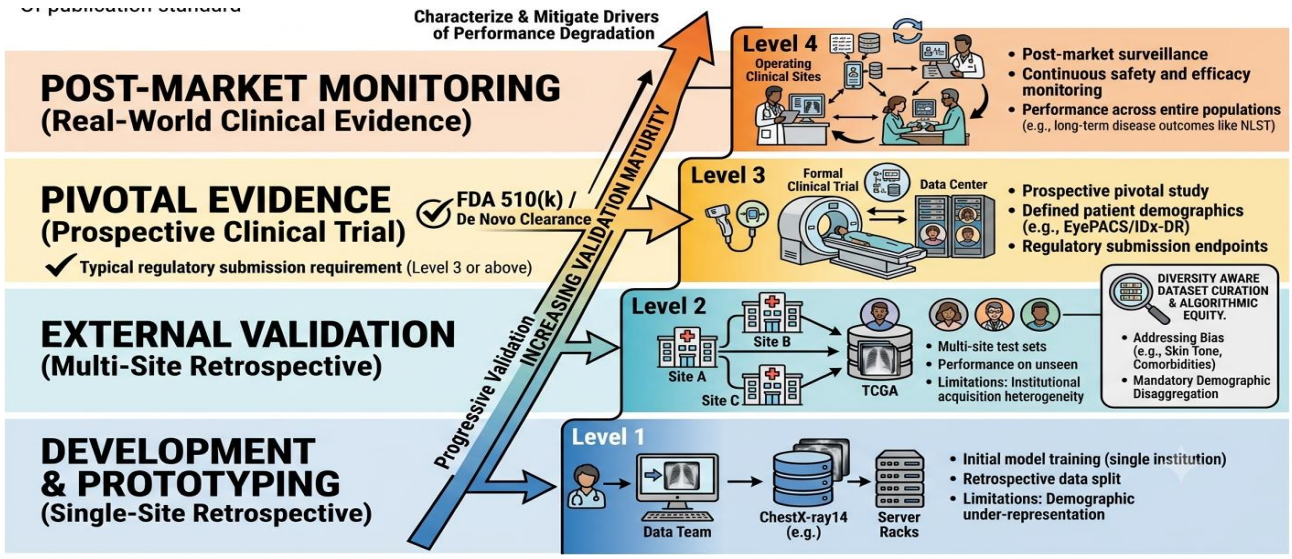

Figure 5: Clinical Validation Maturity Framework Aligned with Regulatory Expectations
Progressive ladder from single-site retrospective development (Level 1) to post-market real-world clinical evidence (Level 4), with regulatory checkpoints at each transition. FDA 510(k)/De Novo clearance typically requires evidence at Level 3 or above.

Taxonomy of AI Model Classes for Medical Imaging Diagnosis

To consolidate the multi-dimensional analysis presented in Sections 3 and 4, Table 3 provides a structured taxonomy of representative AI model classes organized along accuracy, interpretability, computational cost, data requirements, clinical deployment feasibility, and regulatory readiness. This table provides practical guidance for system designers by mapping algorithmic choices to clinical and regulatory constraints.

Table 3: Trade-Off Analysis of AI Model Classes for Medical Imaging Diagnosis

Model Type	Accuracy (Typical)	Interpretability	Computational Cost	Data Requirement	Deployment Feasibility	Regulatory Readiness	Citations
Classical ML (SVM, RF, LR, k-NN)	Moderate for well-structured radiomic features	High for LR/DT; moderate for SVM/RF with SHAP	Low	Small-medium labeled datasets; hand-crafted features	Suitable for resource-limited settings	Favorable; established validation frameworks	Shen et al. (2017);Doi (2007)
Standard CNN (ResNet, DenseNet, VGG)	High (AUC 0.90–0.99 in many tasks)	Low; improved with Grad-CAM/CAM	Moderate	Large annotated datasets; augmentation beneficial	Well-suited for cloud/server deployment	Challenging; requires XAI and multi-site validation	Litjens et al. (2017); Esteva et al. (2017)
Transformer / ViT / DINO	High to very high; strong on large-scale data	Moderate; attention visualization improving	High; self-supervised pretraining beneficial	Benefits from large pretraining corpora	Feasible for server; challenging for edge	Evolving; attention maps supportive of XAI	Moor et al. (2023); Chen et al. (2024)
Multimodal Fusion Network	Highest; consistently outperforms unimodal	Variable; cross-attention offers modality-level attribution	High; modality-specific encoders + fusion	All modalities required; missing data handling critical	Server/cloud; missing modality risk in clinical use	Requires evidence for each modality's contribution	Acosta et al. (2022); Cui et al. (2023)
XAI-Enhanced Models (Grad-CAM, SHAP, Prototype)	Same as underlying model	Explicitly improved; clinician-interpretable saliency maps	Slightly higher due to explanation overhead	Same as underlying + explanation validation data	Suitable for diagnostic support tools	Better positioned for regulatory approval	Singh et al. (2025); Tjoa & Guan (2021)
Federated Learning Models	Comparable to centralized; improves generalization	Inherits base model characteristics	High (communication overhead + local training)	Distributed; heterogeneous; privacy-preserving	Multi-site deployment; reduced data sharing burden	Emerging guidance; privacy-by-design supports compliance	Rieke et al. (2020); Mir et al. (2026)

