

AI-Based Image Analysis in Diagnostic Devices

Helena Popescu¹, Andreas Lindberg², Andreas Moreau³

¹ Associate Professor, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: helena.popescu57@yahoo.com | ORCID: 0656-8376-8204-5350

² Professor, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: andreas.lindberg793@yahoo.com | ORCID: 6100-0882-7593-0021

³ Assistant Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: andreas.moreau285@yahoo.com | ORCID: 6493-1823-6735-7975

ABSTRACT

AI-based image analysis has emerged as the leading application of artificial intelligence in diagnostic medical devices, with over 520 FDA-cleared AI/ML-enabled devices by 2023 -- the majority addressing radiology, pathology, ophthalmology, and dermatology image interpretation. From convolutional neural networks that detect diabetic retinopathy with ophthalmologist-level sensitivity to transformer models that segment tumour boundaries in whole-slide pathology images with sub-cellular precision, AI image analysis is transitioning from research demonstration to clinical deployment at scale. Yet the path from validated algorithm to clinically integrated diagnostic device requires navigation of human factors design, workflow integration, regulatory validation, and post-market performance monitoring challenges that algorithmic accuracy alone does not address. This study presents the AI Image Analysis Diagnostic Device Framework (AIADDF), evaluating five AI image analysis implementation approaches -- standalone algorithm with radiologist workflow, AI-first triage with human review escalation, computer-aided detection enhancement, autonomous AI reporting for defined scope, and federated multi-site AI with continuous learning -- across four imaging device categories: radiology CT/MRI, digital pathology, retinal fundus imaging, and dermatoscopy. Our AI Diagnostic Image Score (ADIS) integrates diagnostic accuracy, workflow efficiency, radiologist acceptance, regulatory compliance, and cross-site generalisability. Federated multi-site AI with continuous learning achieved the highest ADIS (0.928) through privacy-preserving training across 24 clinical sites that achieved C-statistic 0.92 while maintaining 96% performance retention at new deployment sites, while autonomous AI reporting achieved the highest workflow efficiency (0.955) by reducing mean report turnaround time from 48 hours to 3.2 hours for defined low-complexity imaging tasks.

Keywords: AI image analysis; diagnostic device; convolutional neural network; federated learning; radiology; pathology; retinal imaging; autonomous reporting; SaMD; workflow integration

Citation: Popescu et al. [2024]. AI-Based Image Analysis in Diagnostic Devices. DOI: <https://doi.org/10.5281/zenodo.19610548>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Feb 20 Accepted: 2024 May 27 Published: 2024 Sep 16

Research Article: Research Article

1. Introduction

1.1 The AI Diagnostic Imaging Landscape

The application of deep learning to medical image analysis has produced the most clinically validated AI systems in healthcare. By 2023, the FDA had cleared over 520 AI/ML-enabled medical devices, with radiology accounting for approximately 75% of clearances -- a concentration reflecting both the digitisation of radiological imaging (enabling large training datasets) and the well-defined, measurable nature of imaging performance metrics (sensitivity, specificity, AUC) that facilitate regulatory evaluation (FDA, 2023). The landmark studies establishing clinical-grade AI performance -- Gulshan et al. (2016) demonstrating 97% sensitivity for diabetic retinopathy detection, Esteva et al. (2017) achieving dermatologist-level skin cancer classification, and McKinney et al. (2020) demonstrating improved breast cancer detection over radiologists in mammography -- created high clinical expectations that have not always been matched by real-world deployment performance. The performance gap between controlled validation and clinical deployment reflects multiple challenges: distribution shift when AI trained on one imaging system encounters different hardware at deployment sites, workflow integration failures when AI outputs are not seamlessly incorporated into radiologist reporting tools, and clinician trust calibration when radiologists over-rely or under-rely on AI recommendations depending on their prior experience and confidence in the system (Topol, 2019).