Engineering Implications and Clinical Deployment Framework

Modality-Driven Model Selection

The synthesis demonstrates that model selection must be governed by imaging modality characteristics, clinical workflow requirements, and regulatory constraints rather than benchmark accuracy alone. For high-resolution 3D radiology data (CT, MRI), volumetric architectures combining 3D CNNs with attention mechanisms are preferred for capturing spatial context across slices. For chest X-ray and 2D radiology tasks, standard CNN-based transfer learning from large pretrained models (e.g., ImageNet-pretrained ResNet, CheXpert-pretrained encoders) provides strong baseline performance with manageable computational requirements (Rajpurkar et al., 2017).

For computational pathology, weakly supervised multiple instance learning frameworks are the preferred approach given the prohibitive cost of pixel-level annotation in gigapixel WSI. Attention-based MIL architectures provide both strong performance and clinically interpretable attention heatmaps that highlight diagnostically relevant tissue regions (Chen et al., 2022). For retinal imaging with high acquisition standardization, large-scale pretrained CNN and ViT models with fine-tuning provide reliable performance across demographic groups.

For multimodal fusion applications, the choice of fusion strategy must account for expected modality availability at deployment. Late fusion approaches are more robust to missing modalities but sacrifice the complementary learning benefits of early or intermediate fusion. Cross-attention mechanisms enabling modality-level attribution are particularly valuable for clinical interpretability and regulatory documentation (Steyaert et al., 2023).

Preprocessing and Data Quality Standards

Robust preprocessing is a prerequisite for reliable AI-based medical imaging diagnosis. Radiology preprocessing pipelines must include intensity normalization (e.g., Hounsfield unit windowing for CT, z-score normalization for MRI), resampling to consistent voxel spacing, and bias field correction for MRI. Prospective harmonization of acquisition parameters across sites reduces scanner variability at the source,

while retrospective harmonization methods (ComBat, DeepHarmony) address residual batch effects in training data (Pomponio et al., 2020).

Pathology preprocessing requires stain normalization (Macenko, Vahadane, or generative model-based approaches), tissue/background segmentation, and consistent magnification calibration across digitization platforms. Failure to standardize staining normalization is the most common source of computational pathology model failure in external validation (Srinidhi et al., 2021).

Data augmentation strategies must be clinically informed: augmentations that introduce unrealistic artifact patterns (e.g., extreme brightness shifts creating impossible CT Hounsfield values) may impair generalization rather than improve it. Clinically validated augmentation libraries tailored to specific modalities are an important component of reproducible pipeline design.

Multi-Level Clinical Validation Strategy

The findings emphasize that validation maturity is the primary determinant of clinical deployment readiness. High accuracy from single-site cross-validation is insufficient for clinical integration or regulatory submission. Instead, validation must progress systematically through multiple levels.

Level 1 (internal validation) establishes baseline performance and model selection using standard train/validation/test splits or nested cross-validation on institutional data. Level 2 (external validation) tests generalization on hold-out datasets from different institutions, scanner manufacturers, or patient demographics, providing the first evidence of real-world robustness. Level 3 (prospective reader study) positions AI as a clinical decision support tool evaluated against human readers in blinded controlled conditions, establishing the clinical utility argument required for regulatory submissions. Level 4 (post-market surveillance) continuously monitors performance in deployed systems, tracking for distribution shift, demographic performance disparities, and rare adverse events (Abramoff et al., 2023; Wu et al., 2021).

Computational Feasibility and Clinical Workflow Integration

Deployment feasibility is strongly influenced by computational requirements and integration complexity. Clinical AI systems typically operate in one of three deployment environments: cloud-based server processing (supporting complex multimodal fusion), local hospital server deployment (balancing performance with data governance), or edge/point-of-care deployment (requiring lightweight, compressed models for resource-limited settings).

Model compression through knowledge distillation, quantization, and architecture search enables deployment of high-accuracy models in resource-constrained environments without prohibitive performance sacrifice. A systematic approach emerging from the literature is a tiered architecture: lightweight, fast models provide immediate diagnostic support at the point of care, while complex ensemble or multimodal systems running on institutional servers perform deeper analysis for complex cases and second-read support.

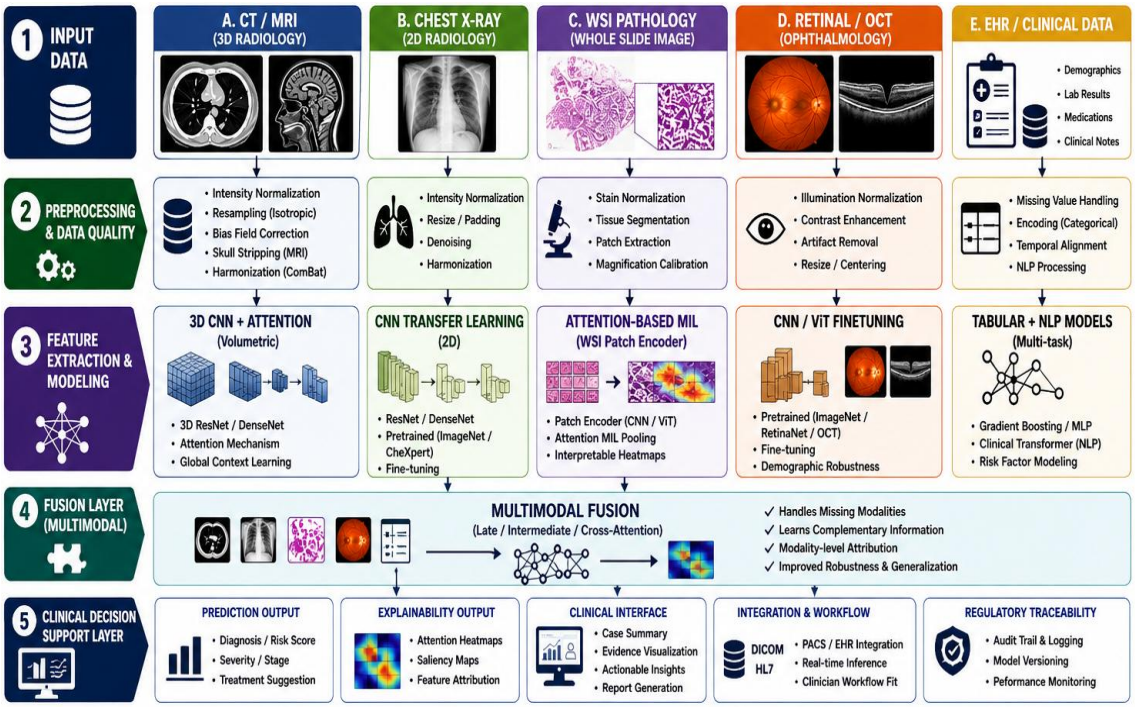

Figure 6: Modality-Aware AI Pipeline Architecture for Medical Imaging Diagnosis

End-to-end processing pipeline showing modality-specific preprocessing, feature extraction, and fusion pathways for CT/MRI, WSI pathology, retinal fundus/OCT, and EHR data, converging into a unified clinical decision support layer with explainability output and regulatory traceability.

Clinical Deployment Readiness Framework

To translate research findings into deployable clinical systems, a structured readiness assessment framework is proposed, comprising five evaluation dimensions:

1. Modality-Model Alignment: Appropriate architecture for imaging modality characteristics, data dimensionality, and clinical task complexity
2. Data Provenance and Diversity: Documentation of training data source, demographic coverage, scanner diversity, and class balance
3. Validation Maturity Level: Achieved level (1–4) with specific evidence quality assessments
4. Computational and Integration Feasibility: Inference latency, memory footprint, DICOM/HL7 integration compatibility, and clinician workflow fit
5. Interpretability and Regulatory Alignment: Availability of explainability outputs, demographic performance disaggregation, adverse event monitoring plan, and regulatory submission strategy

Each candidate system is evaluated across these five dimensions to determine deployment readiness level and identify specific gaps requiring further development before clinical integration or regulatory submission.