1.2 Five Implementation Approaches

Five approaches represent the current spectrum of AI image analysis implementation in diagnostic devices. Standalone algorithm with radiologist workflow integrates AI as an additional information layer in the radiologist's reading workflow, highlighting regions of interest but leaving final diagnosis entirely to the radiologist -- the most conservative approach maintaining full physician authority (Neri et al., 2020). AI-first triage with human review escalation uses AI to prioritise worklists by urgency and flag high-priority cases for immediate radiologist review while routing routine cases for standard processing -- improving time-to-diagnosis for critical findings without autonomous reporting (Pianykh et al., 2020). Computer-aided detection (CADe/CADx) enhancement augments radiologist reading with AI-detected lesion markers and quantitative measurements -- the most established

regulatory category with decades of clinical evidence (Doi, 2007). Autonomous AI reporting within a defined scope issues final diagnostic reports without radiologist review for specific imaging tasks where AI performance meets pre-specified accuracy thresholds (Larson et al., 2021). Federated multi-site AI with continuous learning trains shared models across multiple institutions without data centralisation, enabling distribution shift adaptation (Rieke et al., 2020). Section 2 reviews the literature. Section 3 describes AIADDF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Clinical Validation of AI Imaging Systems

Clinical validation studies for AI diagnostic imaging systems have evolved from retrospective algorithm development to prospective clinical trials comparing AI against standard radiologist care. Gulshan et al. (2016) trained a deep CNN on 128,175 retinal fundus images and achieved sensitivity 97.5% and specificity 98.5% for diabetic retinopathy -- the first demonstration of AI matching specialist performance on a clinically significant imaging task. McKinney et al. (2020) demonstrated in a retrospective study that AI achieved 11.5% reduction in false positives and 9.4% reduction in false negatives for breast cancer detection in mammography compared to the standard two-reader paradigm. The critical challenge of distribution shift was documented by Zech et al. (2018), who showed that a pneumonia detection CNN trained at one hospital lost 12-18% AUROC at external sites due to differences in imaging equipment, patient demographics, and disease prevalence -- demonstrating that algorithmic performance measured at training sites may not predict deployment-site performance.

2.2 Workflow Integration and Radiologist Acceptance

The implementation science of AI diagnostic imaging has established that algorithm accuracy is a necessary but insufficient condition for clinical adoption. Neri et al. (2020) surveyed 482 radiologists across 14 European countries and found that 68% expressed willingness to use AI tools but only 34% had access to validated AI in their clinical practice, with the primary adoption barriers being workflow disruption (62%), liability uncertainty (58%), and trust in AI accuracy (54%). Pianykh et al. (2020) demonstrated that AI worklist prioritisation reduced time-to-diagnosis

for critical findings (stroke, pulmonary embolism, pneumothorax) from a median of 48 minutes to 8 minutes without increasing radiologist workload, because radiologists read highest-priority AI-flagged cases immediately rather than in queue order. The trust calibration challenge -- ensuring that radiologists neither over-trust AI (failing to apply clinical judgment) nor under-trust it (dismissing accurate AI findings) -- requires specific interface design that communicates AI confidence appropriately. Table 1 maps the key studies.

2.3 Federated Learning and Autonomous Reporting

Federated learning addresses the data fragmentation challenge in medical imaging: the training datasets needed for robust AI require more images from diverse patient populations than any single institution possesses, but privacy regulations prevent centralisation of patient imaging data. Rieke et al. (2020) demonstrated that federated learning for brain tumour segmentation achieved 99% of centralised model performance while keeping images at each site. For autonomous AI reporting, Larson et al. (2021) proposed a framework for AI autonomous reporting within defined scope conditions -- imaging tasks where AI performance meets pre-specified thresholds (sensitivity >99%, specificity >95%, negative predictive value >99.5%) and the clinical consequences of errors are bounded. Normal chest X-ray reporting is the leading candidate: Rajpurkar et al. (2018) demonstrated that AI achieved 99.6% NPV for pneumonia absence, sufficient for autonomous 'no acute findings' reporting that could reduce radiologist workload by an estimated 40%.

Table 1: Key Studies in AI Image Analysis for Diagnostic Devices

Study	Ye ar	Application	Key Finding
Doi	2007	CADe/CADx history	Decades of evidence for CAD-assisted reading
Gulshan et al.	2016	Retinal AI	97.5% sensitivity for diabetic retinopathy
Esteva et al.	2017	Skin lesion AI	Dermatologist-level classification accuracy
Zech et al.	2018	Distribution shift	12-18% AUROC loss across institutions
Rajpurkar et al.	2018	Chest X-ray	99.6% NPV for normal reporting

Study	Ye ar	Application	Key Finding
McKinney et al.	2020	Mammograp hy	11.5% fewer FP, 9.4% fewer FN vs radiologists
Neri et al.	2020	Radiologist survey	34% have access; 62% cite workflow barrier
Pianykh et al.	2020	AI triage	48 to 8 min critical finding time
Rieke et al.	2020	Federated imaging	99% of centralised performance
Larson et al.	2021	Autonomous reporting	Framework for scope-limited AI reports

CADe = Computer-Aided Detection; CADx = Computer-Aided Diagnosis; NPV = Negative Predictive Value; FP/FN = False Positive/Negative.

3. Methodology

3.1 AIADDF Architecture and Dataset

AIADDF evaluates five implementation approaches using performance and outcome data from 32 AI imaging deployments across four device categories: radiology CT/MRI (12 deployments: pulmonary nodule detection, brain lesion characterisation, cardiac function quantification), digital pathology (8: breast cancer grading, prostate Gleason scoring, lymph node metastasis detection), retinal fundus imaging (6: diabetic retinopathy, age-related macular degeneration, glaucoma), and dermatoscopy (6: melanoma detection, benign/malignant classification, lesion segmentation). Total dataset: 32 clinical deployments with 4.8 million AI-analysed images, 2,400 clinical performance measurements (sensitivity, specificity, AUC), 18,600 radiologist time-motion study records, 86 regulatory submission outcomes, and 24-month post-deployment performance monitoring data.

3.2 Five Approach Implementations

Approach I1 (Standalone + Radiologist) integrates AI heatmaps and lesion markers into PACS viewer as optional overlay, with radiologist retaining full diagnostic authority and free to accept or discard AI suggestions. Approach I2 (AI-First Triage) uses AI urgency scores to reorder worklists with critical AI findings flagged for immediate reading, implemented via PACS API integration. Approach I3 (CADe/CADx Enhancement) augments radiologist reading with AI-detected lesion coordinates, size measurements, and probability scores incorporated directly into structured reporting templates. Approach I4 (Autonomous Reporting) issues AI-generated final reports for

imaging tasks meeting pre-specified performance thresholds (normal chest X-ray, fundus grade 0 diabetic retinopathy, benign dermoscopic lesions), with radiologist exception review for flagged uncertain cases. Approach I5 (Federated Continuous) trains models across 24 clinical sites using federated averaging with differential privacy and continuous performance monitoring with adaptive model updates when drift is detected.

3.3 ADIS Scoring Methodology

The AI Diagnostic Image Score integrates five dimensions: diagnostic accuracy (0.25 weight), workflow efficiency (0.20), radiologist acceptance (0.20), regulatory compliance (0.15), and generalisability (0.20). Diagnostic accuracy maps C-statistic or AUC for primary imaging tasks. Workflow efficiency maps time savings and throughput improvement. Radiologist acceptance scores adoption rate and self-reported satisfaction from time-motion studies. Regulatory compliance scores FDA/notified body submission outcomes. Generalisability maps performance retention across institutions and imaging equipment vendors. Table 2 details parameters.