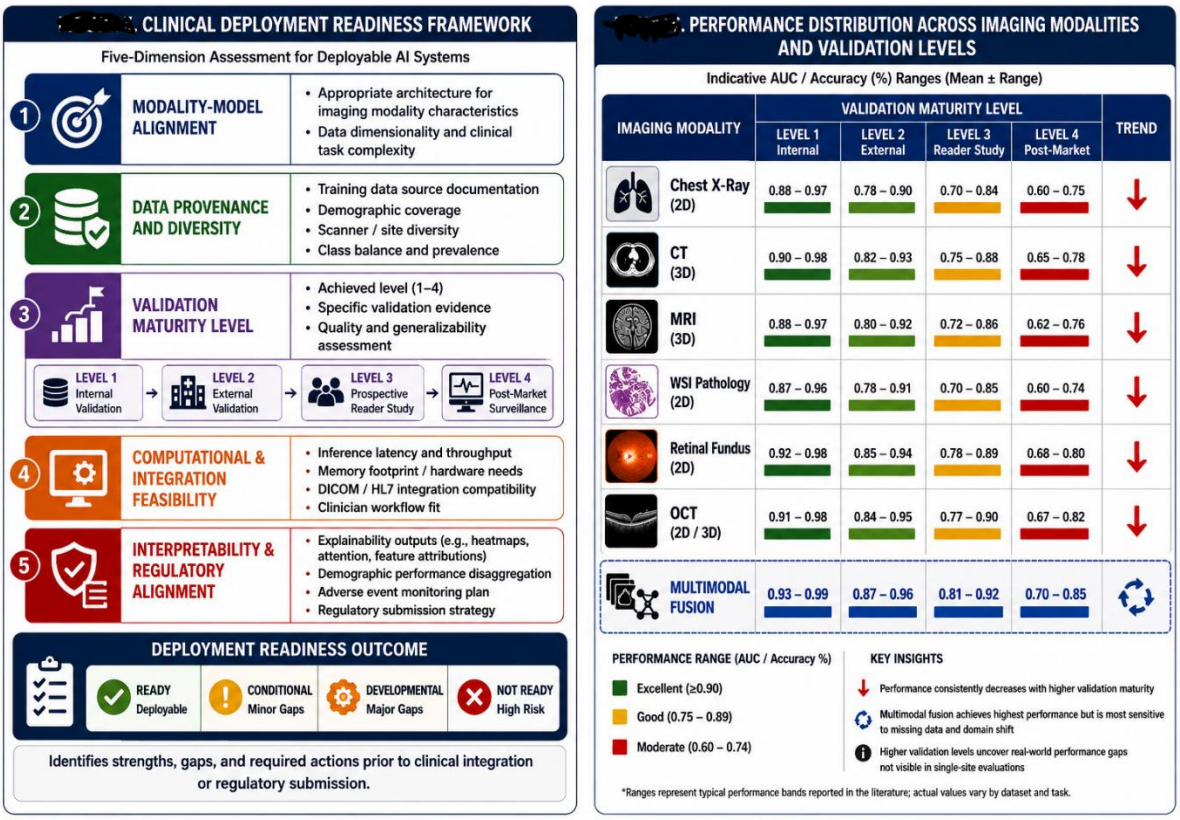

Figure 7: Performance Distribution Across Imaging Modalities and Validation Levels
Indicative AUC/accuracy ranges by modality and validation level. Higher validation maturity consistently reveals performance gaps not apparent in single-site evaluations, with multimodal fusion demonstrating the highest performance but also the highest sensitivity to missing data and domain shift.

Regulatory Perspectives and Certification Pathways

FDA Regulatory Framework for AI/ML Medical Devices

In the United States, AI-based medical imaging software is regulated as Software as a Medical Device (SaMD) under the FDA's device regulatory framework. As of 2023, the FDA had authorized over 692 AI/ML-enabled medical devices, representing a 20-fold increase from the 1995–2015 baseline, with radiology accounting for over 76% of clearances (Pierson et al., 2024; Wu et al., 2021). The primary regulatory pathways are 510(k) premarket notification (for devices substantially equivalent to a legally marketed predicate), De Novo classification (for novel low-to-moderate risk devices without predicates), and Premarket Approval (PMA) for high-risk devices.