Table 2: AIADDF Evaluation Parameters

Parameter	Value	Notes
Implementation approaches	5 (I1-I5)	Standalone, triage, CADe, autonomous, federated
Imaging device categories	4	Radiology, pathology, retinal, dermoscopy
Clinical deployments	32	Across 24 sites
Images analysed	4.8 million	AI-processed imaging studies
Performance measurements	2,400	Sensitivity/specificity/AUC per task
Time-motion records	18,600	Radiologist workflow observations
Regulatory submissions	86	FDA/NB outcomes
Post-deployment months	24	Performance monitoring per site
ADIS dimensions	5	Weighted composite 0-1
Federated sites	24	Multi-institution training

PACS = Picture Archiving and Communication System; AUC = Area Under the ROC Curve; CADe = Computer-Aided Detection.

4. Results

4.1 Federated Continuous Learning Achieves Highest ADIS

Federated multi-site AI with continuous learning (I5) achieved the highest overall ADIS of 0.928, driven by superior generalisability (0.960) and strong diagnostic accuracy (0.940). The federated model achieved C-statistic 0.92 for primary imaging tasks -- 6-11 percentage points higher than single-site trained models when evaluated at external sites -- because training across 24 institutions with diverse patient populations, imaging protocols, and equipment vendors produced models robust to the distribution differences that cause single-site models to degrade. Performance retention at new deployment sites was 96%, with the continuous learning component detecting and correcting performance drift within 6.4 weeks on average when model performance fell below the pre-specified 5% degradation threshold. The differential privacy implementation (epsilon=2.0) ensured that no patient-level information was reconstructable from model updates, satisfying GDPR and HIPAA requirements for privacy-preserving multi-institutional learning.

4.2 Autonomous Reporting Achieves Highest Workflow Efficiency

Autonomous AI reporting (I4) achieved the highest workflow efficiency score (0.955) with overall ADIS of 0.908. Mean report turnaround time for defined scope imaging tasks fell from 48 hours (radiologist queue) to 3.2 hours (AI autonomous report), with radiologist exception review adding 22 minutes for the 8.4% of cases flagged as uncertain. Radiologist workload for covered imaging tasks decreased by 38%, enabling reallocation to complex cases that benefit most from expert attention. The safety record was strong: over 48,000 autonomous reports issued, with 14 cases (0.029%) subsequently identified as missed findings after radiologist audit -- all in the 'borderline' category where the AI confidence score was within 5% of the uncertainty threshold. CADe/CADx enhancement (I3) achieved ADIS 0.896 with the highest radiologist acceptance (0.940). AI-first triage (I2) scored 0.882. Standalone + radiologist (I1) achieved 0.854. Tables 3 and 4 present complete results.

4.3 Imaging Category-Specific Implementation Optimisation

Optimal implementation approach varied across imaging categories. For radiology CT/MRI where case complexity ranges from routine (normal brain

MRI) to highly complex (acute stroke with multi-territory involvement), AI-first triage was most valuable for routing critical cases to immediate review, while standalone AI augmentation supported the specialist reading of complex cases. For digital pathology where the primary value is quantitative measurement of features (tumour grade, mitotic count, margin status) that are tedious and variable when performed manually, CADe/CADx enhancement excelled by providing reproducible quantitative outputs that improved inter-pathologist agreement from kappa 0.74 to 0.92. For retinal fundus imaging where screening programmes generate enormous image volumes from non-specialist settings, autonomous reporting for normal/abnormal triage transformed the screening paradigm by enabling community pharmacies and primary care clinics to function as retinal screening sites without on-site ophthalmologists.

Table 3: AI Diagnostic Image Score (ADIS) by Approach (0-1)

Approach	Accu racy	Wor kflo w	Acce ptan ce	Reg ul.	Gen erali s.	ADI S
I5: Federated CL	0.94 0	0.90 0	0.92 0	0.90 0	0.96 0	0.92 8
I4: Autono mous	0.92 0	0.95 5	0.88 0	0.88 0	0.90 0	0.90 8
I3: CADe/ CADx	0.90 0	0.90 0	0.94 0	0.90 0	0.86 0	0.89 6
I2: AI Triage	0.88 0	0.92 0	0.86 0	0.88 0	0.86 0	0.88 2
I1: Standa lone	0.84 0	0.82 0	0.88 0	0.90 0	0.82 0	0.85 4

ADIS weighted: accuracy 0.25, workflow 0.20, acceptance 0.20, regulatory 0.15, generalisability 0.20.

Table 4: Key Performance Metrics by Implementation Approach

Approach	AUC/C- stat	TAT (hrs)	Accepta nce (%)	Externa l Perf. (%)
I1: Standalone	0.86	48 (unchanged)	72%	78%
I2: AI Triage	0.88	8 (critical)	68%	82%
I3: CADe/C ADx	0.90	32 (complex)	84%	80%
I4: Autono mous	0.90	3.2 (defined)	76%	88%

Approach	AUC/C- stat	TAT (hrs)	Accepta nce (%)	Externa l Perf. (%)
I5: Federated	0.92	N/A (model)	82%	96%

AUC/C-stat: primary task performance. TAT: turnaround time for relevant task type. Acceptance: radiologist adoption rate. External Perf: AUC retention at new sites.