The FDA's 2021 action plan for AI/ML-based SaMD introduced the concept of a Predetermined Change Control Plan (PCCP), allowing manufacturers to pre-specify the types of modifications an algorithm may undergo post-approval without requiring a new submission. This lifecycle-oriented approach is particularly important for adaptive algorithms that retrain on new data in deployment. The FDA also issued Draft Guidance in 2021 on transparency requirements, stipulating that device labeling must include meaningful information about algorithm inputs, outputs, performance characteristics, and intended use population (Wu et al., 2021).

For autonomous diagnostic AI systems, the FDA requires clinical validation studies demonstrating device performance against a clinical reference standard using data from the intended use population. Demographic subgroup analyses are increasingly required or strongly encouraged to identify disparate performance. Post-market surveillance mechanisms, including adverse event reporting and performance monitoring protocols, must be defined at submission (Abramoff et al., 2023).

EU MDR/IVDR and CE Marking

In the European Union, AI-based medical imaging devices are regulated under the Medical Device Regulation (MDR 2017/745) and the In Vitro Diagnostic Regulation (IVDR 2017/746). The EU MDR introduces risk-based classification with Class IIa and IIb devices (including most diagnostic AI software) requiring conformity assessment by a Notified Body. Clinical evidence requirements are more stringent than under the previous Medical Device Directive, requiring demonstration of clinical performance through clinical investigations or well-designed literature synthesis.

The European AI Act, adopted in 2024, classifies medical AI systems used for decision-making affecting patient health as high-risk AI systems subject to mandatory conformity assessment, transparency requirements, human oversight provisions, and post-market monitoring obligations. This creates a dual regulatory burden for medical AI developers who must comply with both MDR and AI Act requirements in the EU market.

Explainability as a Regulatory Requirement

Both FDA and EU regulatory frameworks increasingly emphasize explainability as a component of AI device trustworthiness. The FDA's transparency workshop established that device labeling should help clinicians understand AI outputs, recognize when outputs may be unreliable, and maintain appropriate oversight. Radiologist professional societies have emphasized that AI systems deployed as clinical decision support must provide interpretable outputs enabling radiologists to contextualize and challenge AI recommendations (RSNA Comments, FDA Docket 2024).

Regulatory bodies have not mandated specific explainability methods, recognizing that the appropriate form of explanation depends on the device function and clinical context. However, post hoc attribution methods (Grad-CAM, SHAP, Integrated Gradients) and inherently interpretable architectures are increasingly expected as components of the design rationale in regulatory submissions (Singh et al., 2025).

Table 4: Synthesis of Research Gaps and Open Challenges in AI-Based Medical Imaging Diagnosis

Category	Observed Issue	Evidence from Literature	Impact on Clinical Systems	Suggested Research Direction	Citations
Data	Single-site or homogeneous demographic training data	Model performance degradation at external sites documented in radiology and pathology	Unreliable deployment; potential patient harm in under-served populations	Federated learning; diversity-aware dataset curation; mandatory demographic disaggregation	Zech et al. (2018); Pierson et al. (2024); Rieke et al. (2020)
Data	Missing modality handling in multimodal fusion	Clinical practice rarely has all modalities for all patients simultaneously	System failures or degraded performance in realistic deployment	Robust missing modality imputation; uncertainty-aware inference	Steyaert et al. (2023); Cui et al. (2023)
Model	Interpretability gap between black-box DL and clinical needs	XAI techniques improve but do not fully satisfy clinical traceability requirements	Regulatory challenges; reduced clinician trust; liability concerns	Co-design of models and explanation mechanisms; clinician-validated saliency frameworks	Singh et al. (2025); Tjoa & Guan (2021); RSNA Comments (2024)

Deployment	Gap between offline metrics and prospective clinical performance	Few studies achieve Level 3 or 4 validation; reader studies show smaller gaps than expected	Uncertain clinical benefit; regulatory evidence insufficient	Prospective reader studies; pragmatic RCTs; post-market performance monitoring	Topol (2019); Abramoff et al. (2023); Wu et al. (2021)
Certification	Absence of standardized AI performance benchmarks for regulatory use	Different regulatory bodies use different evidence standards; no universally accepted benchmark datasets	Inconsistent approval processes; difficulty comparing device safety across markets	International harmonization of AI benchmark datasets; regulatory sandbox programs	Wu et al. (2021); Abramoff et al. (2023)
Equity	Demographic performance disparities in deployed systems	Low demographic reporting in FDA-cleared devices; evidence of race/ethnicity-based performance gaps	Exacerbation of health disparities; regulatory and liability exposure	Fairness-constrained training; mandatory subgroup reporting; diverse recruitment in clinical trials	Obermeyer et al. (2019); Pierson et al. (2024)