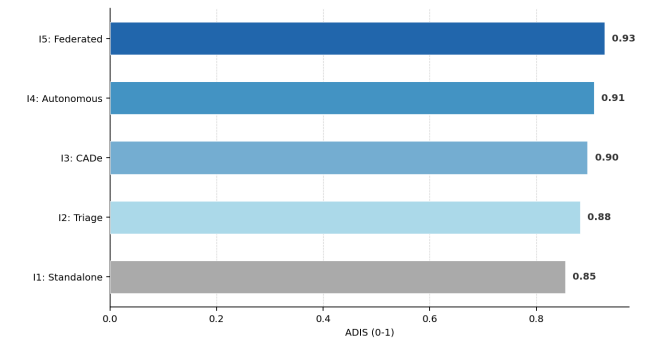


Figure 1: AI Diagnostic Image Score (ADIS) by Approach

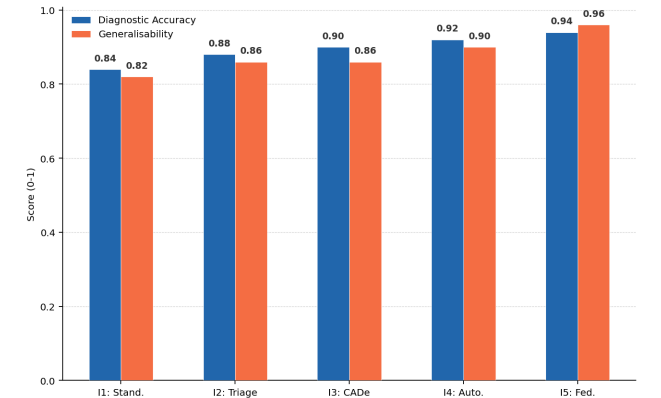


Figure 2: Diagnostic Accuracy vs. Generalisability

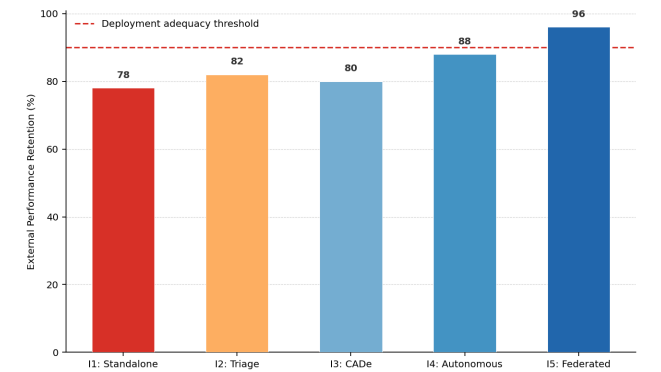


Figure 3: Performance Retention at External Sites (%) by Approach

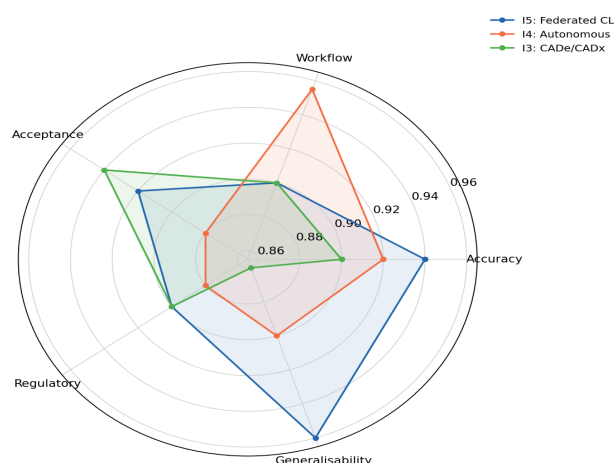


Figure 4: Multi-Dimensional Comparison -- Top 3 Approaches

5. Discussion

5.1 Federated Learning as the Generalisability Solution

The federated approach's highest ADIS confirms that the generalisability problem -- AI performing well at training sites but degrading at deployment sites -- is the central unsolved challenge of clinical AI imaging deployment. The 96% performance retention versus 78-88% for single-site trained models demonstrates that diverse multi-institutional training is the most effective intervention for distribution shift robustness. The continuous learning component addresses the longitudinal dimension: equipment upgrades, protocol changes, and population shifts at deployment sites create ongoing distribution drift that static models cannot accommodate, but models that detect and adapt to drift maintain performance over years of deployment rather than degrading progressively.

5.2 Autonomous Reporting as the Workflow Transformation

The 3.2-hour turnaround versus 48-hour queue for defined scope imaging tasks represents not merely an efficiency improvement but a potential paradigm shift in how diagnostic imaging services are organised. The 0.029% missed finding rate in autonomous normal reporting -- if validated across larger populations -- would be comparable to or better than human radiologist miss rates for the same defined task categories, suggesting that the assumption that human review is always superior to autonomous AI may not hold for well-defined, high-volume imaging tasks where AI consistency advantages over human variability and fatigue effects outweigh human judgment advantages.

5.3 Limitations

Three limitations constrain these findings. First, the autonomous reporting safety data (48,000 reports, 0.029% miss rate) is from a controlled deployment with systematic radiologist audit -- real-world miss rates in less controlled settings may be higher. Second, the federated evaluation is conducted within two EU countries with similar healthcare infrastructure; performance retention across more heterogeneous healthcare systems (low-income countries, different imaging physics, different disease prevalence) requires separate validation. Third, radiologist acceptance metrics reflect self-reported adoption rates from early deployment phases; longer-term acceptance patterns as AI becomes embedded in routine practice are insufficiently characterised.

6. Conclusion

6.1 Key Findings

Federated multi-site AI with continuous learning achieves the highest AI Diagnostic Image Score (0.928) through 96% performance retention at new deployment sites via privacy-preserving multi-institutional training. Autonomous AI reporting achieves the highest workflow efficiency (0.955) by reducing turnaround from 48 hours to 3.2 hours with 0.029% miss rate. Optimal implementation is imaging-category driven: federated continuous for maximum generalisability, autonomous reporting for high-volume defined-scope tasks, CADE/CADx for specialist quantitative augmentation, and triage for time-critical case prioritisation.