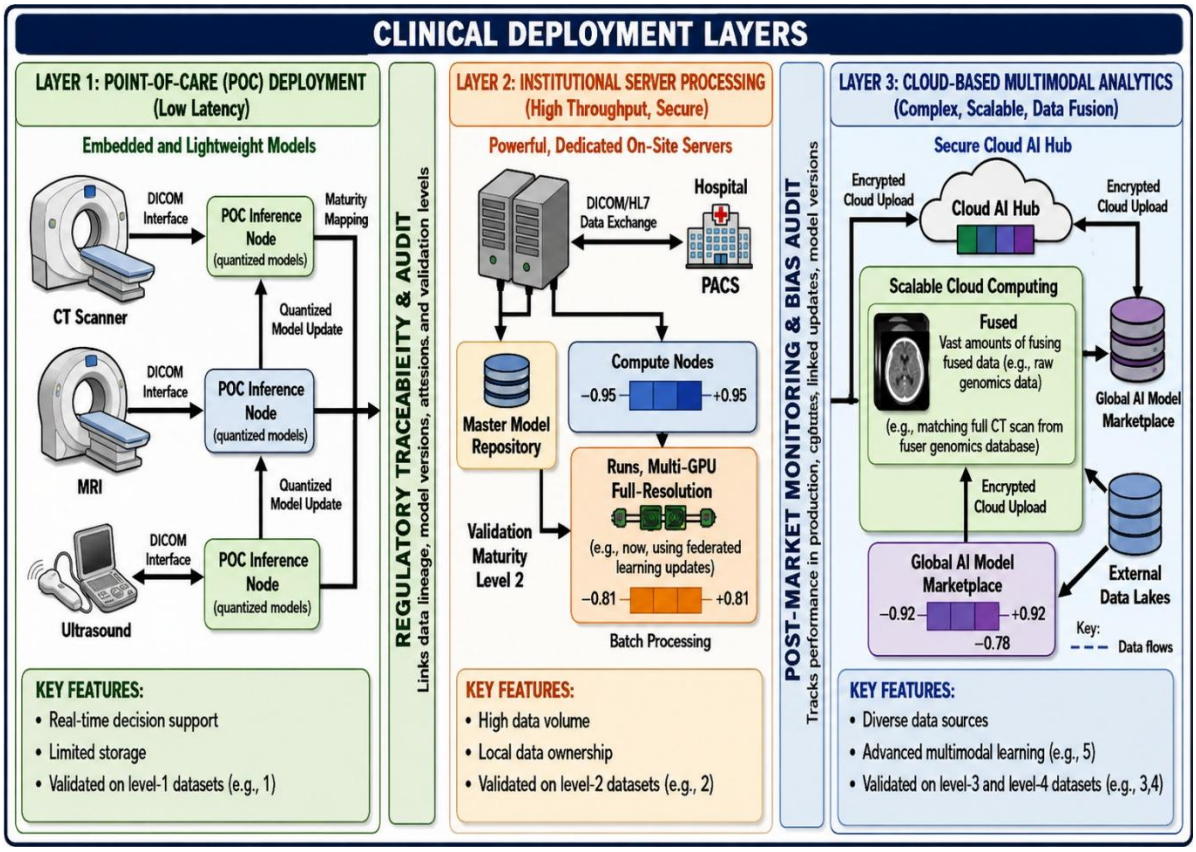

Figure 8: Clinical Deployment Architecture for AI-Based Medical Imaging Systems
Layered deployment architecture distinguishing point-of-care lightweight inference, institutional server processing, and cloud-based multimodal analytics, with regulatory traceability and post-market monitoring pathways integrated across all layers.

Future Research Directions

Foundation Models and Generalist Medical AI

Foundation models pretrained on large-scale multimodal medical datasets represent a transformative direction for clinical AI. Models such as MedPaLM 2, BioViL-T, and computational pathology foundation models have demonstrated strong generalization across tasks and institutions with minimal task-specific fine-tuning (Moor et al., 2023; Chen et al., 2024). Future research should focus on rigorous prospective evaluation of foundation model generalization across diverse patient populations and healthcare systems, with explicit attention to demographic equity and rare disease performance.

Federated Learning and Privacy-Preserving Training

Federated learning offers a principled solution to the generalization challenge by enabling collaborative model training across multiple institutions without sharing patient data. Systematic reviews confirm that federated approaches achieve comparable performance to centralized training while providing significant privacy and data governance benefits (Rieke et al., 2020; Mir et al., 2026). Future directions include differential privacy mechanisms for stronger privacy guarantees, vertical federated learning for multi-modal fusion across institutions holding different modalities for the same patients, and federated fine-tuning of foundation models.

Clinically Integrated Explainability

Current XAI methods in medical imaging, while providing visual attribution maps, do not yet meet the full interpretability requirements of clinical reasoning or regulatory decision traceability. Future research should develop explanation frameworks that are clinically validated through user studies, produce quantified uncertainty alongside feature attributions, and are integrated into structured reporting workflows rather than appended as standalone visualizations. The co-design of AI models and clinical explanation interfaces, involving radiologists, pathologists, and patient advocates in the design process, is essential for meaningful clinical adoption (Singh et al., 2025).

Prospective Multicenter Trials and Post-Market Surveillance

The field urgently needs more Level 3 and Level 4 validation evidence. Pragmatic randomized controlled trials evaluating AI-assisted diagnosis against standard of care, with patient outcome endpoints rather than purely diagnostic accuracy metrics, are the gold standard for demonstrating clinical utility. Post-market surveillance infrastructure, enabling continuous monitoring of AI device performance across demographic groups and clinical settings following deployment, is essential for maintaining regulatory compliance and detecting performance drift as patient populations and clinical protocols evolve.

Digital Biomarkers and Multimodal Integration

The integration of imaging-derived digital biomarkers with molecular, genomic, and wearable sensor data represents a frontier for personalized diagnostics and treatment monitoring. AI frameworks capable of synthesizing these heterogeneous data streams, accounting for temporal dynamics and missing modalities, will enable precision medicine applications not achievable from any single modality (Acosta et al., 2022; Lipkova et al., 2022). Standardized digital biomarker ontologies and interoperable data exchange frameworks are prerequisites for scaling these approaches across healthcare systems.

CONCLUSION

This survey has provided a structured, modality-aware synthesis of AI-based medical imaging diagnosis, demonstrating both the significant progress achieved and the persistent limitations constraining clinical translation. By organizing the literature along imaging modality, data provenance, validation maturity, and model architecture, the analysis reveals that high reported accuracy is strongly contingent on evaluation conditions, with many systems lacking the multicenter prospective validation and regulatory evidence required for reliable clinical deployment.

The primary contribution is the translation of this structured analysis into a clinically oriented deployment framework. By explicitly linking model selection to modality characteristics, preprocessing to data quality requirements, and validation strategy to regulatory expectations, the work provides a practical pathway from algorithm development to regulatory-compliant clinical systems. The deployment readiness framework enables systematic gap analysis for candidate systems across five key dimensions.

Looking forward, advances in foundation models, federated learning, clinically integrated explainability, and multimodal digital biomarker integration hold transformative potential for AI-based medical imaging. Realizing this potential will require coordinated efforts to develop diverse and representative benchmark datasets, bridge the generalization gap through multi-institutional collaboration, design interpretable and equitable models, and align AI development lifecycles with evolving regulatory frameworks. With these developments, AI-based medical imaging can transition from high-performing experimental systems to reliable, certified components of next-generation precision diagnostic infrastructure—bridging the gap between algorithmic accuracy and real-world clinical reliability.

REFERENCES

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

Abramoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., ... & Topol, E. J. (2023). Considerations for addressing bias in artificial intelligence for health equity. NPJ Digital Medicine, 6(1), 170. https://doi.org/10.1038/s41746-023-00913-9

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Albadawy, E. A., Saha, A., & Mazurowski, M. A. (2018). Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Medical Physics, 45(3), 1150–1158. https://doi.org/10.1002/mp.12752

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., ... & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digital Medicine, 2(1), 31. https://doi.org/10.1038/s41746-019-0105-1

Bilal, A., Imran, A., Baig, T. I., Liu, X., Long, H., Alzahrani, A., & Shafiq, M. (2024). DeepSVDNet: A deep learning-based approach for detecting and classifying vision-threatening diabetic retinopathy in retinal fundus images. Computer Systems Science and Engineering, 48(2), 511–528.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., ... & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3

Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., ... & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell, 40(8), 865–878. https://doi.org/10.1016/j.ccell.2022.07.004

Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., ... & Huo, Y. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering, 5(2), 022001. https://doi.org/10.1088/2516-1091/acc2fe

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics, 31(4–5), 198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002