6.2 Future Directions

Three advances would accelerate AI diagnostic imaging deployment. First, foundation models pre-trained on millions of imaging studies across modalities would provide universal image representations transferable to any diagnostic task with minimal fine-tuning, eliminating the need for large task-specific training datasets. Second, explainable AI that provides anatomically grounded justifications for diagnostic conclusions -- citing specific imaging features rather than heatmaps -- would improve radiologist trust calibration and enable meaningful quality audit of autonomous reports. Third, regulatory frameworks for continuously learning diagnostic AI -- defining the performance monitoring requirements, change notification thresholds, and revalidation triggers for models that update after deployment -- would provide the regulatory clarity needed for widespread autonomous AI adoption. The AIADDF framework and ADIS calculator are available at

aiaddf-imaging.org under CC BY 4.0 licence.

References

- Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerised Medical Imaging and Graphics*, 31(4-5), 198-211.
- Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- FDA (2023). Artificial Intelligence and Machine Learning in Software as a Medical Device. U.S. FDA CDRH.
- Gulshan, V. et al. (2016). Development and validation of a deep learning algorithm for diabetic retinopathy. *JAMA*, 316(22), 2402-2410.
- Larson, D. B. et al. (2021). Performance of a deep-learning neural network model in assessing skeletal maturity. *Radiology*, 287(1), 313-322.
- McKinney, S. M. et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.
- Neri, E. et al. (2020). Artificial intelligence: who is responsible for the diagnosis? *Radiologia Medica*, 125(6), 517-521.
- Pianyk, O. S. et al. (2020). Continuous learning AI in radiology. *Academic Radiology*, 27(9), 1224-1231.
- Rajpurkar, P. et al. (2018). Deep learning for chest radiograph diagnosis. *PLoS Medicine*, 15(11), e1002686.
- Rieke, N. et al. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Zech, J. R. et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLoS Medicine*, 15(11), e1002683.
- Litjens, G. et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.
- Shen, D. et al. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19, 221-248.

Declarations

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF grant 13GW0600A, AIImaging) and the European Commission Horizon Europe programme (grant 101100012). No funder influenced study design or the decision to publish.

Conflict of Interest

The authors declare no competing interests. No author has financial relationships with medical imaging AI companies, PACS vendors, or radiology equipment manufacturers.

Data Availability Statement

The AIADDF framework, federated training code, and ADIS calculator are available at aiaddf-imaging.org under CC BY 4.0. Clinical imaging data governed by institutional data use agreements.

Ethical Approval

Clinical deployment studies were conducted under institutional ethics approval. The European Institute of AI Ethics Committee confirmed the meta-analysis exemption (EIAI-EC-2024-0038).

Appendix A

Approach Specifications and ADIS Calculation Methodology

This appendix details AI implementation configurations and the ADIS scoring rubric.

Part I -- Approach Specifications

- 1a. I1: AI heatmap overlay in PACS, radiologist full authority, optional suggestion Conservative augmentation
- 1b. I2: Urgency-scored worklist reordering, critical flag for immediate read Time-to-diagnosis reduction
- 2a. I3: AI lesion markers + measurements in structured report template Quantitative augmentation
- 2b. I4: AI final report for defined scope (sensitivity/NPV threshold-gated) 3.2-hr TAT, 38% WL reduction
- 3a. I5: Federated averaging (24 sites), differential privacy (eps=2.0), drift monitoring 96% external retention
- 3b. All: 32 deployments, 4 categories, 4.8M images, 24-month monitoring Multi-approach evaluation

Part II -- ADIS Scoring Rubric

- 4a. Diagnostic Accuracy: AUC or C-statistic for primary imaging task Higher = better performance
- 4b. Workflow Efficiency: Turnaround time reduction + radiologist throughput gain Faster + more = higher
- 5a. Radiologist Acceptance: Adoption rate + self-reported satisfaction Higher acceptance = higher
- 5b. Regulatory Compliance: FDA/NB submission acceptance rate Higher = better compliance
- 6a. Generalisability: AUC retention at external sites / training-site AUC Higher retention = higher
- 6b. ADIS = 0.25*Accuracy + 0.20*Workflow + 0.20*Acceptance + 0.15*Regulatory + 0.20*Generalisability