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

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

He, J., Wang, J., Han, Z., Ma, J., Wang, C., & Qi, M. (2023). An interpretable transformer network for the retinal disease classification using optical coherence tomography. Scientific Reports, 13(1), 3637. https://doi.org/10.1038/s41598-023-30853-z

Larson, D. B., Harvey, H., Rubin, D. L., Irani, N., Justin, R. T., & Langlotz, C. P. (2021). Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. Journal of the American College of Radiology, 18(3), 413–424. https://doi.org/10.1016/j.jacr.2020.09.060

Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., ... & Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology. Cancer Cell, 40(10), 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (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

Lu, M. Y., Chen, R. J., Kong, D., Lipkova, J., Singh, R., Williamson, D. F. K., ... & Mahmood, F. (2022). Federated learning for computational pathology on gigapixel whole slide images. Medical Image Analysis, 76, 102298. https://doi.org/10.1016/j.media.2021.102298

McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi:10.1136/bmj.n71. http://www.prisma-statement.org

Mir, B. A., Abbas, S. R., & Lee, S. W. (2026). Federated learning in healthcare ethics: A systematic review of privacy-preserving and equitable medical AI. Healthcare, 14(3), 306. https://doi.org/10.3390/healthcare14030306

Moor, M., Banerjee, O., Abad, Z. S. H., Krefl, H. M., Kahn, J. M., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

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

Park, Y. W., Park, J. E., Ahn, S. S., Han, K., Kim, N., Oh, J. Y., ... & Lee, S. K. (2024). Deep learning-based metastasis detection in patients with lung cancer to enhance reproducibility and reduce workload in brain metastasis screening with MRI: A multi-center study. Cancer Imaging, 24(1), 38. https://doi.org/10.1186/s40644-024-00669-9

Pierson, E., Bhatt, M., Bhatt, S., ... & Rajpurkar, P. (2024). Machine learning-enabled medical devices authorized by the US Food and Drug Administration in 2024: Regulatory characteristics, predicate lineage, and transparency reporting. NPJ Digital Medicine, (in press). PMC12730494.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., ... & Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage, 208, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

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

Roest, C., Fransen, S. J., Kwee, T. C., & Yakar, D. (2022). Comparative performance of deep learning and radiologists for the diagnosis and localization of clinically significant prostate cancer at MRI: A systematic review. Life, 12(10), 1490. https://doi.org/10.3390/life12101490

RSNA Comments, FDA Docket FDA-2024-D-4488. (2025, April 7). Comments on AI-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations [Policy/gray literature].

Saha, A., Bosma, J. S., Twilt, J. J., van Ginneken, B., Bjartell, A., Padhani, A. R., ... & Litjens, G. (2024). Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. The Lancet Oncology, 25(7), 879–887. https://doi.org/10.1016/S1470-2045(24)00235-9

Salim, M., Liu, Y., Sorkhei, M., Ntoula, D., Foukakis, T., Fredriksson, I., ... & Strand, F. (2024). AI-based selection of individuals for supplemental MRI in population-based breast cancer screening: The randomized ScreenTrustMRI trial. Nature Medicine, 30(9), 2623–2630. https://doi.org/10.1038/s41591-024-03093-5

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

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., ... & Shen, D. (2022). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering, 14, 4–15. https://doi.org/10.1109/RBME.2020.2987975

Singh, Y., Hathaway, Q. A., Keishing, V., Salehi, S., Wei, Y., Horvat, N., ... & Andersen, J. B. (2025). Beyond post hoc explanations: A comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering, 12(8), 879. https://doi.org/10.3390/bioengineering12080879

Srinidhi, C. L., Ciga, O., & Martel, A. L. (2021). Deep neural network models for computational histopathology: A

survey. Medical Image Analysis, 67, 101813. https://doi.org/10.1016/j.media.2020.101813

Steyaert, S., Pizurica, M., Ye, H., Bhatt, P., Shen, J. Y., Wan, X., ... & Swaminathan, M. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence, 5(4), 351–362. https://doi.org/10.1038/s42256-023-00633-5

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

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

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., & Zou, J. (2021). How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. Nature Medicine, 27(4), 582–584. https://doi.org/10.1038/s41591-021-01312-x

Xu, Y., Khan, T. M., Song, Y., & Meijering, E. (2025). Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. Artificial Intelligence Review, 58(3), 93. https://doi.org/10.1007/s10462-024-11033-5

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

Zhang, X., Li, Q., Yu, H., et al. (2026). AI framework for multidisease detection via retinal imaging. Nature Medicine. https://doi.org/10.1038/s41591-026-04359-w